Performance of Hidden Markov Models in Recovering the Standard Classification of Glycoside Hydrolases

Mariana Fonseca Rossi, Beatriz Mello and Carlos G Schrago

Department of Genetics, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.

Evolutionary Bioinformatics
Volume 13: 1–5

DOI: 10.1177/1176934317703401

SAGE

ABSTRACT: Glycoside hydrolases (GHs) are carbohydrate-active enzymes that assist the hydrolysis of glycoside bonds of complex sugars into carbohydrates. The current standard GH family classification is available in the CAZy database, which is based on the similarities of amino acid sequences and curated semi-automatically. However, with the exponential increase in data availability from genome sequences, automated classification methods are required for the fast annotation of coding sequences. Currently, the dbCAN database offers automatic annotations of signature domains from CAZy-defined classifications using a statistical approach, the hidden Markov models (HMMs). However, dbCAN does not contain the entire set of CAZy GH families. Moreover, no evaluation has been conducted so far of the viability of using HMM profiles as a means of automatically assigning GH amino acid sequences to the standard CAZy GH family classification itself. In this work, we performed a meta-analysis in which amino acid sequences from CAZy-defined GH families were used to build HMM family-specific profiles. We then queried a set with ~300 000 GH sequences against our database of HMM profiles estimated from CAZy families. We conducted the same evaluation against the available dbCAN HMM profiles. Our analyses recovered 65% of matches with the standard CAZy classification, whereas dbCAN HMMs resulted in 61% of matches. We also provided an analysis of the types of errors commonly found when HMMs are used to recover CAZy-based classifications. Although the performance of HMM was good, further developments are necessary for a fully automated classification of GH, allowing the standardization of GH classification among protein databases.

KEYWORDS: Automatic annotation, CAZy, dbCAN, protein family, HMM profiles, protein classification

RECEIVED: November 29, 2016. **ACCEPTED:** March 9, 2017.

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 418 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: CGS was funded by grants 310974/2015-1 and 44954/2016-9 and BM by grant 158819/2014-4, all from the National Research Council (CNPq).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Carlos G Schrago, Department of Genetics, Federal University of Rio de Janeiro, Rio de Janeiro 21941-617, Brazil.
Email: carlos.schrago@gmail.com

Introduction

Glycoside hydrolases (GHs) are Carbohydrate-Active enZymes (CAZymes) that assist the breakdown of glycoside bonds of complex sugars into carbohydrates.[1] They are extensively studied because of their role in the degradation of cellulose, hemicellulose, and lignin.[2] Over the past 2 decades, studies have been conducted aiming to organize GH diversity[3–7] based on their mechanism of action,[8] enzyme structure,[7–10] type of substrates, and the domains of life in which GHs occur.[11,12] Glycoside hydrolases were first classified based on substrate specificity and the type of catalytic reaction.[13] However, these classifications failed to account for both structural and evolutionary features. Therefore, a new classification scheme was proposed by Henrissat,[3] which is currently widely used and may be regarded as the standard GH classification. The Henrissat classification assigned GH to families based on the similarity of amino acid sequences and, consequently, on their secondary structure, combining methods of alignment and hydrophobic cluster analysis.[3,14,15] Glycoside hydrolase families were later compiled in the CAZy database according to the Henrissat classification, which currently encompasses >130 families.[6] Since then, this classification of CAZymes has been used in studies dealing with a diverse array of subjects, from the structure, specificity, and efficiency of GHs to the analysis of metagenomes.[10,16–18]

Currently, an automatic method of assigning sequences to GH groups is required[19] because of the huge generation of new data from genome projects via high-throughput sequencing technologies. The CAZy database employs a semi-automatic annotation using Pfam hidden Markov model (HMM) profiles[20] and BLAST (Basic Local Alignment Search Tool),[21] which are posteriorly curated by experts manually.[1] Although this procedure has generated a widely implemented GH classification, it is difficult to reproduce automatically. Several CAZy GH families are absent in databases such as Pfam, which makes difficult a cross-correspondence between protein databases.[20]

Recently, statistical methods based on HMMs have shown good performance in homology inference of amino acid sequences and have been widely used in evolutionary studies of protein families.[22–24] Thus, HMM may result in a viable alternative to the semi-automated approach of CAZy. As a consequence of the HMM method, a profile consisting of amino acid variation probabilities along alignment sites is inferred, indicating possible insertions and deletions.[25] In addition, HMM profiles indicate conserved and nonconserved positions within members of the same family.[22] An extensive collection of annotated HMM profiles is currently available in protein databases, such as Pfam,[20] which use this statistical approach for improvement of the classification criteria used to group protein families into clans.[20] Families belonging to the same clan share similarities in amino acid sequences, structure, and HMM profile.[20]

An automatic classification of GH families based on the CAZy database was proposed by Yin et al[26] and is currently available in the dbCAN database. For each CAZy GH family, dbCAN identifies signature domains and builds an HMM model to represent them.[26] However, dbCAN does not contain all GH families, and no overall evaluation has ever been conducted of the performance of the HMM method in recovering the standard GH classification implemented in CAZy.

In this study, we conducted the first comparative meta-analysis of the degree of correspondence between HMM-inferred profiles and their corresponding CAZy database classifications, which is largely regarded as the standard GH classification. When HMM profiles correctly recovered the GH family classification, we considered this family to be automated using this statistical approach. We showed that HMM profiles recovered the CAZy family classification well. Although mismatches between HMM and CAZy were found, they were not random, and errors could be systematically associated with specific GH families. This indicates that further developments in the automatic assignment of amino acid sequences to CAZy protein families may be targeted at those particular cases, allowing for a consistent classification of GH sequences among protein databases.

Methods

We downloaded, from GenBank, approximately 300 000 GH sequences classified by the CAZy database as of January 2017. This data set consisted of all available GH sequences from cellular organisms, including 5469 sequences from Archaea, 240 650 sequences from Bacteria, and 55 080 sequences from Eukarya. A total of 127 out of the 135 GH families listed in CAZy were studied. Eight families were excluded from the analysis because they were absent in cellular organisms. For the sake of comparison, we also analyzed the performance of HMM models readily available in the dbCAN database.[26]

To evaluate the recovery rate of the standard CAZy GH classification using HMM profiles, multiple alignments were conducted for amino acid sequences from each CAZy GH family, using the Clustal Omega algorithm.[27] The quality of GH alignments was evaluated using the transitive consistency score (TCS) statistics available in T-Coffee platform.[28] As TCS assigns a score for each site in an alignment, it allows verification of the overall alignment's quality. We were then able to check whether poor-quality alignments affected the inference of HMM profiles. The TCS of more than 500 was regarded as well supported.[28] To evaluate whether the frequency of gaps affected the inference of HMM profiles, the proportion of gaps in each alignment was calculated using Perl scripts. The TCS and gap proportions were estimated for both the alignments of complete sequences from CAZy GH families and the alignments of family domains from dbCAN.

The resulting 127 individual CAZy GH family alignments were then loaded into the HMMER 3.1 platform[29] to build their respective HMM profiles using the *hmmbuild* software. The inferred CAZy-based HMM profiles were posteriorly

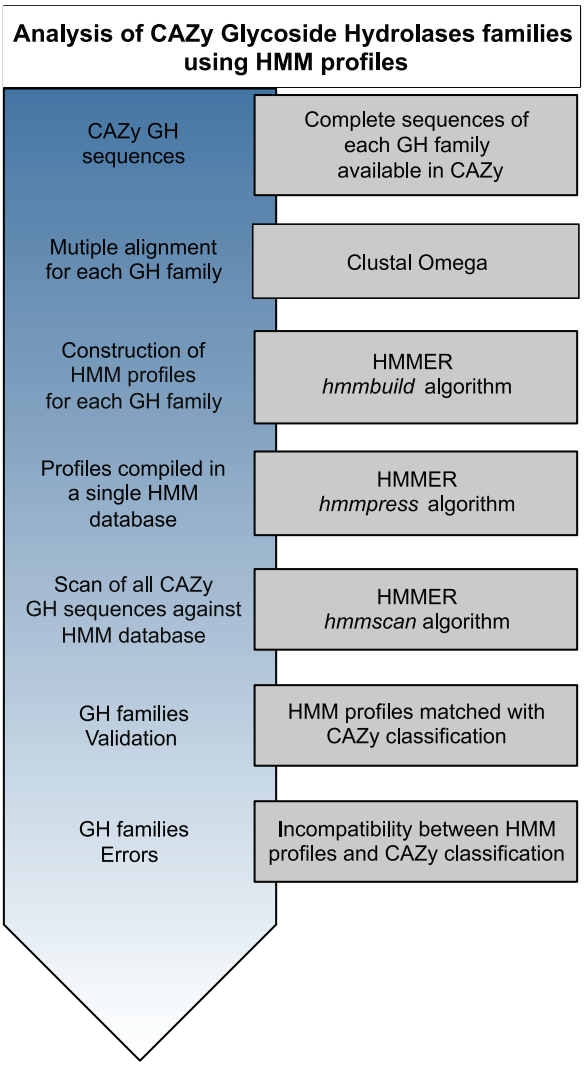

Figure 1. Steps used to evaluate the performance of retrieving CAZy GH families' classifications using HMM. GH indicates glycoside hydrolase; HMM, hidden Markov model.

combined into a single HMM database using the *hmmpress* software. This database of HMM profiles was hereafter referred to as *HMM-CAZy*. Finally, all GH amino acid sequences from cellular organisms available in CAZy, regardless of their CAZy-based classification, were queried against the complete *HMM-CAZy* database using the *hmmscan* software. We expected, thus, a one-to-one match between the CAZy GH classifications and their correspondent HMM profiles. Statistically, this was a measure of the power of the HMM method in recovering the CAZy classification. We also downloaded all available individual HMM profiles from dbCAN and built a single database of profiles, hereafter referred to as *HMM-dbCAN*, using the same strategy employed with the *HMM-CAZy*. We thus inferred the power of the HMM profiles built solely from protein domains in recovering the CAZy classification. In all steps, default HMMER parameters were used.

If an amino acid sequence assigned by CAZy to a given GH family matched the corresponding HMM profile, built using sequences from the same CAZy-defined GH family, it was counted as a correct match for this GH family because the

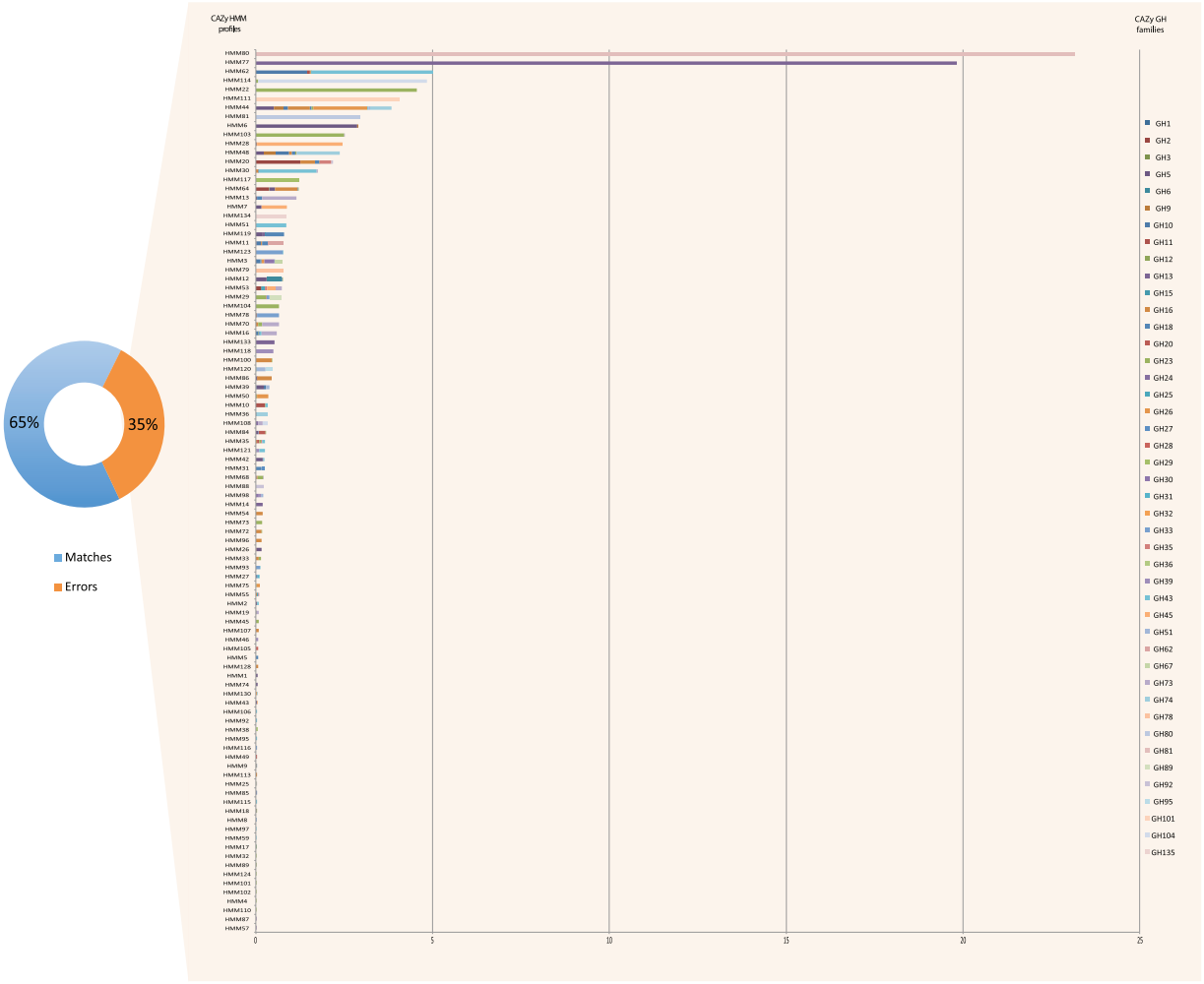

Figure 2. Frequency of errors, as shown in the left pie chart, and discrimination of GH families that presented incongruences between CAZy GH classification and HMM profiles inferred in this study. For each CAZy family, horizontal bars indicate the percentage of mismatches between the CAZy-defined family and the HMM profile. GH indicates glycoside hydrolase; HMM, hidden Markov model.

HMM assignment recovered the standard GH classification. However, a mismatch was computed when an amino acid sequence, defined by CAZy as a member of a given GH family, was not assigned to the respective HMM profile estimated for that same family. If a mismatch was found, we registered the HMM profile with the best hit for that sequence. We summarized the frequency of matches as well as mismatches for each family independently. Information on the steps of the HMM analysis is detailed in Figure 1.

Results and Discussion

When GH amino acid sequences were queried against the *HMM-CAZy* and *HMM-dbCAN* databases, positive matches between CAZy classifications and HMM profiles were recovered for 65% and 61% of GH families, respectively. Mismatches between the standard classification and HMM profiles were thus found in 35% and 39% of CAZy families when queried against the profiles of the *HMM-CAZy* and *HMM-dbCAN* databases, respectively (Figures 2 and 3, Supplementary Tables S1 and S2). Furthermore, in 24% of cases, identical mismatches were found when analyzing both HMM databases (*HMM-CAZy* and *HMM-dbCAN*) (Supplementary Table S3).

Mismatches between the CAZy protein family classification and the retrieved HMM profiles consisted of 2 major types (Figure 4). In the first type, several amino acid sequences from a given CAZy GH family were consistently assigned to an HMM profile estimated for a different CAZy GH family (Figures 2 and 3). For these cases, the greatest discrepancies were observed in GH families 73 and 81, in which the queries against the *HMM-CAZy* database resulted in a high frequency of incorrect matches with profiles inferred for families 70 and 80, respectively (Figure 2). When we analyzed the *HMM-dbCAN* database, a high frequency of inconsistencies were recovered for GH families 80 and 134, which resulted in incorrect matches with profiles inferred for families 81 and 135, respectively (Figure 3). We named this type of mismatch Error 1, which accounted for 38% of the mismatches found when using the *HMM-CAZy* database and 41% when using the *HMM-dbCAN* database (Figures 2 to 4). This error is equivalent to a measure of statistical bias because amino acid sequences were consistently assigned to an incorrect HMM profile.

We also noticed another recurrent type of mismatch, named Error 2, which was recovered when amino acid sequences from a given CAZy GH family were assigned, without any consistent

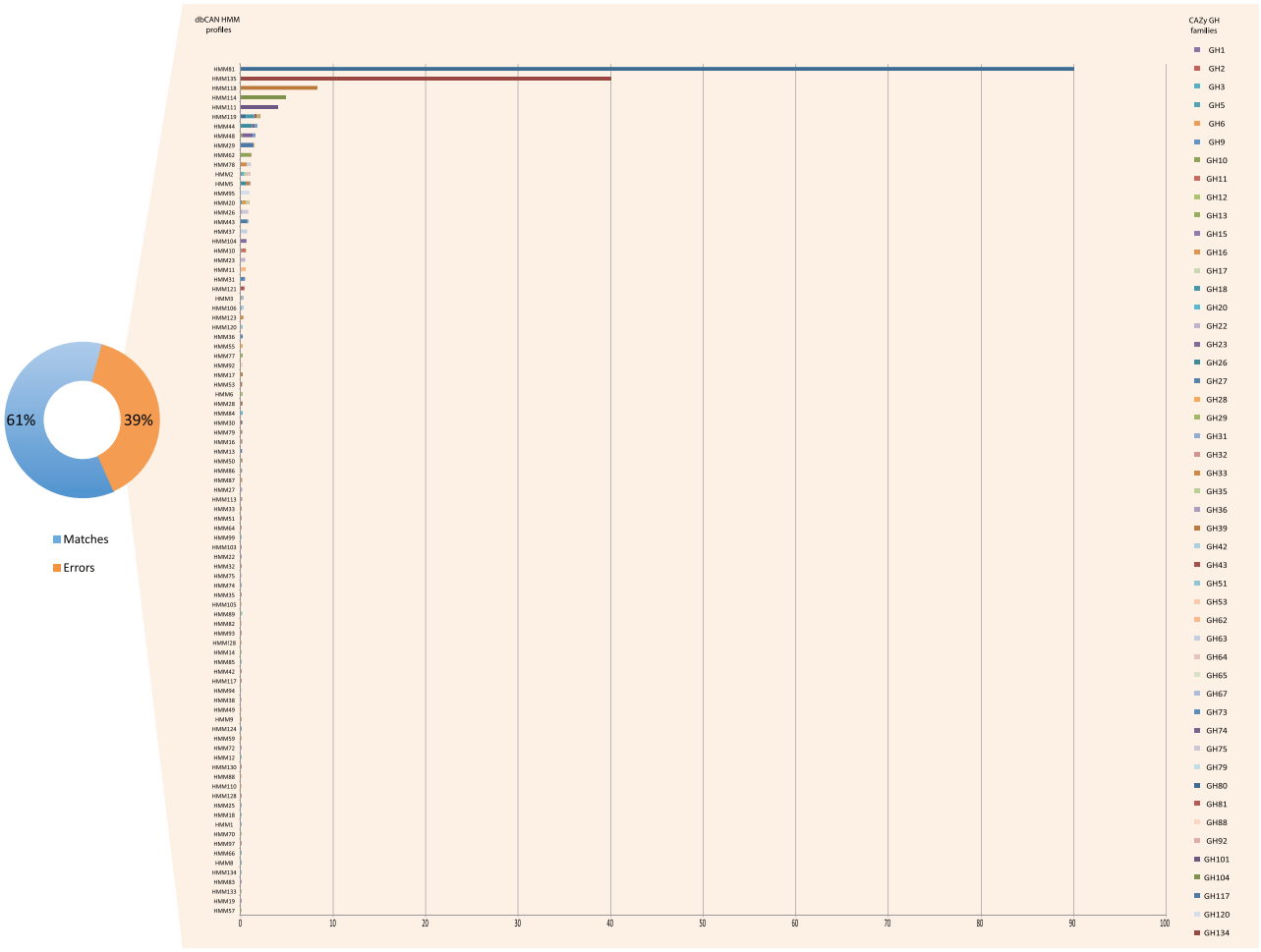

Figure 3. Frequency of errors, as shown in the left pie chart, and discrimination of GH families that presented incongruences between CAZy GH classification and HMM profiles downloaded from the dbCAN database. For each CAZy family, horizontal bars indicate the percentage of mismatches between the CAZy-defined family and the HMM profile from the dbCAN database. GH indicates glycoside hydrolase; HMM, hidden Markov model.

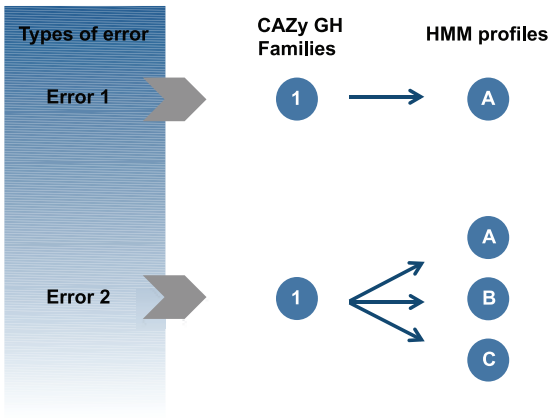

Figure 4. Types of errors found between CAZy GH family classification and HMM profiles. GH indicates glycoside hydrolases; HMM, hidden Markov model.

association, to several HMM profiles that were all different from the CAZy classification for those sequences (Figure 4). The frequency of occurrence of Error 2 was higher when using the *HMM-CAZy* database compared with *HMM-dbCAN* (62% and 59%, respectively) (Figures 2 and 3). The CAZy families with the highest frequency of Error 2 in both HMM databases were families 5, 16, 18, and 43 (Figures 2 and 3). Amino acid sequences from these families all matched at least 9 HMM profiles that were different to their original CAZy classification (Figures 2 and 3).

Mismatches found in our study were possibly caused by the multimodular structure of GH sequences, which are frequently composed of several protein domains.[1] Because we conducted HMM analyses based on complete GH sequences, an impact on the family assignment was expected when using HMM profiles built from CAZy families. This issue was not attenuated when we adopted the *HMM-dbCAN* profile database, in which independent HMM profiles were created using only alignments of GH domains.[26] Estimation of HMM profiles from alignments of complete amino acid sequences resulted in a higher frequency of matches between the HMM search and the CAZy GH classification.

A better performance of our HMM profiles, built from complete sequence alignments, was recovered, despite their lower alignment scores when compared with GH domain sequences downloaded from dbCAN. The low alignment

scores obtained for the HMMs from complete sequences were probably due to the higher percentage of gaps in the alignments (circa 1.6×) in comparison with dbCAN domain alignments. In general, the mismatches observed in both HMM data sets were obtained for GH families with the lowest alignment scores (Supplementary Table S4). We hypothesize that the classification criteria of sequence similarity, adopted by CAZy when defining GH families, generate sequence groups that are not evolutionarily related and, consequently, low-quality alignments are obtained. As dbCAN alignments were inferred from sequence domains only, they presented higher alignment scores and lower gap proportions (average of 2%).[26] Nevertheless, the performance of dbCAN HMM profiles in recovering the CAZy classification was poorer.

We conclude that the automated strategy adopted here; that is, using HMM profiles to recover the standard GH classification implemented in CAZy was effective for most families. Moreover, although mismatches were found, they did not occur randomly, as error types were recurrent in specific families. This important finding means that a focus may thus be directed to these problematic GH families, aimed at obtaining a complete correspondence between CAZy and HMM-based methods. Our classificatory approach using HMM was a useful tool to automatize the CAZy classification of GHs; in principle, sequence assignments to GH families can be made objectively using CAZy-derived HMMs built from complete sequences. This is a significant step toward a higher congruence between the classifications adopted by distinct protein databases, which directly affects future studies on biochemical and evolutionary aspects of GHs.

Acknowledgements

This study is a partial fulfillment of MFR's requirements for a doctoral degree in Genetics at the Federal University of Rio de Janeiro.

Author Contributions

MFR and CGS conceived and designed the experiments and wrote the first draft of the manuscript. MFR, BM, and CGS analyzed the data; contributed to the writing of the manuscript; agreed with manuscript results and conclusions; jointly developed the structure and arguments for the paper; and made critical revisions and approved final version. All authors reviewed and approved the final manuscript.

REFERENCES

1. Lombard V, Ramulu GH, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res*. 2014;42:490–495.
2. André I, Potocki-Veronese G, Barbe S, Remaund-Simeon M. CAZyme discovery and design for sweet dreams. *Curr Opin Chem Biol*. 2014;19:17–24.
3. Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J*. 1991;280:309–316.
4. Henrissat B, Bairoch A. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J*. 1993;293:781–788.
5. Henrissat B, Bairoch A. Updating the sequence-based classification of glycosyl hydrolases. *Biochem J*. 1996;316:695–696.
6. Coutinho PM, Henrissat B. Carbohydrate-active enzymes: an integrated database approach. In: Gilbert HJ, Davies G, Henrissat H, Svensson B, eds. *Recent Advances in Carbohydrate Bioengineering*. Cambridge, UK: The Royal Society of Chemistry; 1999:3–12.
7. Henrissat B, Davies GJ. Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol*. 1997;7:637–644.
8. Davies G, Henrissat B. Structures and mechanisms of glycosyl hydrolases. *Structure*. 1995;3:853–859.
9. Bourne Y, Henrissat B. Glycoside hydrolases and glycosyltransferases: families and functional modules. *Curr Opin Struct Biol*. 2001;11:593–600.
10. Henrissat B, Sulzenbacher G, Bourne Y. Glycosyltransferases, glycoside hydrolases: surprise, surprise! *Curr Opin Struct Biol*. 2008;18:527–533.
11. Pope PB, Denman SE, Jones M, et al. Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proc Natl Acad Sci U S A*. 2010;107:14793–14798.
12. Mahajan C, Chadha BS, Nain L, Kaur A. Evaluation of glycosyl hydrolases from thermophilic fungi for their potential in bioconversion of alkali and biologically treated *Parthenium hysterophorus* weed and rice straw into ethanol. *Bioresour Technol*. 2014;163:300–307.
13. *Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry on the Nomenclature and Classification of Enzyme-Catalysed Reactions*. London, England and New York, NY: Academic Press; 1984.
14. Naumoff DG. Hierarchical classification of glycoside hydrolases. *Biochemistry (Mosc)*. 2011;76:622–635.
15. Davies GJ, Sinnott ML. Sorting the diverse: the sequence-based classification of carbohydrate-active enzymes. *Biochem J*. 2008;30:26–32.
16. Pope PB, Mackenzie AK, Gregor I, et al. Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci. *PLoS ONE*. 2012;7:e38571.
17. Fushinobu S, Alves VD, Coutinho PM. Multiple rewards from a treasure trove of novel glycoside hydrolase and polysaccharide lyase structures: new folds, mechanistic details, and evolutionary relationships. *Curr Opin Struct Biol*. 2013;23:652–659.
18. Chothia C, Lesk AM. The relation between the divergence of sequences and structure in proteins. *EMBO J*. 1986;5:823–826.
19. Gnanavel M, Mehrotra P, Rakshambikai R, Martin J, Srinivasan N, Bhaskara RM. CLAP: a web-server for automatic classification of proteins with special reference to multi-domain proteins. *BMC Bioinformatics*. 2014;15:343.
20. Finn RD, Coggill P, Eberhardt RY, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 2016;44:D279–D285.
21. Altschul S, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–410.
22. Söding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. 2005;21:951–960.
23. Qi M, Wang P, O'Toole N, et al. Snapshot of the eukaryotic gene expression in muskoxen rumen—a metatranscriptomic approach. *PLoS ONE*. 2011;6:1–12.
24. Mimouni N, Lunter G, Deane C. *Hidden Markov Models for Protein Sequence Alignment*. Oxford, UK: University of Oxford; 2004:1–26.
25. Karplus K, Barret C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*. 1998;14:846–856.
26. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2012;40:445–451.
27. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539.
28. Chang JM, Tommaso PD, Notredame C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol Biol Evol*. 2014;31:1625–1637.
29. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39:29–37.