Opinion

Cell PRESS

Cellulases: ambiguous nonhomologous enzymes in a genomic perspective

Leonid O. Sukharnikov[1,2,4], Brian J. Cantwell[1,4], Mircea Podar[1,2,4] and Igor B. Zhulin[1,3,4,5]

[1] BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[2] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[3] Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[4] Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN 37996, USA
[5] Department of Microbiology, University of Tennessee, Knoxville, TN 37996, USA

The key material for bioethanol production is cellulose, which is one of the main components of the plant cell wall. Enzymatic depolymerization of cellulose is an essential step in bioethanol production, and can be accomplished by fungal and bacterial cellulases. Most of the biochemically characterized bacterial cellulases come from only a few cellulose-degrading bacteria, thus limiting our knowledge of a range of cellulolytic activities that exist in nature. The recent explosion of genomic data offers a unique opportunity to search for novel cellulolytic activities; however, the absence of clear understanding of structural and functional features that are important for reliable computational identification of cellulases precludes their exploration in the genomic datasets. Here, we explore the diversity of cellulases and propose a genomic approach to overcome this bottleneck.

Cellulose and cellulases

The dramatic rate of fossil fuels depletion and the resultant global economic and environmental consequences have spurred the search for alternative renewable energy sources such as biofuels. One of the promising materials for biofuel production is plant biomass [1], which contains large amounts of the sugar polymers cellulose (a polymer of β-1,4 linked glucose) and hemicelluloses (polymers composed of xylose, mannose, galactose, rhamnose, arabinose and other sugars [2]. These can be broken down by a mixture of enzymes into simple sugars that are fermentable to produce ethanol [3]. Although cellulose is largely present as crystalline fibers that are highly resistance to hydrolysis, its biomass content is typically larger than that of hemicellulose [3], and consequently, cellulases are the key enzymes for bioethanol production. Individual cellulose polymers form rigid microfiber structures that are stabilized by inter- and intramolecular hydrogen bonds and van der Waals interactions between glucose residues in the fibers, which significantly contributes to its resistance [4]. This network of bonds leads to a mostly uniform arrangement of fibers and the resultant crystalline cellulose lacks enzyme-accessible surface morphologies, further enhancing resistance to hydrolysis [5].

All cellulases are glycoside hydrolase (GH) enzymes that utilize the same catalytic mechanism of acid–base catalysis, with inversion or retention of glucose anomeric configuration [6]. There are two common types of the cellulase active sites. GHs with open (groove, cleft) active sites typically exhibit endocellulolytic activity (endocellulases), binding anywhere along the length of the cellulose molecule and hydrolyzing the β-1,4 glycosidic linkage, whereas those with tunnel-like active sites exhibit exocellulolytic activity (cellobiohydrolases), binding at the ends of the cellulose molecule [7] and producing unit-length oligosaccharide products. Typically, exocellulases are processive enzymes, that is, they are attached to the cellulose chain until it is completely hydrolyzed [7,8], whereas endocellulases can be both processive and non-processive [7]. Efficiency of processive cellulases can greatly contribute to the rate-limiting step of cellulose hydrolysis [8]. Cellulases with endo- mode of action appear to be represented by a larger number of protein folds (Table 1). This indicates that endocellulases are either more evolutionarily diverse or many novel exocellulases are yet to be found [9]. Many cellulases are multidomain proteins, and in addition to the catalytic domain, have accessory domains such as carbohydrate binding modules (CBMs) connected by a flexible linker [10]. The main role of CBMs is to help cellulases bind cellulose, although they might also participate in initial disruption of cellulose fibers [11]. Cellulases preferentially bind to the amorphous or somewhat disordered (e.g. through acid pretreatment) regions on the surface of the crystalline cellulose fiber [12]. Endocellulases (sometimes along with CBMs) help to disrupt the cellulose fibers and create accessible ends, whereas cellobiohydrolases continue the degradation by removing di- and oligosaccharides (usually 2–4 residues) from the ends of the disrupted cellulose fibers [13].

Lack of standards in cellulase enzymology

Several biochemical methods are commonly used to determine the substrate-specificity of cellulases and the endo-/exo-

Corresponding author: Zhulin, I.B. (joulineib@ornl.gov).

Table 1. Examples of cellulases with endo- and exo- modes of action

CAZy family	Accession Number	Fold	Mode of action	Reference
GH5	Q47916	$(\beta/\alpha)_8$	Endo	[19]
GH5	CAB76938.1	$(\beta/\alpha)_8$	Exo	[51]
GH6	Q53488	atypical β/α barrel	Endo	[52]
GH6	AAA62211.1	atypical β/α barrel	Exo	[53]
GH7	P56680	β-jelly roll	Endo	[54]
GH7	A7LN91	β-jelly roll	Exo	[27]
GH8*	AAA73867.1	$(\alpha/\alpha)_6$	Endo	[55]
GH9	Q02934	$(\alpha/\alpha)_6$	Endo	[56]
GH9	Q6RSN8	$(\alpha/\alpha)_6$	Exo	[57]
GH12*	O33897	β-jelly roll	Endo	[25]
GH23*	2XQO_A	α_8 superhelical	Endo	[33]
GH44*	Q934F9	$(\beta/\alpha)_8$	Endo	[58]
GH45*	Q9P868	β_6-barrel	Endo	[59]
GH48	P37698	$(\alpha/\alpha)_6$	Endo	[60]
GH48	Q8KKF7	$(\alpha/\alpha)_6$	Exo	[61]
GH51*	P77865	$(\beta/\alpha)_8$	Endo	[26]
GH61*	O14405	β-sandwich with an Ig-like topology. $\beta_9\alpha_5$	Endo	[62]
GH74*	Q9WYE1	sevenfold β-propeller	Endo	[63]

*GH8 does not have confirmed cellulases with exo- mode of action in CAZy [30]

mode of action. The reducing sugar assays involve incubating purified enzyme with cellulose-containing substrates along with a binding reagent (i.e. dinitrosalicylic acid), which reacts with glucose, released during the incubation, to create fluorescent compounds, which are then detected spectrophotometrically [14]. In the halo assay, a gene that codes for a putative cellulase is introduced into a non-cellulolytic microorganism, such as *Escherichia coli*, which is then grown on cellulose substrates stained with Congo Red. Colonies that carry cellulase genes are screened by formation of halo plaques that result from degradation of the stained cellulose by the bacterial colony [15]. Viscosimetry and TLC assays are commonly used to determine exo- versus endo- modes of action, for example, exocellulases reduce viscosity of solutions of carboxymethyl cellulose (CMC) much slower than endocellulases, whereas running incubation products of a cellulase on a gel shows whether shorter, such as glucose, cellobiose (exo-mode of action) or longer oligosaccharides, such as cellotriose, cellotetraose (endo- mode of action) are present [16,17].

Adding to the challenge of biochemical characterization of cellulases is the multisubstrate specificity. Many of the biochemically confirmed cellulases are active on a variety of substrates in addition to cellulose, such as xylan, lichenan and mannan. For example, cellulase Cel5E from *Pseudomonas fluorescens* is active on CMC, lichenan, Avicel (or microcrystalline cellulose) and acid-swollen cellulose but completely inactive on xylan [18]. Cellulase CelG form *Fibrobacter succinogenes* belongs to the same GH family 5 but shows high activity on CMC and xylan, and is completely inactive on Avicel and lichenan [19]. By contrast, some cellulases are active only on cellulose derivatives. For example, a GH family 5 cellulase cel5B from *Termobifida fusca* is able to degrade only cellulose-containing substrates [CMC, Avicel and MN300 (native fibrous cellulose)], but is completely inactive on other substrates [20].The vast majority of researchers use CMC degradation as an indication of cellulolytic activity. Therefore, here we consider documented CMC hydrolysis as the minimum requirement for an enzyme to be annotated as a biochemically confirmed cellulase. The multisubstrate specificity of cellulases and the persistent lack of data about activity on substrates other than CMC emphasize the need for adoption of a universal methodology for cellulase validation and characterization (Box 1).

Box 1. Current problems in cellulase studies and proposed solutions

1. Experimental

2.2 Lack of standardization in the use of certain assays and substrates for experimental cellulase determination:

- *Devise a standard assay or a set of assays for unambiguous and reliable identification of cellulases.*

1.2 Poor taxonomic representation among experimentally studied organisms:

- *Obtain genome sequences and biochemically characterize potential cellulases from taxonomically diverse organisms*

2. Computational

2.2 Cellulases are found in 12 unrelated protein families

- *Develop a natural classification system for each cellulase-containing protein family*

2.2 There are multiple substrate specificities other than cellulose in each of the cellulase-containing families. There are no known genomic markers for cellulases. Current models for genomic identification of cellulases are not specific:

- *Identify class-specific genomic markers for cellulases*
- *Develop sensitive, cellulase-specific models*
- *Validate models via iterative experiment–computation approach*

Known cellulolytic bacteria: a few of the many

Bacteria that are either known to be or potentially could be cellulolytic are widely distributed in nature. However, the best studied cellulose degraders, such as *Clostridium thermocellum*, *Clostridium cellulolyticum* and *Caldicellulosiruptor bescii* (previously known as *Anaerocellum thermophilum*) belong to the same phylum, the Firmicutes. Despite numerous studies of microbial cellulolytic apparatus [21–24], only about 20 genomes of known cellulose degraders have been fully sequenced to date. Recent genomic studies have identified many bacteria that contain arrays of various GHs (many of which could be cellulases [21,22]). Therefore, it is likely that only a small fraction of the cellulolytic world has been annotated and studied to date, and more experimental and genomic investigation of potential cellulase degraders from diverse taxa and habitats is needed.

CAZy database: a bridge from enzymology to genomics

The CAZy (Carbohydrate-Active Enzymes) database provides classification of enzymes (e.g. GHs, glycosyl transferases) and substrate-binding modules involved in various types of carbohydrate metabolism based on sequence comparison. All known cellulases are found within 12 GH families of the CAZy database, and can be described with two enzyme commission numbers: EC 3.2.1.4 (endoglucanase) and EC 3.2.1.91 (cellobiohydrolase). Families GH5 and GH9 appear to have the largest number of biochemically characterized cellulases. This could be partly because cellulases from these families are abundant in the model cellulolytic bacteria. Yet, many enzymes that effectively hydrolyze cellulose belong to other, smaller CAZy families, for example, the Cel12A cellulase from *Rhodothermus marinus* (GH12) [25], endoglucanase F from *F. succinogenes* S85 (GH51) [26], and CbhI from *Fusicoccum* sp.(GH7) [27]. This indicates that the search for potential efficient cellulases should be substantially broadened.

Although the collection of carbohydrate enzyme data in CAZy provides a very useful resource for enzymologists, annotations could be significantly improved. For example, the term 'characterized' in CAZy is applied equally to proteins that have been characterized biochemically and to those for which the functions have been predicted computationally. As we show, computational predictions for cellulases are currently unreliable; therefore knowing the source of information for annotation would be helpful. Nevertheless, CAZy provides a much needed connection between enzymology and genomics and can be considerably enhanced with improved computational models.

Challenges of genomic identification of cellulases

To search for cellulases in the ever-increasing genomic and metagenomic data, reliable sequence-based methods for their identification must be available. Current computational methodologies require that proteins should be conserved sufficiently in sequence to carry out full-length sequence similarity searches (e.g. BLAST) or they should have specific markers, such as distinctive protein domains and domain combinations, motifs and accessory proteins (see [28] for details), to yield reliable predictions. To illustrate the problems of genomic identification of cellulases, we compare their relevant features to those of another common enzyme involved in carbohydrate metabolism, hexokinase (the first enzyme of the glycolysis pathway). BLAST searches with a hexokinase seem quite reliable, whereas those with confirmed cellulases produce much more ambiguous results, in which similar sequences can be annotated with a variety of definitions other than cellulase. Automated annotation of new genomes depends heavily on the identification of similar proteins by BLAST, therefore, this ambiguity greatly complicates identification of potential novel cellulases.

From a structural perspective, hexokinases belong to a single protein fold (Figure 1). All proteins that catalyze the ATP-dependent conversion of aldo- and keto-hexose sugars to the hexose-6-phosphate [29] have the same ribonuclease-H-like motif fold and belong to the same protein family, hexokinase. By contrast, proteins that catalyze the hydrolysis of the β-1,4 glucoside bond using the same mechanism of acid–base catalysis (cellulases) belong to at least eight unrelated protein folds (Figure 1), which further differentiates into even more protein families [30]. For example, cellulase Cel5E from *P. fluorescens* has an $(\beta/\alpha)_8$ fold and belongs to GH family 5 [18] (family classification according to the CAZy database [30]); cellulase Egl-257 from *Bacillus circulans* has an $(\alpha/\alpha)_6$ barrel fold and belongs to GH family 8 [31]; and cellulase cel44a from *C. thermocellum* has a TIM-like barrel and β-sandwich domain fold and belongs to GH family 44 [32]. Recent biochemical and genomics studies have identified cellulases in 11 or 13 CAZy families [9,30,33]. Cellulases therefore are representatives of a large class of nonhomologous isofunctional enzymes [34], that is, proteins that catalyze the same biochemical reaction, which have evolved independently and are unrelated in sequence and structure. Therefore, in contrast to hexokinase, cellulases from each protein family must be treated as independent cases in any type of genomic analysis. This is a potential problem, which is easily resolved, although it dramatically increases the amount of data analysis.

In addition to pairwise sequence similarity searches, the second powerful tool used in automated annotation is protein domain architecture, which is identified using domain-specific profile Hidden Markov Models (HMMs). HMMs are built from multiple sequence alignments and represent probabilities of certain amino acids being located at certain positions in a domain. Again, hexokinases can be easily distinguished from other enzymes based on their domain architecture (Figure 2). Nearly all hexokinases display a conserved combination of two protein domains termed 'hexokinase_1' (Pfam accession PF00349) in the N terminus and 'hexokinase_2' (PF03727) in the C terminus. Detection of these domains in any protein sequence unambiguously identifies it as a hexokinase. There is essentially no diversity in the domain architecture of hexokinases: <10% of sequences exhibit a duplicated version of the dual domain protein (Figure 2) and <1% contain other unrelated domains.

By contrast, identification of cellulases by domain architecture is problematic because of two characteristics. First, cellulases display an extremely wide diversity of domain architectures even within the same protein family (Figure 2). Second, and more importantly, the HMMs

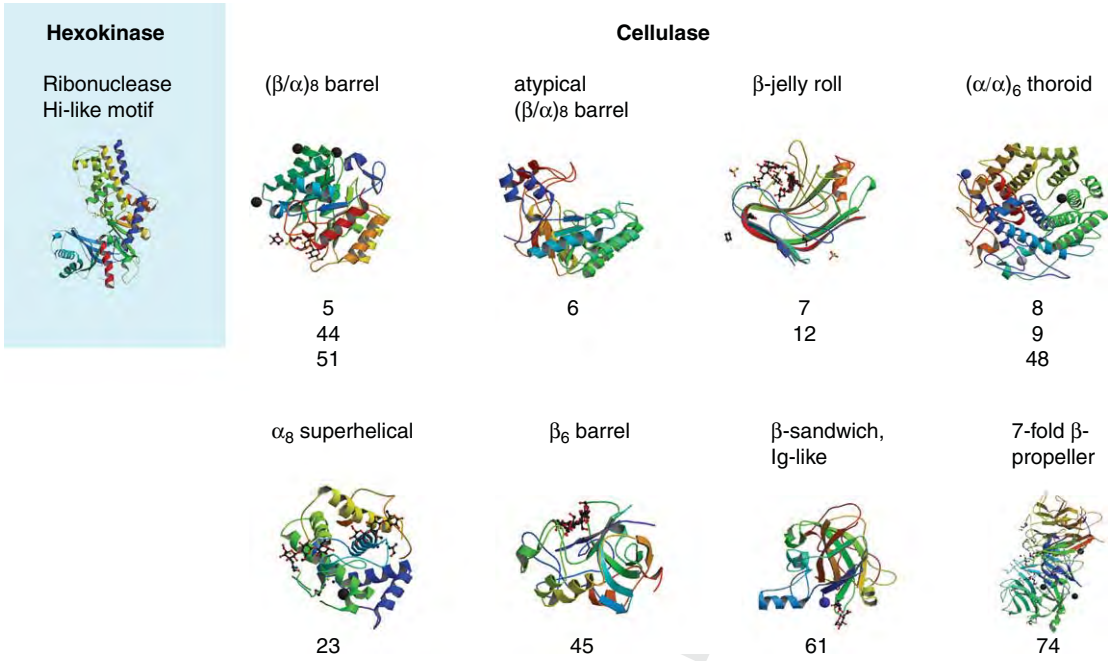

Figure 1. Hexokinase and cellulases: structural conservation and diversity. Corresponding CAZy families are listed below structures (images are taken from the RCSB PDB (www.pdb.org). The following labels correspond to PDB accession numbers: 'Hexokinase' - 1ig8[43]; 'GH5, GH44, GH51' - 1e5j[44]; 'GH6' - 2boe[45]; 'GH7, GH12' - 2jen[46]; 'GH8, GH9, GH48' - 1ia6[47]; 'GH23' – 2xqo[33]; 'GH45' - 4eng[48]; 'GH61' - 2vtc[49]; 'GH74' - 2cn2[50].

currently available to recognize cellulases are built from multiple alignments that include cellulases and similar-in-sequence non-cellulases, and thus are not able to differentiate between members of the same protein family that have different substrate specificities. To illustrate this problem, we compare known activities of enzymes that belong to GH5 family (Pfam PF00150, *Cellulase)* to that of GH19 family (Pfam 00182, *Glyco_hydro_19*). GH19 is a large family (>1000 sequences in current databases) in which all 165 experimentally studied enzymes exhibit a single activity – chitinase (EC 3.2.1.14). GH5 family is comparable in size (just over 2000 sequences), however, among 373 experimentally studied enzymes from this family, at least 12 different activities other than cellulase have been reported (data from the CAZy database). Therefore, datasets retrieved with the current *Cellulase* domain model [35] might contain primarily non-cellulases and therefore would not be helpful to experimentalists.

Metagenomes: 'gold mines' that need sluicing rather than panning

Metagenomic exploration of environments where lignocellulose is being effectively decomposed is the most promising path towards discovery of novel cellulases. Recent advances in metagenomics have resulted in generating genomic datasets from diverse environments, including fresh water [36], the ocean [37], guts of insects [38], ruminants [39], and even human intestines [40]. Such datasets have great potential to reveal novel cellulolytic capabilities. For example, the recent metagenomic study of a cow rumen has uncovered tens of thousands of putative cellulases [41], thus truly becoming a gold mine for their future exploration. However, the same computational problems that we have outlined above have prevented unambiguous identification of true cellulases in this dataset; investigators have had to narrow down their list of targets for experimental validation randomly, and the reported success rate is around 50% [41]. Clearly, a more efficient and cost-effective method of mining is urgently needed.

Proposed computational solutions

Natural classification systems based on evolutionary relationships between sequences are instrumental in dealing with complex biological systems [28]. Cellulases are found in protein families with different evolutionary histories and belong to different protein folds, therefore, the evolutionary path of each cellulase-containing protein family must be evaluated independently. To build a natural classification system for cellulases, classes must be defined using a phylogenomic approach, in which related sequences of enzymatic domains are collected, properly aligned using available structural information, and then clustered (e.g. via phylogenetic tree construction). Independent genomic markers, such as specific combinations of enzymatic and accessory domains, and genome neighborhoods, must be identified for each individual class. To link biochemical activities to genomically identified classes, all available information on substrate specificity of individual sequences must be mapped onto individual classes.

An effective natural classification scheme will assist in searching for novel cellulolytic activities in genomic datasets by identifying markers that can be used to differentiate cellulases from related enzymes with different substrate specificity. Although it is difficult to discern a pattern of accessory domains when looking at all sequences of a given GH family, focusing on a class of related proteins

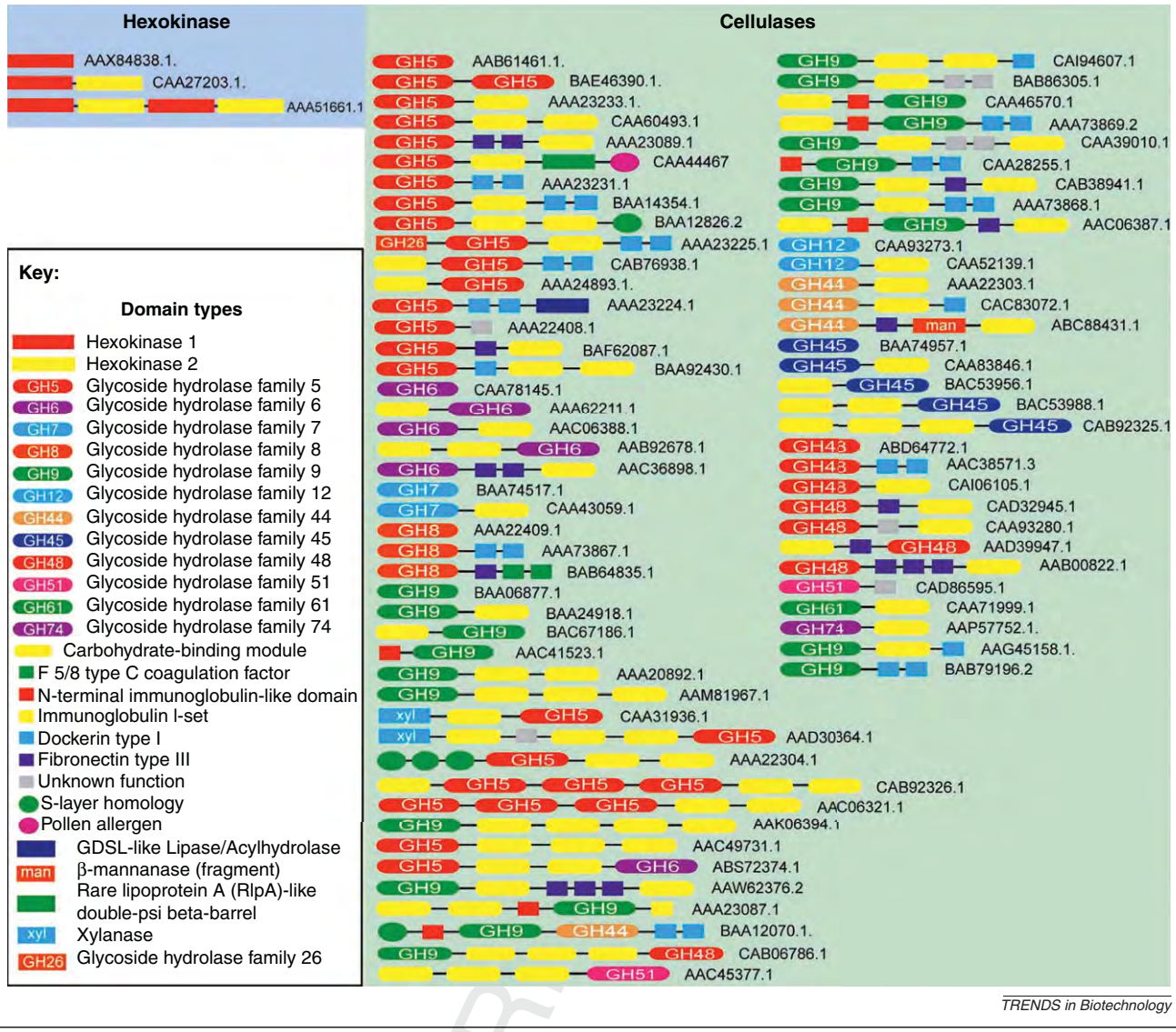

Figure 2. Hexokinase and cellulases: conservation and diversity of domain architectures. Accession numbers for sequences are shown.

within a family might reveal specific accessory domains associated with that class. Most of the biochemically confirmed cellulases have carbohydrate-binding module domains (Figure 2), and cellulases with the same catalytic domains tend to degrade resistant crystalline cellulose more efficiently if they contain a larger number of CBMs [27,42]. Thus, identifying CBMs that are class-specific should be productive for better classification of the catalytic domains. Similarly, analysis of genome neighborhoods might reveal certain types of genes that are consistently found in proximity to genes that encode biochemically confirmed cellulases. Then, the presence of these genes in proximity to genes that encode unknown GHs suggest that it might be a cellulase (a 'guilty-by-association' approach). Lastly, analysis of the aligned sequences can identify class-specific patterns of conserved amino acids, whose potential role in substrate specificity can be revealed by mapping onto available 3D structures and homology models. Aligned sequences of specific classes can also be turned into specific and sensitive domain models (e.g. HMM) for each of the catalytic domains or, where appropriate, for their combinations with auxiliary domains. Such models could become an essential tool, to search specifically for cellulases in ever-increasing genomic and metagenomic datasets. With new, refined models it should be possible to reduce the search space for cellulases by orders of magnitude, and to provide experimentalists with a short list of enzymes that are more likely to be a true cellulase. Newly developed cellulase-specific models should be deposited to relevant databases (e.g. Pfam and CAZy) to ensure their availability to the scientific community.

The need for specific cellulase models is pressing. We now have hundreds of environmental sequencing samples that contain >1 billion sequences, including datasets from such cellulolytic environments as termite gut [38] and cow rumen [41]. Together with still largely unexplored complete genomes of cellulose degraders, metagenomic data create a great reservoir for finding novel cellulolytic activities. There is also a need for a much closer collaboration between experimentalists and computational scientists in this area. The existing biochemical characterization has

been performed on a small subset of closely related organisms; therefore, a substantial number of experiments will be needed to fill gaps on substrate specificity within newly identified classes of cellulase-containing families. Better standardization of cellulase assays and more thorough assessment of activity on a variety of carbohydrate polymers will greatly improve our ability to link sequence classes to enzyme activities.

Conclusion

In conclusion, we would like to point out several contentious areas in practical biotechnology that might be addressed using computational genomics in the near future. First, there is a clear difference between enzymes in their ability to hydrolyze cellulose substrates, such as untreated, raw and pretreated plant material (e.g. switchgrass and wood pulp). Such differences could be caused by inherent enzymatic domain properties (e.g. K_{cat}, product release) or associations with accessory domains that enhance substrate binding (e.g. CBM). Thus, one of the targets for computational studies is associating the experimentally determined characteristics of various cellulases with catalytic site conservation and accessory domain architecture. The more enzymes with known sequence, structure and biochemical activities that are available, the more powerful associations and therefore predictions can be made. The resultant data could be applicable to enzyme engineering as well, to search for better catalysts within a reduced sequence and structure space.

Second, many challenges are posed by the engineering of cellulases to be robust under harsh industrial settings (e.g. temperature, solvents, and ionic conditions). Hence, better understanding of the cellulase active site and enzymatic functions at the sequence level could enable protein engineering that can maintain catalytic properties while enhancing protein robustness.

Finally, better communication between leading world cellulase researchers must be established to enable standardization of experimental and computational approaches to studies of cellulases. One way of accomplishing this goal would be creation of a freely available internet resource that would include internationally accepted methodologies for biochemical and computational cellulase studies, and a curated and updatable database of confirmed cellulases. To improve accessibility to such a resource, we recommend merging it with already existing web resources, such as the CAZy database mentioned above.

Acknowledgments

The BioEnergy Science Center (BESC) is a United States Department of Energy Bioenergy Research Center that is supported by the Office of Biological and Environmental Research in the Department of Environment Office of Science. We thank numerous colleagues across BESC for stimulating discussions.

References

1 Lynd, L.R. *et al.* (2008) How biotech can transform biofuels. *Nat. Biotechnol.* 26, 169–172
2 Popper, Z. *et al.* (2011) Evolution and diversity of plant cell walls: from algae to flowering plants. *Annu. Rev. Plant Biol.* 62, 8.1–8.24
3 Ragauskas, A.J. *et al.* (2006) The path forward for biofuels and biomaterials. *Science* 311, 484–489
4 Cheng, G. *et al.* (2011) Transition of cellulose crystalline structure and surface morphology of biomass as a function of ionic liquid pretreatment and its relation to enzymatic hydrolysis. *Biomacromolecules* 12, 933–941
5 Zhou, W. *et al.* (2009) Cellulose hydrolysis in evolving substrate morphologies I: a general modeling formalism. *Biotechnol. Bioeng.* 104, 261–274
6 Davies, G. and Henrissat, B. (1995) Structures and mechanisms of glycosyl hydrolases. *Structure* 3, 853–859
7 Kurasin, M. and Väljamäe, P. (2011) Processivity of cellobiohydrolases is limited by the substrate. *J. Biol. Chem.* 286, 169–177
8 Beckham, G.T. *et al.* (2011) Applications of computational science for understanding enzymatic deconstruction of cellulose. *Curr. Opin. Biotechnol.* 22, 231–238
9 Gilbert, H.J. (2010) The biochemistry and structural biology of plant cell wall deconstruction. *Plant Physiol.* 153, 444–455
10 Fontes, C.M. and Gilbert, H.J. (2010) Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. *Annu. Rev. Biochem.* 79, 655–681
11 Shoseyov, O. *et al.* (2006) Carbohydrate binding modules: biochemical properties and novel applications. *Microbiol. Mol. Biol. Rev.* 70, 283–295
12 Canilha, L. *et al.* (2011) A study on the pretreatment of a sugarcane bagasse sample with dilute sulfuric acid. *J. Ind. Microbiol. Biotechnol.* DOI: 10.1007/s10295-10010-10931-10292
13 White, A.R. and Brown, R.M., Jr (1981) Enzymatic hydrolysis of cellulose: visual characterization of the process. *Proc. Natl. Acad. Sci. U.S.A.* 78, 1047–1051
14 Miller, G.L. (1959) Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Anal. Chem.* 31, 426–428
15 Teather, R.M. and Wood, P.J. (1982) Use of Congo-Red polysaccharide interactions in enumeration and characterization of cellulolytic bacteria from bovine rumen. *Appl. Environ. Microbiol.* 43, 777–782
16 Irwin, D.C. *et al.* (1993) Activity studies of eight purified cellulases: specificity, synergism, and binding domain effects. *Biotechnol. Bioeng.* 42, 1002–1013
17 Cohen, R. *et al.* (2005) Processive endoglucanase active in crystalline cellulose hydrolysis by the brown rot basidiomycete *Gloeophyllum trabeum*. *Appl. Environ. Microbiol.* 71, 2412–2417
18 Hall, J. *et al.* (1995) The non-catalytic cellulose-binding domain of a novel cellulase from *Pseudomonas fluorescens subsp. cellulosa* is important for the efficient hydrolysis of Avicel. *Biochem. J.* 309, 749–756
19 Iyo, A.H. and Forsberg, C.W. (1996) Endoglucanase G from *Fibrobacter succinogenes* S85 belongs to a class of enzymes characterized by a basic C-terminal domain. *Can. J. Microbiol.* 42, 934–943
20 Posta, K. *et al.* (2004) Cloning, characterization and phylogenetic relationships of *cel5B*, a new endoglucanase encoding gene from *Thermobifida fusca*. *J. Basic Microbiol.* 44, 383–399
21 Weiner, R.M. *et al.* (2008) Complete genome sequence of the complex carbohydrate-degrading marine bacterium Saccharophagus degradans strain 2-40 T. *PLoS Genet.* 4, e1000087
22 Barabote, R.D. *et al.* (2009) Complete genome of the cellulolytic thermophile *Acidothermus cellulolyticus* 11B provides insights into its ecophysiological and evolutionary adaptations. *Genome Res.* 19, 1033–1043
23 Moraïs, S. *et al.* (2010) Cellulase–xylanase synergy in designer cellulosomes for enhanced degradation of a complex cellulosic substrate. *MBio* 1, e00285-00210
24 Rincon, M.T. *et al.* (2010) Abundance and diversity of dockerin-containing proteins in the fiber-degrading rumen bacterium Ruminococcus flavefaciens FD-1. *PLoS ONE* 5, e12476
25 Crennell, S.J. *et al.* (2002) The structure of *Rhodothermus marinus* Cel12A, A highly thermostable family 12 endoglucanase, at 1. 8 Å resolution. *J. Mol. Biol.* 320, 883–897
26 Malburg, S.R. *et al.* (1997) Catalytic properties of the cellulose-binding endoglucanase F from *Fibrobacter succinogenes* S85. *Appl. Environ. Microbiol.* 63, 2449–2453
27 Kanokratana, P. *et al.* (2008) Identification and expression of cellobiohydrolase (CBHI) gene from an endophytic fungus. Fusicoccum sp. (BCC4124) in Pichia pastoris. *Protein Expr. Purif.* 58, 148–153
28 Wuichet, K. and Zhulin, I.B. (2010) Origins and diversification of a complex signal transduction system in prokaryotes. *Sci. Signal.* 3, ra50

29 Finn, R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.* 38, D211–222
30 Cantarel, B.L. *et al.* (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* 37, D233–238
31 Hakamada, Y. *et al.* (2002) Enzymatic properties, crystallization, and deduced amino acid sequence of an alkaline endoglucanase from Bacillus circulans. *Biochim. Biophys. Acta* 1570, 174–180
32 Kitago, Y. *et al.* (2007) Crystal structure of Cel44A, a glycoside hydrolase family 44 endoglucanase from *Clostridium thermocellum*. *J. Biol. Chem.* 282, 35703–35711
33 Bras, J.L.A. *et al.* (2011) Structural insights into a unique cellulase fold and mechanism of cellulose hydrolysis. *Proc. Natl. Acad. Sci. U.S.A.* 108, 5237–5242
34 Omelchenko, M.V. *et al.* (2010) Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol. Direct* 5, 31
35 Zhou, F. *et al.* (2009) Large-scale analyses of glycosylation in cellulases. *Genomics Proteomics Bioinform.* 7, 194–199
36 Debroas, D. *et al.* (2009) Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget-France). *Environ. Microbiol.* 9, 2412–2424
37 Yooseph, S. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5, e16
38 Warnecke, F. *et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450, 560–565
39 Brulc, J.M. *et al.* (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1948–1953
40 Qin, J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65
41 Hess, M. *et al.* (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 331, 463–467
42 Baba, Y. *et al.* (2005) Alternative splicing produces two endoglucanases with one or two carbohydrate-binding modules in *Mucor circinelloides*. *J. Bacteriol.* 187, 3045–3051
43 Kuser, P.R. *et al.* (2000) The high resolution crystal structure of yeast hexokinase PII with the correct primary sequence provides new insights into its mechanism of action. *J. Biol. Chem.* 275, 20814–20821
44 Fort, S. *et al.* (2001) Mixed-linkage cellooligosaccharides: a new class of glycoside hydrolase inhibitors. *Chembiochem* 2, 319–325
45 Larsson, A.M. *et al.* (2005) Crystal structure of *Thermobifida fusca* endoglucanase Cel6A in complex with substrate and inhibitor: the role of tyrosine Y73 in substrate ring distortion. *Biochemistry* 44, 12915–12922
46 Gloster, T.M. *et al.* (2007) Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families Gh5 and Gh12. *J. Biol. Chem.* 282, 19177–19189
47 Parsiegla, G. *et al.* (2002) Crystal structure of the cellulase Cel9 M enlightens structure/function relationships of the variable catalytic modules in glycoside hydrolases. *Biochemistry* 41, 11134–11142
48 Davies, G.J. *et al.* (1996) Structure determination and refinement of the *Humicola insolens* endoglucanase V at 1.5 A resolution. *Acta Crystallogr. D: Biol. Crystallogr.* 52, 7–17
49 Karkehabadi, S. *et al.* (2008) The first structure of a glycoside hydrolase family 61 member, Cel61B from the Hypocrea jecorina, at 1. 6 A resolution. *J. Mol. Biol.* 383, 144–154
50 Martinez-Fleites, C. *et al.* (2006) Crystal structures of *Clostridium thermocellum* xyloglucanase, Xgh74A, reveal the structural basis for xyloglucan recognition and degradation. *J. Biol. Chem.* 281, 24922–24933
51 Zverlov, V.V. *et al.* (2002) A newly described cellulosomal cellobiohydrolase, CelO, from Clostridium thermocellum: investigation of the exo-mode of hydrolysis, and binding capacity to crystalline cellulose. *Microbiology* 148, 247–255
52 3 Lin, F. *et al.* (1994) Cloning and sequencing of an endo-beta-1,4-glucanase gene mcenA from *Micromonospora cellulolyticum* 86W-16. *J. Ind. Microbiol.* 13, 344–350
53 Zhang, S. *et al.* (1995) Characterization of a *Thermomonospora fusca* exocellulase. *Biochemistry* 34, 3386–3395
54 Davies, G.J. *et al.* (1997) Oligosaccharide specificity of a family 7 endoglucanase: insertion of potential sugar-binding subsites. *J. Biotechnol.* 57, 91–100
55 Fierobe, H-P. *et al.* (1993) Purification and characterization of endoglucanase C from *Clostridium cellulolyticum*. Catalytic comparison with endoglucanase A. *Eur. J. Biochem.* 217, 557–565
56 Hazlewood, G.P. *et al.* (1993) Gene sequence and properties of CelI, a family E endoglucanase from *Clostridium thermocellum*. *J. Gen. Microbiol.* 139, 307–316
57 Schubot, F.D. *et al.* (2004) Structural basis for the exocellulase activity of the cellobiohydrolase CbhA from *Clostridium thermocellum*. *Biochemistry* 43, 1163–1170
58 Rincon, M.T. *et al.* (2001) EndB, a newly identified family 44 cellulase from the rumen cellulolytic bacterium *Ruminococcus flavefaciens* 17, binds to cellulose via a novel cellulose binding domain and to a 130kDa *R.flavefaciens* protein via a dockerin domain. *Appl. Environ. Microbiol.* 67, 4426–4431
59 Eberhardt, R.Y. *et al.* (2000) Primary sequence and enzymic properties of two modular endoglucanases, Cel5A and Cel45A, from the anaerobic fungus Piromyces equi. *Microbiology* 146, 1999–2008
60 Reverbel-Leroy, C. *et al.* (1996) Molecular study and overexpression of the *Clostridium cellulolyticum* celF cellulase gene in *Escherichia coli*. *Microbiology* 142, 1013–1023
61 Sanchez, M.M. *et al.* (2003) Exo-mode of action of cellobiohydrolase Cel48C from *Paenibacillus sp*. BP-23. *Eur. J. Biochem.* 270, 2913–2919
62 Karlsson, J. *et al.* (2001) Homologous expression and characterization of Cel61A (EG IV) of *Trichoderma reesei*. *Eur. J. Biochem.* 268, 6498–6507
63 Chhabra, S.R. *et al.* (2001) 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