Glycobiology Advance Access published August 9, 2010

doi: 10.1093/glycob/cwq106

CAZymes Analysis Toolkit (CAT): Web-service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database

Byung H. Park [1,*,§], Tatiana V. Karpinets[2,3,§], Mustafa H. Syed[2,§], Michael R. Leuze[1], and Edward C. Uberbacher[2]

1. Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
2. Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
3. Department of Plant Sciences, University of Tennessee, Tennessee, USA

* To whom correspondence should be addressed.

§ These authors contributed equally.

Key words: carbohydrate-active enzymes/ protein families/ computational annotation/ biofuel

Abstract

The Carbohydrate-Active Enzyme (CAZy) database provides a rich set of manually annotated enzymes that degrade, modify, or create glycosidic bonds. Despite rich and invaluable information stored in the database, software tools utilizing this information for annotation of newly sequenced genomes by CAZy families are limited. We have employed two annotation approaches to fill the gap between manually curated high-quality protein sequences collected in the CAZy database and the growing number of other protein sequences produced by genome or metagenome sequencing projects. The first approach is based on a similarity search against the entire non-redundant sequences of the CAZy database. The second approach performs annotation using links or correspondences between the CAZy families and protein family domains. The links were discovered using the association rule learning algorithm applied to sequences from the CAZy database. The approaches complement each other and in combination achieved high specificity and sensitivity when cross-evaluated with the manually curated genomes of *Clostridium thermocellum ATCC 27405* and *Saccharophagus degradans 2-40.* The capability of the proposed framework to predict the function of unknown protein domains (DUF) and of hypothetical proteins in the genome of *Neurospora crassa* is demonstrated. The framework is implemented as a Web service, the CAZymes Analysis Toolkit (CAT), and is available at http://cricket.ornl.gov/cgi-bin/cat.cgi.

Introduction

Carbohydrate-Active Enzymes (CAZymes) participate in many important biological processes including cell wall synthesis, signaling, and energy production. There is a general correlation between the number of glycosidase and glycosyltransferase-related genes and the total number of genes in an organism (Coutinho, P.M., Stam, M., et al. 2003). Plants produce especially many CAZymes. For example, almost 800 genes encoding such enzymes (>3.3% of all genes) were found in the *Arabidopsis thaliana* genome (Coutinho, P.M., Stam, M., et al. 2003). The involvement of CAZymes, glycosidases and glycosyltransferases, in the formation and modification of the wood cell walls makes CAZymes abundant in the woody plants including *Populus* and *Eucalyptus* (Geisler-Lee, J., Geisler, M., et al. 2006, Mellerowicz, E.J. and Sundberg, B. 2008). These enzymes are of special interest for studies aiming to reduce the difficulty in hydrolyzing biomass in bioenergy crops (McCann, M.C. and Carpita, N.C. 2008). CAZymes involved in the hydrolysis of the glycosidic bond, glycoside hydrolases and carbohydrate-binding modules, **a**re important for the degradation of the complex polysaccharides of the plant cell wall by microorganisms producing biofuel (Lynd, L.R., van Zyl, W.H., et al. 2005). Thus, all groups of CAZymes have attracted a great attention of the biofuel research. Complex carbohydrate utilization is also of great interest in medical studies and in studies of plant-microbe interactions. Carbohydrates are abundant in diverse environments of higher eukaryotes including humans and plants, and may serve as carbon sources for beneficial or pathogenic microbes in an organisms (Moyrand,

F., Fontaine, T., et al. 2007). Therefore, predicting genes encoding proteins involved in complex carbohydrates acquisition and utilization is crucial for characterizing bacterial pathogenic potential, anti-bacterial defense, and the potentially beneficial effects of host-microbe interactions.

The Carbohydrate-Active Enzyme (CAZy) database (https//www.cazy.org) (Cantarel, B.L., Coutinho, P.M., et al. 2009) is the most comprehensive repository of CAZymes. The database, through frequent updates, provides rich sets of manually curated information about all groups of CAZymes, including their names, genbank accessions, EC numbers, 3D structure, and taxonomy. This information, when properly utilized, can serve as an invaluable source to identify CAZymes in newly sequenced genomes and classify them. In particular, through an assignment of a CAZy family to a new sequence, the catalytic mechanism of the enzyme, the active site residues and possible substrates can be unveiled. At present, however, CAZy database does not provide a similarity search capability for protein sequences, and furthermore, sequences of interest cannot be directly downloaded. This complicates the use of the database not only for annotation but also for computational analysis of the collected information on CAZymes.

In this study, as an effort to better utilize the CAZy database for systematic annotation, we present the Web-based CAZymes Annotation Tools (CAT). The tools are based on the information collected in the CAZy database, analysis of this information and its supplementation with information from other databases. In addition to common annotation tools based on a sequence similarity search using uni-directional and bi-

directional BLAST, we have proposed and evaluated a novel approach to annotate a sequence with CAZy families. This approach is based on the association rules inferred by 1) linking the annotations of CAZymes in the CAZy database (Cantarel, B.L., Coutinho, P.M., et al. 2009) and their annotation using the Protein Family database (Finn, R.D., Tate, J., et al. 2008), and 2) by further analysis of the data using the "association rule learning" algorithm. A validation of the rules discovered in this study indicated high sensitivity of the rule-based approach for CAZy family annotation of new sequences and its utility for assigning functions to Domains of Unknown Function (DUF) and hypothetical proteins.

The association rule learning employed in the study produces confident and conserved links between different CAZy families and protein family (Pfam) domains within the multi-domain CAZyme architecture. Most CAZymes involved in degradation of complex carbohydrates, particularly those that hydrolyze cellulose and hemicelluloses (Hilden, L. and Johansson, G. 2004), have a complex modular architecture and may include one or two functional domains and several noncatalytic domains known as carbohydrate-binding modules (CBMs) (Boraston A.B., Bolam D.N., et al. 2004). It is believed that the presence of CBMs in the structure of the CAZymes is important to target appended catalytic domains to the polysaccharide; change its conformation, packing and the surface structure (Shoseyov, O., Shani, Z., et al. 2006). In the CAZy database each module is predicted separately by considering its similarity to its pre-computed Hidden Markov Model (HMM), and the same enzyme may be assigned to several families or be annotated by several Pfam domains. How conserved the link is between the CAZy families, or the

co-occurrence of different domains in the structure of CAZymes, is not clear. Using the "association rule learning" algorithm we find many conserved rules underlying the modular CAZyme architecture. We have also developed an interface to query the associations between CAZymes and Pfam domains in terms of different taxonomic groups and enzymes. Evaluation of the proposed annotation tools, which are based on the similarity search and on the association rules between CAZy families and Pfam domains, have been performed using two genomes, *Clostridium thermocellum* ATCC 27405 and *Saccharophagus degradans* 2-40. Applicability of CAT is further illustrated by annotation of recently updated genomes of *Populus trichocarpa*, *Escherichia Coli*, and *Neurospora crassa*. CAT is accessible as a part of the Bioenergy Science Center (BESC) public portal at http://cricket.ornl.gov/cgi-bin/cat.cgi.

Results

Reference set of sequences from CAZy

Processing of the CAZy HTML Web pages (as of September 30, 2009) as described in the Method section produced 167,469 records of annotated CAZymes, among which 167,401 were found to have Genbank accessions. A further removal of redundant sequences (proteins with resubmissions), keeping only the latest submissions, leaves 109,744 unique sequences. This final set is used for the similarity search, analysis of CAZymes, and the extraction of association rules between Pfam domains and CAZymes.

Assignment of CAZy family to a sequence using similarity search against the CAZy database and the Protein families database

We have implemented and evaluated two approaches for the annotation of a sequence or a set of sequences with CAZy families. The first approach, using similarity search, is based on the BLAST search, either uni-directional or bi-directional. This algorithm directly transfers the CAZy family annotation of the most similar sequence in the reference set to a query sequence. The second approach assigns a CAZy family to a protein sequence indirectly. The assignment is based on annotation of the sequence with Pfam domains and then through links between Pfam domains and CAZy families. The Pfam protein families database (Finn, R.D., Tate, J., et al. 2008) is a comprehensive, regularly updated and manually curated collection of protein domains and families, represented as multiple sequence alignments, and as HMM or Position Specific Scoring Matrices (PSSM). Some CAZy families in the CAZy database have one to one correspondence with a Pfam model. Therefore, Pfam models, when they are found in a sequence, can be indirectly used for assignment of the CAZy family to the sequence. Potentially, the high quality of the models available in the Pfam protein families database can provide a high specificity and sensitivity for CAZy family assignments. However, at the current time, only 142 CAZy families out of 290 (less than 50%) are known to have a corresponding Pfam domain in the CAZy database. The limited number of the available links essentially decreases specificity and sensitivity of the CAZy family annotation using the Pfam protein family database. According to our analysis, only ~60 % of sequences annotated by a CBM family in the CAZyreference set are found using the

links. The sensitivity can be increased by loosening the threshold for Pfam domain search, but this significantly decreases the specificity of the CAZy family assignment.

We have addressed the problem of the poor linkage between the CAZy and the Pfam databases by revealing a comprehensive set of highly confident "association rules" between CAZy families and Pfam domains. Each rule in the set suggests a correspondence between CAZy families and Pfam domains (Supplementary Table 1). The CAZy reference set of sequences was used as the target of this learning algorithm, which is described in the Methods section in more detail.

The analysis produced 1,052 rules with the level of support from 5 to 13,442 and with the level of confidence >= 0.8. After removal of redundant rules (those producing the same annotations), we generated 104 non-redundant rules with support >= 40 and confidence >= 0.84 (Supplementary Table 1). These rules are used to assign CAZy families to a new sequence given the Pfam annotations of the sequence. The confidence and support of a rule denote an expected accuracy of the rule and the number of sequences that satisfy the rules (for details, please refer the Method section).

Evaluation of the of CAZy family assignments

Evaluation of the proposed annotation methods was conducted with two test genomes: *Clostridium thermocellum ATCC 27405* and *Saccharophagus degradans 2-40*, which we abbreviate as CTHE and SD2.40, respectively. CTHE includes 3,191 proteins among

which 179 proteins are annotated by CAZy. SD2.40 includes 4,007 proteins among which 315 proteins are annotated by CAZy. The annotation algorithms are evaluated at different parameter values that specify the sensitivity of search. For the similarity search, the "bit-score" of the BLAST search is varied from 50 to 300. For the association rule method, rules with support >= 10 observations are used with bit-score range from 50 to 200 for Pfam search (Note Pfam domains should be searched first to apply rules). No significant difference was found between the outputs from uni-directional and bi-directional similarity search of each organism, but the sensitivity and the specificity of the CAZy family assignment were very different between the two search algorithms and the Pfam based algorithm. Specificity and sensitivity plots for each algorithm are shown in Figure 1 and 2. According to Figure 1, the similarity search algorithm is very accurate. Its specificity is close to 100%, i.e. it does not produce false CAZy family assignments. The search, however, is not very sensitive. Its sensitivity is about 60%. It means that more than 1/3 of the CAZymes were not identified by the similarity search. The Pfam based annotation (Fig. 2) is more sensitive and identifies more than 80% of CAZymes in the genome, but it is not as accurate as the similarity search. Thus, the combination of the Pfam based annotation with the direct similarity search may provide an optimal balance of specificity and sensitivity for CAZy family assignment.

CAZymes Analysis Toolkit (CAT)

We have developed a Web-based interface 1) for an automatic assignment of CAZy families to new sequences using the proposed algorithms and 2) for an advanced search

of the CAZy reference sequences. A user can submit as an input either a sequence or a set of sequences (typically a whole genome). Using the interface provided in CAT, a user can select the annotation algorithm and a threshold value related to the sensitivity level of the algorithm. Optionally, the user can request a composite output that compares Pfam domain structures and the protein lengths of the input sequences and its CAZy hit sequence side by side. For large submissions a user is advised to provide an e-mail address. In this case a link to the Web page, where the results are stored, is sent to the user when the processing is completed. Results are stored for 2 days after which an access will be no longer valid.

CAT provides a diverse set of search options. "CAZymes Search" allows users to browse the CAZy reference set by enzyme, organism, taxonomic group, and CAZy families or any combination of these. For example, a specific product name or activity such as "xylanase" or "endo-1,4-beta glucanase" can be used to specify a search. For a broader search, like "Eukaryotae", the search will return a list of Eukaryotic sequences in the reference set. The search results are supplemented with the product name, taxonomic information downloaded from Genbank and by Pfam domain annotation produced as described in Methods.

"Associations search" in CAT provide a way to search the CAZy reference set in terms of associations between CAZy families. A single protein may be annotated with several CAZy families. Carbohydrate binding domain "CBM48", for example, is often found with glycoside hydrolase "GH13" in a number of bacterial proteins. CAT provides an

interface to search for such associations between different CAZy families. A search can be made by an enzyme, a CAZy family, an organism, a superkingdom, or a taxonomy. The result of a search is a list of associations, the number of proteins (linked to the list of proteins) in CAZy database, and association rules derived from proteins showing these associations.

Examples of CAT applications

CAT is employed to predict CAZymes in recent submissions of the *Populus trichocarpa* genome and in the *Escherichia coli K12-MG1655* genome.

A Supplementary Table 1 provides an example from a similarity search of a recent submission of the *P. trichocarpa* genome (Jul 30 2009) against our CAZy reference set using uni-directional BLAST with the recommended thresholds. The uni-directional BLAST search returned 7,207 hits representing 4671 unique proteins in the poplar genome. 1,979 hits from this list (984 unique proteins) also had consistent protein length and domain architecture with the CAZy reference sequences.

An annotation of the *E. coli* genome (Accession version NC_000913, Date: 30-JUL-2009) using the inferred association rules is provided in Supplementary Table 2. The algorithm predicted 114 CAZymes in the genome. Only 67 of them could be found in CAZy. Forty seven proteins predicted by the algorithm were not previously annotated by a CAZy family. To validate the annotation we compared EC-numbers and product

annotation predicted by CAZy with annotation of the proteins available in KEGG, EcoCyc and GenBank (Supplementary Table 3). Manual comparison of the annotation confirmed 28 CAZy family predictions for these proteins.

Predicting functions for unknown protein domains

CAT was also applied to predict functional activity of protein Domain of Unknown Function. In some cases the discovered confident associations between CAZy families and Pfam domains directly link the unknown domains to a specific CAZy family. Such associations allowed us to assign functional activity, taxonomic groups and CAZy families to several unknown domains including DUF2029, DUF297, DUF303, DUF847 and several others that are listed and characterized in Supplementary Table 8. For example, DUF2029, which is known as a putative transmembrane protein from bacteria, is linked to CAZy family GT87. All sequences except one (104 records) annotated in CAZy by family GT8 have DUF2029 in their Pfam domain architecture (the sequential order of protein domains). Figure 3 a, which is generated using CAT, demonstrates that the domain is located in the center of the enzyme and accounts for more than 50% of its length. According to a description of the family, the domain likely represents alpha α-1,2-mannosyltransferase (EC 2.4.1.-), an enzyme that uses polyprenol-P-mannose as the sugar donor. A taxonomic characterization of the sequences annotated by DUF2029 reveals that they all belong to bacteria. Another hypothetical domain, DUF297, was linked to GH114 (Figure 3, b), and likely associates with bacteria or fungal endo-α-1,4-polygalactosaminidase (EC 3.2.1.109). Another interesting association between an

unknown Pfam domain and a CAZy family involves DUF847. The domain is found in 162 out of 165 sequences annotated in CAZy by glycoside hydrolase family 108 (Supplementary Table_8). The Pfam domain architecture of the sequences (Figure 3, c) demonstrates that the domain is located in their N-termini and as a rule coincides with a C-terminal domain PG_binding_3, which is a potential peptidoglycan binding domain according to the Pfam description. This domain architecture clearly attributes the enzymatic activity of GH108, N-acetylmuramidase (EC 3.2.1.17), to DUF847 and the binding activity of the enzyme to PG_binding_3. DUF303 associates with the family 6 of carbohydrate esterases (CE6) and likely represents acetyl xylan esterase (EC 3.1.1.72). Forty one out of forty three carbohydrate esterases have DUF303 in their architecture. In most of these enzymes (30 out of 43), DUF303 is the only domain that constitutes the Pfam architecture and occupies more than one third of the enzymes length (Figure 3 d). This information provides additional evidence to the link between CE6 and DUF303.

Predicting functions for hypothetical proteins

The applicability of CAT to predict function of hypothetical proteins is demonstrated by annotation of the *Neurospora crassa* genome using CAT. *Neurospora crassa* is a filamentous fungus used as a model laboratory organism that can degrade cellulose. A recent analysis of plant cell wall and cellulose degradation by this organism using a shotgun proteomics approach (Tian, C.G., Beeson, W.T., et al. 2009).(A recent work of Beeson *et al.* [reference needed] has identified 50 secreted proteins that very likely represent CAZymes, from which only 26 proteins were computationally predicted as

CAZymes. Our analysis of the secreted proteins using CAT tools predicted 7 additional CAZymes previously annotated as hypothetical (Supplementary Table 5).

Discovering conserved associations between protein domains within the modular structure of CAZymes

With the inferred association rules, CAT discovers regularities in the domain architecture of the CAZymes. Most enzymes involved in degradation of complex carbohydrates, particularly those that hydrolyze cellulose and hemicellulose, have a complex modular architecture and may include 1-2 functional domains and several noncatalytic domains known as carbohydrate-binding modules (CBMs). In CAZy database we have found 634 enzymes annotated by 2 families representing catalytic activities and 460 enzymes with 2 different CBM domains. Supplementary Table 6 provides 33 examples from the CAZy database of enzymes annotated by 3 different catalytic activities and by 3 different binding activities. How conserved the link is between the CAZy families, or the degree of co-occurrence of different domains in the structure of CAZymes, has not been clearly understood. CAT answers this question by identifying many conserved rules underlying the modular CAZyme architectures. Supplementary table (Suppl_Association_Taxa_Product.xls) provides a list of 470 associations between families found in CAZy. The Web interface to the table available in CAT can be employed to reveal some conserved associations between CAZy families and to identify taxonomic groups where the associations can be found. For example, the association of GH72 with CBM43 found in 1,3-beta-glucanosyltransferase is typical for fungi. The

association of GH19 and CBM18 is typical for chitinases in plants. But the association GH18 and CBM5 can be found only in bacterial chitinases (Suppplementary Table 7).

Discussion

The proposed analytical tools have several potential applications in studies of carbohydrate active enzymes. One of the basic analyses of new sequences implemented by CAT and not available for researches before is the BLAST search of a sequence or sequences against the CAZy database. The BLAST search is implemented by many databases including Genbank (Benson, D.A., Karsch-Mizrachi, I., et al. 2009) and KEGG (Moriya, Y., Itoh, M., et al. 2007) and provides the basis for functional annotation of new proteins by transferring an annotation from well-annotated sequences in the database to similar sequences. This annotation, however, has certain pitfalls. It was shown that functionality of enzymes diverges quickly when their sequence identity is less than 70% (Tian, W.D. and Skolnick, J. 2003). But even 70% identity can produce 10% false positive annotation. More accurate computational prediction is produced by using the bi-directional best hit by comparison of two genomes using the BLAST program. This approach was used to infer clusters of orthologous groups (Tatusov, R.L., Koonin, E.V., et al. 1997) and to assign the KEGG orthology identifiers for proteins in new genomes (Moriya, Y., Itoh, M., et al. 2007). Both types of the search, regular and bi-directional, are available in CAT, and the results are supplemented by annotation of sequences with Pfam domains.

Although both similarity search algorithms are very accurate when they are used with a recommended bit-score threshold, they are not sensitive. About 40% of CAZymes can be missed in the genome annotation by the similarity search along. The low sensitivity of the algorithm is likely because the search is focused on the global sequence similarity. Many CAZymes, however, have a complex modular architecture. And only one of the modules, that has a specific structure and function, is responsible for the CAZy family assignment (Cantarel, B.L., Coutinho, P.M., et al. 2009). Thus, another approach to annotation of a sequence by a CAZy family is based on the Pfam search and links between Pfam domains and CAZy families inferred by the association rules. The approach is more sensitive than the similarity search and can predict more than 80% of CAZymes with specificity 80-90%. In this study we infer links between Pfam domains and CAZy families for the Pfam based annotation using a novel approach based on the association rules learning (Agrawal, R. and Shafer, J.C. 1996, Sriphaew, K. and Theeramunkong, T. 2004). This data mining technique finds causal relations among a set of variables in large datasets and has been applied for analysis of some biological datasets mainly produced by microarrays (Martinez, R., Pasquier, N., et al. 2008, Nam, H., Lee, K., et al. 2009). An association rule identifies two groups of variables (in our case they are CAZy families and Pfam domains) where an occurrence of one implies an occurrence of the other. Such a rule is produced by examining co-occurrence patterns of variables in a large dataset, which is the CAZy reference set of sequences in our case. Although the technique gives statistical measures to characterize the scope (support) and the precision (confidence) of a rule, there are no widely accepted approaches to find thresholds for the measures. By changing the thresholds we can change the number of inferred rules and, in our case, the

number of links between a CAZy family and a Pfam domain. In this study we have identified thresholds for the level of support and confidence by comparing all the links found between CAZy families and Pfam domains by the algorithm with the set of rules given in the CAZy database. By using the set of CAZy rules as a benchmark, we have chosen the minimum level of support produced by a CAZy rule and consider this level as a threshold. Only rules with the level of support that is greater that the threshold were used for validation of the Pfam based annotation. The threshold identified by this approach is rather stringent and can limit a sensitivity of the Pfam based CAZy family assignment. By reducing the threshold we can increase sensitivity of the prediction, but, at the same time, we will also increase the number of false positive annotations. Considering this limitation we have provided in CAT an option to input a threshold for the level of support and, thus, to try different balance between specificity and sensitivity of the Pfam based predictions.

Some links between CAZy families and Pfam domains inferred in the study include Pfam domains of unknown function or DUF. In the paper we have provided examples how these links can propose a functional activity and taxonomic groups to unknown domains (Figure 3, Supplementary table_DUF). Most of such rules have a high level of support, and high percentage of sequences annotated in CAZy by the family have the associated DUF along or in an association with another domain. In addition, CAT also provides an option to align sequences and to visualize their Pfam architecture (Figure 3) for all or selected sequences annotated by the family or/and by the associated DUF. Although the visualization in combination with the computational prediction is helpful in proposing a

functional activity for unknown domain, this activity must be further evaluated experimentally or using already published studies. For example, the predicted link between DUF1083 and CBM9 (Supplementary Table_DUF) can be confirmed by studies of glycoside hydrolase family 10 xylanase from Thermotoga maritime (Boraston, A.B., Creagh, A.L., et al. 2001, Verjans, P., Dornez, E., et al. 2010).

Availability of the Pfam annotation and the protein sequences in the CAZy reference dataset create an opportunity for researches to engage a number of sequence based or Pfam based computational tools to study CAZymes, like PfamAlyzer (Hollich, V. and Sonnhammer, E.L.L. 2007), CDART (Geer, L.Y., Domrachev, M., et al. 2002), ArchSchema (Tamuri, A.U. and Laskowski, R.A. 2010), and Pfam domain architectures of proteins from the UniProt database (Bairoch, A., Consortium, U., et al. 2009). Such tools can be very helpful in finding new CAZymes or their new architecture. For example, the domain DUF303 was linked in our study to acetyl xylan esterase (EC 3.1.1.72) annotated in CAZy as CE6. A search for this domain in CAT returns 41 annotated enzymes from this family. But with a click on the domain name in any of these sequences, CAT redirects the query to Pfam database and extracts all architectures with DUF303 of proteins available in the UniProt database. This search returns 354 sequences with only DUF303 in the architecture and 141 sequences with a combination of DUF1737 (a small domain at the N-terminus), DUF303 (in the middle), and PB000458 (at the C-terminus), which is an automatically generated protein family. Extracting the sequences and their annotations using the Pfam interface, we can see that they belong to putative uncharacterized bacterial proteins from UniProt. Considering the association of DUF303 with CE6, we can suggest that these proteins may represent a new architecture

or even family of CAZymes associated with unknown protein domains DUF1737 and PB000458.

In conclusion, analysis and processing of the data downloaded from the CAZy database allowed us to develop a set of tools for CAZymes annotation, search and query. Using the association rule learning algorithm we have discovered many conserved rules between CAZy families and Pfam domains, which are very powerful for assigning new proteins to CAZy families. To the best of our knowledge, this is the first study that has systematically explored rules underlying domain architecture of the carbohydrate active enzymes.

Methods

Obtaining data from the CAZy database

The information content of each CAZy family is directly extracted from its HTML pages, and populated into the local database. The HTML pages obtained through a GET request for a family are parsed to associate the family with the Genbank accession number, related CAZy families, known activities, EC numbers, and available cross references to other databases including UniProt (Consortium, T.U. 2009) Pfam (Finn, R.D., Tate, J., et al. 2008) and PDB (Kirchmair, J., Markt, P., et al. 2008). The latest download was made on 30 Sep 2009. The local database was built using MySQL, and Perl scripts developed in house are used to create and update the database.

The protein sequences of genes in CAZy database were downloaded in batch through the eUtils Web-service of Genbank (http://eutils.ncbi.nlm.nih.gov) and stored in FASTA format with Genbank accession as IDs. The sequences are further processed to remove redundancies by keeping only the latest submission each sequence to Genbank. This set of unique sequences is used as the reference set for the similarity search and will be further referred as the CAZy reference set.

Similarity search algorithms

The search algorithm compares a query protein sequence or a set of sequences with the CAZy reference set of sequences using either uni-directional or bi-directional BLASTp program (Altschul, S.F., Madden, T.L., et al. 1997) with specified parameters. A BLAST search compares a query sequence with the target sequences, producing a list of similar sequences. With uni-directional approach, sequence S_t with the best bit-score is selected as the best hit target sequence for query sequence S_q (Note S_t can be a set of sequences). With bi-directional approach, S_q should also be the best hit target sequence for S_t. Thus a bi-directional BLAST search is typically applied when the best reciprocally similar pairs between the query and target sets are desired. The query and the reference sequence in such a pair are considered as an orthologous pair. The pair will be given as output of the BBH algorithm. Both algorithms were implemented using Perl script.

Protein family annotation of the protein sequences

Seed alignments from the Pfam database of alignments and PSSM are used to annotate the query sequences and the set of reference sequences with protein families. The search is done through the Reverse Position Specific BLAST (RPS-BLAST) (Altschul, S.F., Madden, T.L., et al. 1997) with a set of default parameter values and the e-value of 0.1.

For a given query sequence, the protein family annotation suggests a set of Pfam domains and their location. The identified domains are concatenated according to their starting locations in the sequence. The associations between Pfam domains and CAZy families, which are available in the CAZy database, or inferred in the study, are used to assign CAZy families to the query sequence.

Finding conserved associations between CAZy families and Pfam domains using the association rules

An association rule denotes an interesting relation between variables in a large data (Agrawal, R. and Shafer, J.C. 1996, Sriphaew, K. and Theeramunkong, T. 2004). A rule has an implication form of $A \Rightarrow B$, which reads "*the presence of A implies the presence of B.*" A rule is assessed by *support, confidence, and lift.* The support of a rule $A \Rightarrow B$ is simply the number of instances (or records) in the data in which both *A* and *B* appear. While support is a quantitative measure of a rule, confidence is a qualitative measure of a rule. It illustrates to which degree a rule is true. For example, a rule $A \Rightarrow B$ is 95% confident if *B* occurs in 95% of the records in which *A* occurs. The lift of a rule denotes

the ratio of the observed support of a rule to the support expected by chance. In general, a rule with high support and confidence is favored. In this study, we restrict association rules to bear relations between Pfam domains and CAZy families. More specifically, for any rule $A \Rightarrow B$, A is a set of Pfam domains and B is a CAZy family. Thus, for a query sequence, if its Pfam annotation includes all the domains in A, it is assigned CAZy family B. In this study,

Two sets of association rules are created: one that is readily available in CAZy (CAZy database already assigns each CAZy family with Pfam domains) and an additional set of rules inferred from the Pfam annotations and CAZy family assignments of the entire protein set in CAZy. In particular, for each of 109,744 sequences in the CAZy database, its Pfam annotation together with the CAZy family assignments are created. These lists of augmented annotations, which are often called transaction records in data mining community, are then analyzed to infer the rules described above.

Evaluation of the CAZy family assignment

CAZy family annotations of two genomes, *Clostridium thermocellum* ATCC 27405 and *Saccharophagus degradans* 2-40, from the CAZy database are used to evaluate the proposed methods. Prior to the evaluation, all protein sequences of each genome are removed from the reference data, and the CAZy family assignments are made from the rest of the sequences. The performance is measured based on the number of correct assignments (true positive), the number of incorrect assignments (false positive), and the

number of missing assignments (false negative) for sequences in each organism. True positives are defined as the number of sequences of which CAZy family predictions exactly match their true annotations. For example, if a sequence is predicted with families *A* and *B,* and the true annotation is also *A* and *B.* it is considered a true positive. For all other cases, such as *A* and *C,* are considered a false positive. False negatives are the cases when the algorithm cannot give any predictions for sequences that in fact have annotations. Sensitivity and specificity are measured at different levels of bit scores. For Pfam based annotation, four different support levels (10, 15, 20, and 25) are tested. The similarity search algorithm is tested only for the case of bi-directional best hit.

Acknowledgments

This research was supported by the BioEnergy Science Center and by the Genomic Science Program of the Office of Biological and Environmental Research, U.S. Department of Energy (DOE).

Reference

Agrawal, R. and Shafer, J.C. (1996) Parallel mining of association rules. IEEE Transactions on Knowledge and Data Engineering, 8, 962-969.

Altschul, S.F., Madden, T.L., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25, 3389-3402.

Consortium, U., et al. (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Research, 37, D169-D174.

Benson, D.A., Karsch-Mizrachi, I., et al. (2009) GenBank. Nucleic Acids Research, 37, D26-31.

Boraston A.B., Bolam D.N., et al. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem J 2004. 382,769-781

Boraston, A.B., Creagh, A.L., et al. (2001) Binding specificity and thermodynamics of a family 9 carbohydrate-binding module from Thermotoga maritima xylanase 10A. Biochemistry, 40, 6240-6247.

Cantarel, B.L., Coutinho, P.M., et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Research, 37, D233-238.

The UniProt Consortium, T.U. (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Research, 37, D169-174.

Coutinho, P.M., Stam, M., et al. (2003) Why are there so many carbohydrate-active enzyme-related genes in plants? Trends Plant Sci, 8, 563-565.

Finn, R.D., Tate, J., et al. (2008) The Pfam protein families database. Nucleic Acids Research, 36, D281-288.

Geer, L.Y., Domrachev, M., et al. (2002) CDART: Protein homology by domain architecture. Genome Research, 12, 1619-1623.

Geisler-Lee, J., Geisler, M., et al. (2006) Poplar carbohydrate-active enzymes. Gene identification and expression analyses. Plant Physiology, 140, 946-962.

Hilden, L. and Johansson, G. (2004) Recent developments on cellulases and carbohydrate-binding modules with cellulose affinity. Biotechnology Letters, 26, 1683-1693.

Hollich, V. and Sonnhammer, E.L.L. (2007) PfamAlyzer: domain-centric homology search. Bioinformatics, 23, 3382-3383.

Kirchmair, J., Markt, P., et al. (2008) The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery. J Med Chem, 51, 7021-7040.

Lynd, L.R., van Zyl, W.H., et al. (2005) Consolidated bioprocessing of cellulosic biomass: an update. Curr Opin Biotechnol, 16, 577-583.

Martinez, R., Pasquier, N., et al. (2008) GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. Bioinformatics, 24, 2643-2644.

McCann, M.C. and Carpita, N.C. (2008) Designing the deconstruction of plant cell walls. Curr Opin Plant Biol, 11, 314-320.

Mellerowicz, E.J. and Sundberg, B. (2008) Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. Curr Opin Plant Biol, 11, 293-300.

Moriya, Y., Itoh, M., et al. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research, 35, W182-185.

Moyrand, F., Fontaine, T., et al. (2007) Systematic capsule gene disruption reveals the central role of galactose metabolism on Cryptococcus neoformans virulence. Mol Microbiol, 64, 771-781.

Nam, H., Lee, K., et al. (2009) Identification of temporal association rules from time-series microarray data sets. Bmc Bioinformatics, 10, -.

Shoseyov, O., Shani, Z., et al. (2006) Carbohydrate binding modules: biochemical properties and novel applications. Microbiol Mol Biol Rev, 70, 283-295.

Sriphaew, K. and Theeramunkong, T. (2004) Fast algorithms for mining generalized frequent patterns of generalized association rules. Ieice Transactions on Information and Systems, E87d, 761-770.

Tamuri, A.U. and Laskowski, R.A. (2010) ArchSchema: a tool for interactive graphing of related Pfam do-main architectures. Bioinformatics.

Tatusov, R.L., Koonin, E.V., et al. (1997) A genomic perspective on protein families. Science, 278, 631-637.

Tian, C.G., Beeson, W.T., et al. (2009) Systems analysis of plant cell wall degradation by the model filamentous fungus Neurospora crassa. Proceedings of the National Academy of Sciences of the United States of America, 106, 22157-22162.

Tian, W.D. and Skolnick, J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? Journal of Molecular Biology, 333, 863-882.

Verjans, P., Dornez, E., et al. (2010) Truncated derivatives of a multidomain thermophilic glycoside hydrolase family 10 xylanase from Thermotoga maritima reveal structure related activity profiles and substrate hydrolysis patterns. Journal of Biotechnology, 145, 160-167.

Figures

Figure. 1 Evaluations of the similarity search algorithm using CAZy annotation of *Clostridium thermocellum ATCC 27405* (left) and *Saccharophagus degradans 2-40* (right) as benchmarks. In each case, Circular and triangular shaped plots denote specificites with and without additional evaluation of Pfam domains in the compared proteins, respectively. Likewise, rectangular and diamond shaped plots denote sensitivites with and without Pfam domains evaluation, respectively. For the analyzed organisms, no difference in the outputs was found between uni-directional and bi-directional blast search.

Figure. 2 Evaluation of the CAZy family annotations based on the association rules between CAZy families and Pfam domains using CAZy annotation of *Clostridium thermocellum ATCC 27405* (left) and *Saccharophagus degradans 2-40* (right) as benchmarks. In each case, diamond and rectangular shaped plots denote specificity and sensitivity, respectively. In each case, rules with the minimum support of 10 are used.

Figure 3. Visualization of the domain architecture in CAT for selected Glycosyltransferases family 8 (a), glycoside hydrolases family 114 (b), glycoside hydrolases family 108 (c), and carbohydrate esterases family 6 (d) confirming links of the CAZy families with domains of unknown function.

Table I Non-redundant association rules between CAZy families and Pfam domains inferred from the CAZy reference dataset

Pfam	CAZy family	Support[a]	Confidence[b]	Lift[c]	Number of sequences[d]	Support in family[e]
Glycos_transf_2	GT2	10623	0.99	8	13726	77.39
Glycos_transf_1	GT4	9296	0.87	9	10307	90.19
SLT	GH23	2690	1	35	3357	80.13
Transpeptidase	GT51	2446	1	34	2893	84.55
Glyco_transf_5	GT5	1809	1	44	2897	62.44
CBM_48	CBM48	1190	0.99	60	1746	68.16
CBM_48	GH13	1180	0.98	19	5367	21.99
Glyco_hydro_2_N	GH2	1146	0.99	81	1339	85.59
Esterase	CE1	1094	1	74	1426	76.72
Glyco_hydro_3_C	GH3	997	1	55	1802	55.33
Alpha-amylase_C	GH13	958	1	19	5367	17.85
Chitin_synth_2	GT2	936	1	8	13726	6.82
Glyco_hydro_2_C	GH2	906	0.99	81	1339	67.66
Chitin_synth_1	GT2	849	1	8	13726	6.19
Glyco_hydro_43	GH43	703	1	138	755	93.11
Cellulose_synt	GT2	642	1	8	13726	4.68
Glycos_transf_N	GT30	578	1	170	581	99.48
Amidohydro_1	CE9	556	1	177	862	64.5
Glyco_hydro_32C	GH32	504	1	118	841	59.93
3D	GH102	387	0.98	248	390	99.23

Bgal_small_N	GH2	357	1	81	1339	26.66
Glyco_hydro_38C	GH38	322	1	275	362	88.95
Glyco_hydro_65N	GH65	311	1	299	329	94.53
Glyco_hydro_92	GH92	277	1	355	278	99.64
XET_C	GH16	260	1	104	1091	23.83
SBP_bac_3	GH23	238	1	35	3357	7.09
Alpha-L-AF_C	GH51	230	1	418	271	84.87
Abhydrolase_3	CE10	227	0.99	331	297	76.43
Rod-binding	GH73	219	1	134	967	22.65
BiPBP_C	GT51	219	1	34	2893	7.57
Alpha-mann_mid	GH38	218	1	275	362	60.22
Glyco_hydro_42M	GH42	214	0.97	437	218	98.17
PG_binding_1	GH103	207	0.81	370	672	30.8
MLTD_N	GH23	198	1	35	3357	5.9
Esterase_phd	CE1	196	1	74	1426	13.74
Chitin_bind_1	GH19	178	0.95	113	943	18.88
DUF1205	GT1	170	1	57	2154	7.89
Alpha-amylase_N	GH13	170	0.99	19	5367	3.17
Gly_transf_sug	GT32	165	0.85	340	303	54.46
DUF847	GH108	162	1	603	165	98.18
Trehalose_PPase	GT20	157	0.84	126	654	24.01
Branch	GT14	154	1	638	172	89.53
COesterase	CE10	154	0.99	330	297	51.85
GT36_AF	GH94	153	0.99	503	234	65.38
MGDG_synth	GT28	151	1	86	1158	13.04
Alg14	GT1	144	0.99	56	2154	6.69
Lyase_8_N	PL8	143	1	575	173	82.66
DUF1972	GT4	142	1	10	10307	1.38

Sucrose_synth	GT4	132	1	10	10307	1.28
Glyco_hydro_20b	GH20	131	0.99	181	566	23.14
DUF1975	GT4	130	1	10	10307	1.26
PilZ	GT2	130	1	8	13726	0.95
X8	CBM43	128	0.98	658	149	85.91
Glyco_transf_8N	GT8	126	1	121	848	14.86
DUF1957	GH57	125	1	256	386	32.38
DUF2029	GT87	124	1	793	125	99.2
Glyco_hydro_65C	GH65	121	1	299	329	36.78
Bac_rhamnosid_N	GH78	117	1	540	185	63.24
PG_binding_3	GH108	115	1	603	165	69.7
Glyco_hydro_53	GH53	112	0.99	870	112	100
ChitinaseA_N	GH18	110	0.99	53	1997	5.51
PUD	CBM41	106	1	768	128	82.81
Fuc4NAc_transf	GT56	105	1	936	105	100
Glyco_hydro_97	GH97	104	1	945	104	100
CBM_X	GH94	103	0.98	497	234	44.02
PUD	GH13	102	0.96	19	5367	1.9
AXE1	CE7	97	0.94	954	99	97.98
MIR	GT39	97	1	568	203	47.78
Glyco_transf_36	GH94	97	1	507	234	41.45
Caps_synth	GT32	92	1	399	303	30.36
PUD	CBM48	92	0.87	52	1746	5.27
Chitin_synth_1N	GT2	89	1	8	13726	0.65
Lyase_8_C	PL8	88	1	575	173	50.87
AMPKBI	CBM48	88	1	60	1746	5.04
GtrA	GT2	88	1	8	13726	0.64
NodZ	GT23	86	1	1143	118	72.88

Glycogen_syn	GT3	82	1	1198	85	96.47
CHB_HEX	GH20	82	1	182	566	14.49
Alginate_lyase2	PL7	76	0.96	1244	76	100
NAGidase	GH84	76	1	1293	76	100
Gal_Lectin	GH35	75	1	271	377	19.89
Glyco_hydro_42C	GH42	74	1	451	218	33.94
Glyco_hydro_67M	GH67	72	1	1365	72	100
Glyco_hydro_67C	GH67	71	1	1365	72	98.61
CST-I	GT42	70	1	1404	70	100
Raffinose_syn	GH36	68	1	287	362	18.78
Pec_lyase	PL10	65	1	1512	65	100
CelD_N	GH9	64	1	261	412	15.53
Glyco_transf_52	GT52	63	1	1560	63	100
Glucodextran_N	GH15	63	1	172	587	10.73
CHB_HEX_C	GH20	62	1	182	566	10.95
GFO_IDH_MocA	GH109	61	0.98	1585	65	93.85
S6PP	GT4	61	0.9	9	10307	0.59
DUF1083	CBM9	60	1	1536	78	76.92
Mannosyl_trans2	GT76	58	1	1694	59	98.31
DUF563	GT61	57	1	1724	130	43.85
PAE	CE13	56	1	1755	60	93.33
DUF821	GT90	55	1	1755	169	32.54
PMEI	CE8	55	1	293	338	16.27
Sialidase	GH33	55	0.98	1322	417	13.19
CgtA	GT82	53	1	1854	53	100
RGP	GT75	49	1	2006	50	98
DUF303	CE6	41	1	2340	43	95.35
Hyaluronidase_1	PL16	40	1	2457	40	100

[a] Support of a rule is the number of sequences in the CAZy reference dataset in which both Pfam domain and the associated CAZy family in the rule appear.

[b] Confidence is an estimated probability that a sequence belongs to a CAZy family specified by the rule if the sequence has the Pfam domain specified by the rule.

[c] Lift is a ratio of the observed support of the rule to the support expected by chance. For example, if lift is 50, it means that by chance we would find 50 times less sequences annotated by the Pfam domain and CAZy family specified by the rule.

[d] Number of sequences of a rule is the total number of sequences annotated by either CAZy family or by Pfam domains, or by both.

[e] Support in family is the percent of sequences supporting the rule in the reference set annotated by the family specified by the rule.