In silico Identification of Glycoside Hydrolase 9 Proteins from a Publicly Available Metagenome Database

William Wysocki

Introduction

Cellulose is a component of the plant cell wall and exists as a vital portion of the earth's biomass. Because of this, cellulose degradation is important in plants for physiological processes such as fruit expansion and cell division. It's also important to microorganisms for the utilization of monomeric sugars as a food source that results from cellulose breakdown. This can be accomplished by hydrolyzing the glycoside linkages present between the monomeric sugars in cellulose. Glycoside hydrolases (GH) are enzymes that potentially catalyze this reaction. There currently exist 61 families of GH enzymes recognized by the CAZy website (http://www.cazy.org) (Henrissat, 1997). Of these families, glycoside hydrolase family 9 (GH9) is the second largest and consists largely of cellulases. A cellulase is defined as an enzyme which uses cellulose as a biochemical substrate (Sukharnikov *et al.*, 2012).

Cellulases can also be an important economically commodity. These enzymes can catalyze the breakdown of cellulose sources such as plant materials such as leaf litter and unusable crop waste into smaller components. These sugars can be utilized in the microbial production of biofuels such as ethanol and butanol (Wilson, 2009). Identification of novel cellulases can aid in the development of a more efficient biofuel production as well as augment our understanding of the catalytic process.

Here we utilize a variety of software to identify GH9 domains banked within the NCBI database of nonredundant proteins and within the JGI metagenomic database. A phylogenetic analysis was also performed to determine which domains had not been previously banked and to search for new potential subfamilies of GH9 cellulases within publicly available metagenomic samples.

Materials and Methods

Locating Banked GH9 Domains

Accession numbers were extracted from MS Excel by copying all pages from the CAZy GH9 website (http://www.cazy.org/GH9.html) and copying only the column containing the accession numbers into a blank document using the text editor nano (FILE1). Shell commands were then used to extract only the lines that contained accession numbers (COM1). This was done by deleting the empty lines and then by extracting only the lines that contained numerical characters (FILE2). The resulting accession numbers were then uploaded to the NCBI Batch Entrez and the results were downloaded in fasta format (FILE3). After the GH9 Hidden Markov Model was downloaded offline (http://cys.bios.niu.edu/dbCAN/family.php?ID=GH9) (FILE4),

the hmmsearch command in the HMMER package (Finn *et al.*, 2008) was used to search for homologous domains within the NCBI fasta file (COM2). The 'domtblout' argument allowed all homologous domains to be located rather than the default option which finds only one best domain fit in each sequence (FILE5). Using a Perl script (SCRIPT1), only the accessions were extracted from the HMM output which had at least 80% coverage and an E-value of less than $1X10^{-5}$ (FILE6). Another Perl script (SCRIPT2) extracted the fasta sequences that corresponded to each accession (FILE7). A Python script (SCRIPT3) was used to extract only the domain sequence from each fasta sequence by reading the domain start and end numbers in the HMM output (FILE8). The Python script also generated a color file which assigned the color blue to each fasta sequence used.

Locating Metagenomic GH9 Domains

The JGI database had been downloaded by the instructor in its entirety in fasta format. The GH9 Hidden Markov Model was used once again to perform an hmmsearch command to search for homologous domains within the JGI fasta file (COM3). As before, the 'domtblout' argument allowed all homologous domains to be located (FILE9). Using a Perl script modified from before (SCRIPT4), only the accessions were extracted from the HMM output which had at least 80% coverage and an E-value of less than $1X10^{-5}$ (FILE10). Another Perl script (SCRIPT5) extracted the fasta sequences that corresponded to each accession (FILE11). A Python script (SCRIPT6) was used to extract only the domain sequence from each fasta sequence by reading the domain start and end numbers in the HMM output (FILE12). The Python script also generated a color file which assigned the color red to each fasta sequence used.

Alignment and Phylogenetic Analysis

The two fasta sequence outputs from the both python scripts (FILE12, FILE8) were concatenated using shell commands (COM4) (FILE13). They were then aligned using the sequence alignment tool MAFFT (COM5) (FILE14) (Katoh and Kuma, 2002). A phylogenetic tree was then generated using the FastTree program (COM6) (Price *et al.*, 2009) and was exported in Newick format (FILE15). This tree was then uploaded to iTOL (Letunic and Bork, 2006) along with the color file. The resulting colored phylogeny was then visually inspected to locate novel clades of GH9 proteins that were found in the metagenomic files.

Table 1: Shell commands used to parse data at various steps of the procedure. Numbers refer to labels in materials and methods section denoted by 'COM'.

1	less GH9.accessions.raw | sed '/^$/d' | grep [0-9] > GH9.accessions.clean
2	hmmsearch --domtblout GH9.NCBI.HMM.out GH9.hmm.ps gh9.fasta &
3	hmmsearch --domtblout GH9.JGI.HMM.out GH9.hmm.ps /home/yyin/db/jgi-v3.5.pr.fa &
4	cat ncbi.output jgi.output > totalregions.fasta
5	mafft –auto totalregions.fasta > totalregions.fasta.aln
6	/home/mrupani/Downloads/Fastree totalregions.fasta.aln > treeproject1.nwk

Table 2: Names of parsed data files and files present at intermediate steps, denoted by 'FILE'.

1	GH9.accessions.raw
2	GH9.accessions.clean
3	gh9.fasta
4	GH9.hmm.ps
5	GH9.ncbi.hmm.out
6	GH9.ncbi.hmm.out.4.1.1.hit_ids
7	GH9.ncbi.hmm.out.4.1.1.fa
8	ncbi.output
9	GH9.jgi.hmm.out
10	GH9.jgi.hmm.out.4.1.1.hit_ids
11	GH9.jgi.hmm.out.4.1.1.fa
12	jgi.output
13	totalregions.fasta
14	totalregions.fasta.aln
15	treeproject1.nwk

Table 2: Names of scripts used to parse data at different steps, denoted by 'SCRIPT'.

1	4_1.pl
2	Retrieve_ncbi_fasta.pl
3	Step5NCBI.py
4	4_1.pl
5	retrieve_fasta.pl
6	Step5JGI.py

Results

The Hidden Markov Model search against the banked NCBI GH9 proteins yielded 2,102 domain hits while the search against the JGI database yielded 15,428 domain hits. After they were filtered for coverage and E-value, the GH9 domains from the NCBI yielded 1,452 hits while the JGI database yielded 1,777 hits. The banked GH9 domains were then grouped by phylum and then by species. The top represented phyla can be found in Figure 3 while the most represented species can be found in Figure 2.

Figure 1 shows the phylogenetic tree generated using FastTree and colored using the iTOL website. Blue nodes indicate previously banked GH9 domains while red nodes indicate GH9 domains extracted from metagenomic data from the JGI database. Several clades completely composed of metagenomic samples are visible but four (denoted with arrows in Figure 1) contain enough members to be novel subfamilies of GH9 proteins.

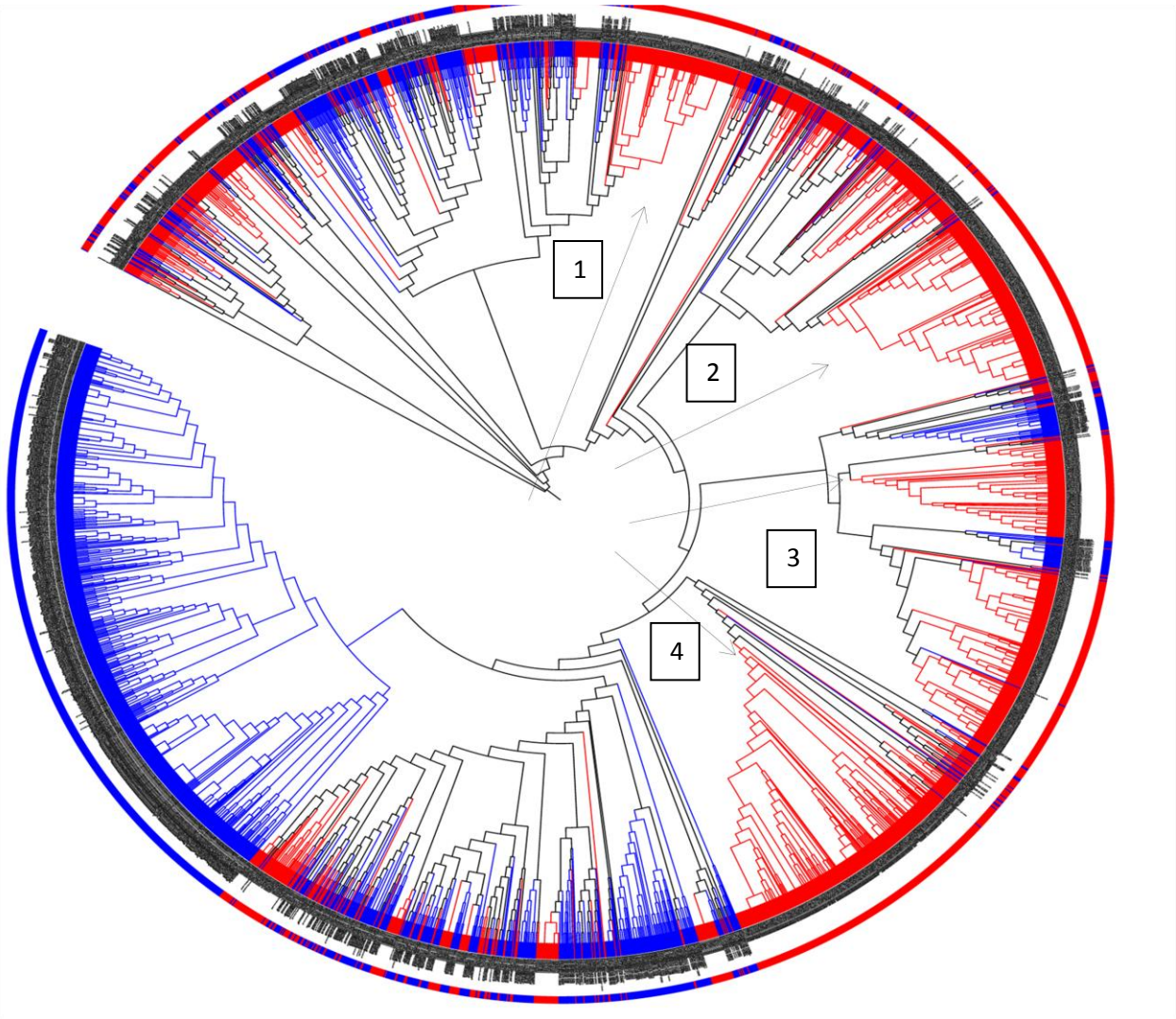


Figure 1: Phylogenetic tree of GH9 proteins. Blue taxa have previously been recorded in the NCBI while red taxa were identified from metagenomic samples in the JGI. Arrows denote potential novel subfamilies of GH9 proteins and are labeled 1-4.

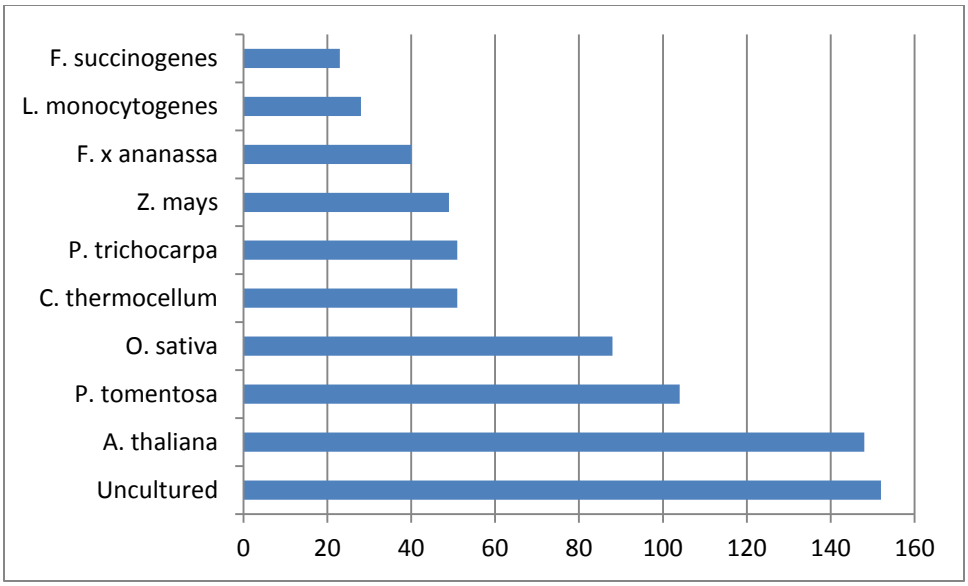


Figure 2: Top ten organisms that contained the most GH9 domains banked in the NCBI. Notice that even within the NCBI, a large amount of GH9 domains were discovered from uncultured organisms.

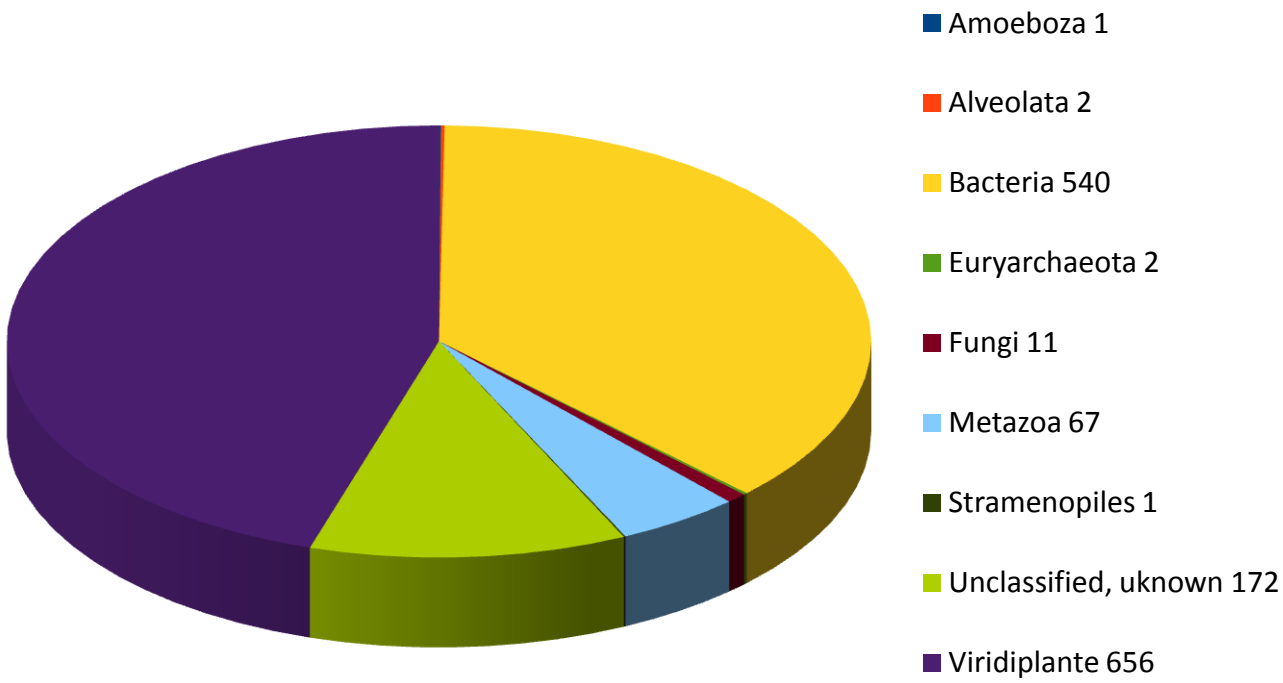


Figure 3: A representation of the phyla represented by searching known GH9 domains.

Discussion

Figures 4-7 demonstrate that most of the metagenomically identified GH9 domains are either from bacterial sources or have been synthetically constructed. The presence of mostly uncultured sources of GH9 domains illustrates the previous identification efforts of other studies and can also be seen in Figure 2. The non-monophyly of the GH9 domains may be an indication of horizontal gene transfer in this phylogeny. Certain GH9 domains show high sequence similarity but are present in distantly related phyla which also potentially indicates a horizontal gene transfer event. This is consistent with the results found by Sukharnikov *et al.* (2012) that microbial GH48 proteins may have exhibited horizontal gene transfer. Since GH9 domains are known to work synergistically with GH48 domains, it would make evolutionary sense that they would be inherited together or at least in the same manner.

Tracing the potential novel GH9 domains back to the metagenomes they were extracted from will allow scientists to locate sources of cellulase degrading activity. New GH9 domains were located in stool samples, the guts of animals such as reindeer and wallabies, hot springs, and compost piles. Even though most of the organisms cannot be conventionally cultured, knowledge of the sequence can allow for amplification using specific primers and expression through cloning.

Although four new potential monophyletic GH9 subfamilies were located, additional experiments need to be performed in order to verify the enzyme's identity as a cellulase. Sukharnikov *et al.* (2012) demonstrated that members of the glycoside hydrolase family 48, most of which are cellulases, could have completely different functionality such as chitinase activity, despite containing very similar sequence identity and even similar three dimensional structures. Like the previously mentioned study, experimental verification of novel GH9 protein cellulase activity would need to be associated to a certain motif or domain to verify which of these enzymes are actual cellulases.

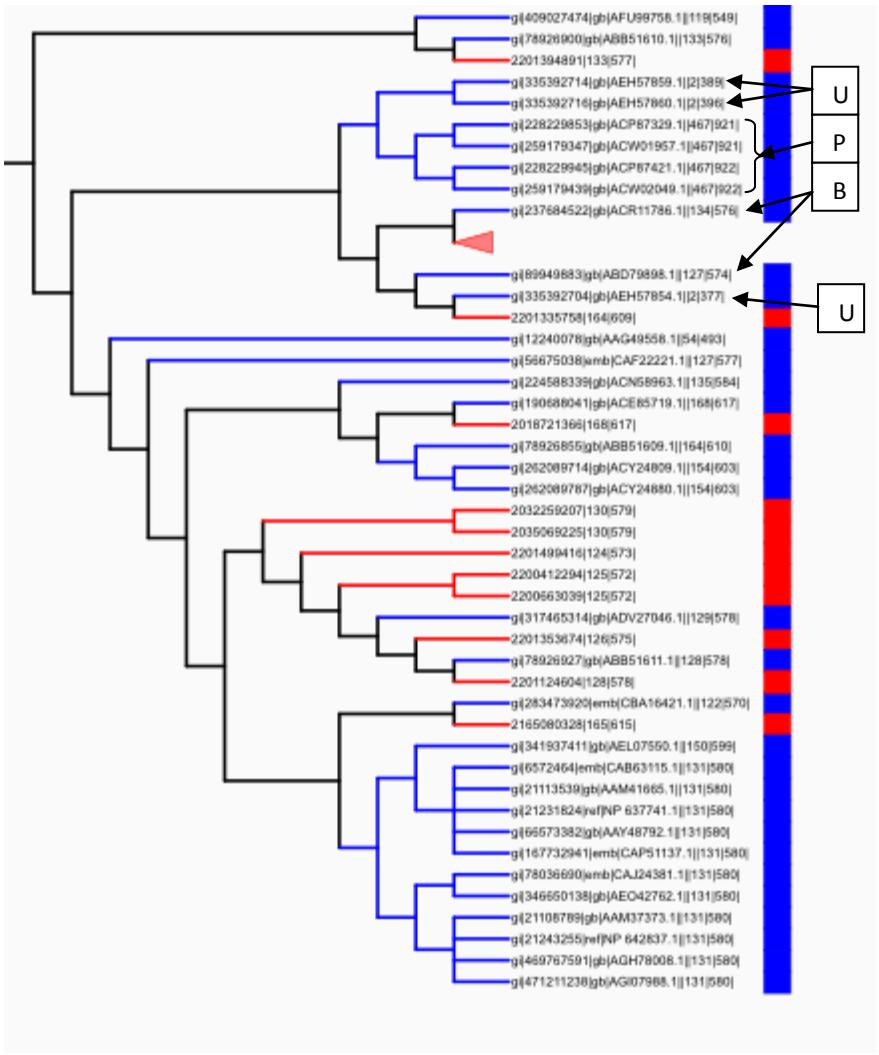


Figure 4: A cladogram that includes the potentially novel protein subfamily 1 (Labeled on Figure 1). The subfamily clade has been collapsed and the OTUs closely related to this clade are classified into bacteria (B), uncultured (U), synthetic (S) or patents (P).

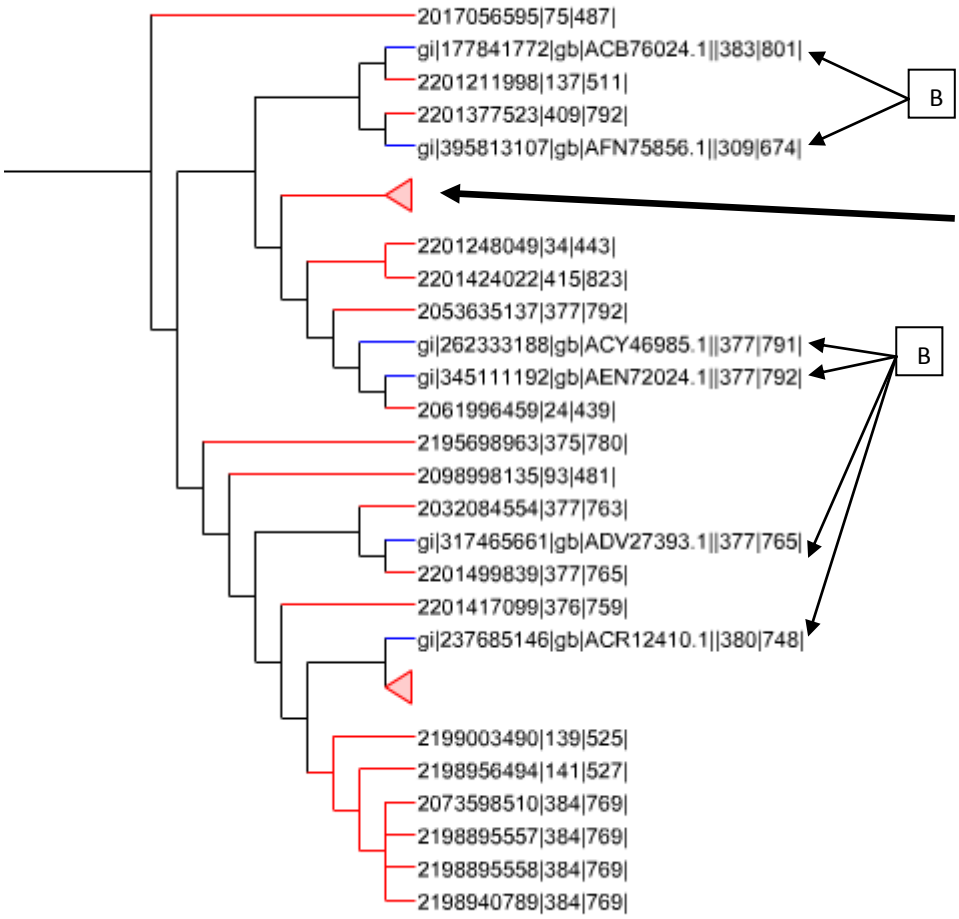


Figure 5: A cladogram that includes the potentially novel protein subfamily 2 (Labeled on Figure 1). Two clades has been collapsed and the largest is labeled with an arrow. The key is the same as Figure 4.

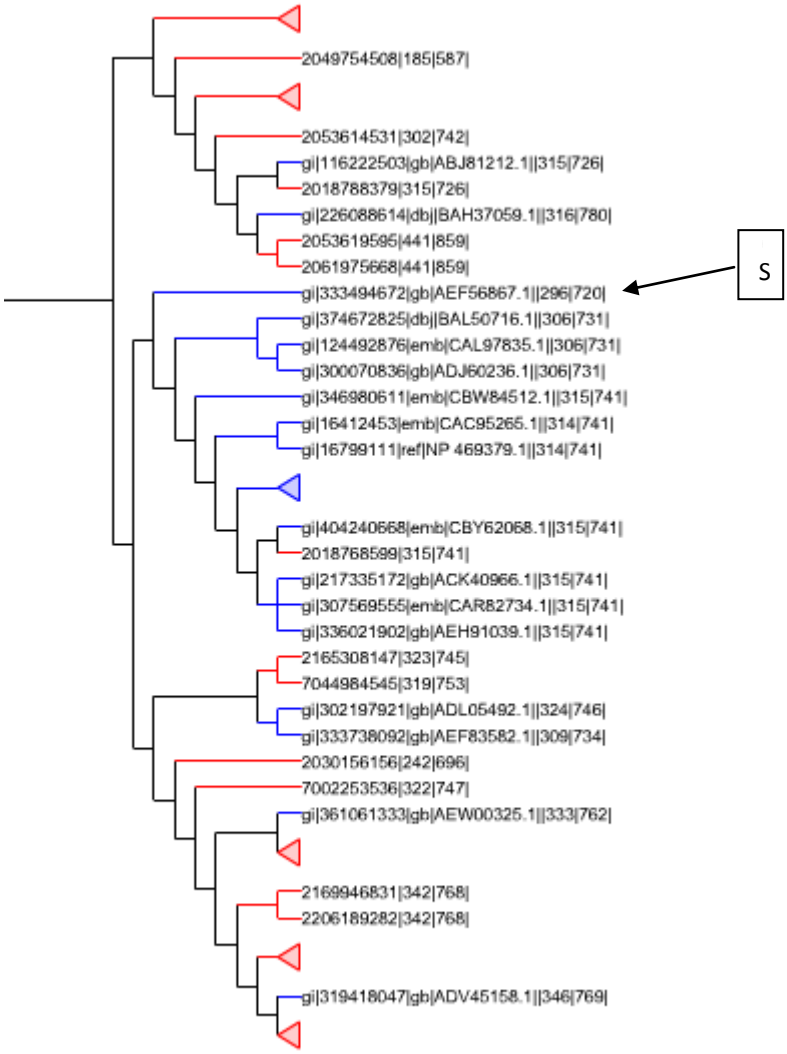


Figure 6: A cladogram that includes potentially novel protein subfamily 3 (Labeled on Figure 1). Several clades of varying sizes have been collapsed and every known OTU here is classified as a bacterium except for the labeled OTU.

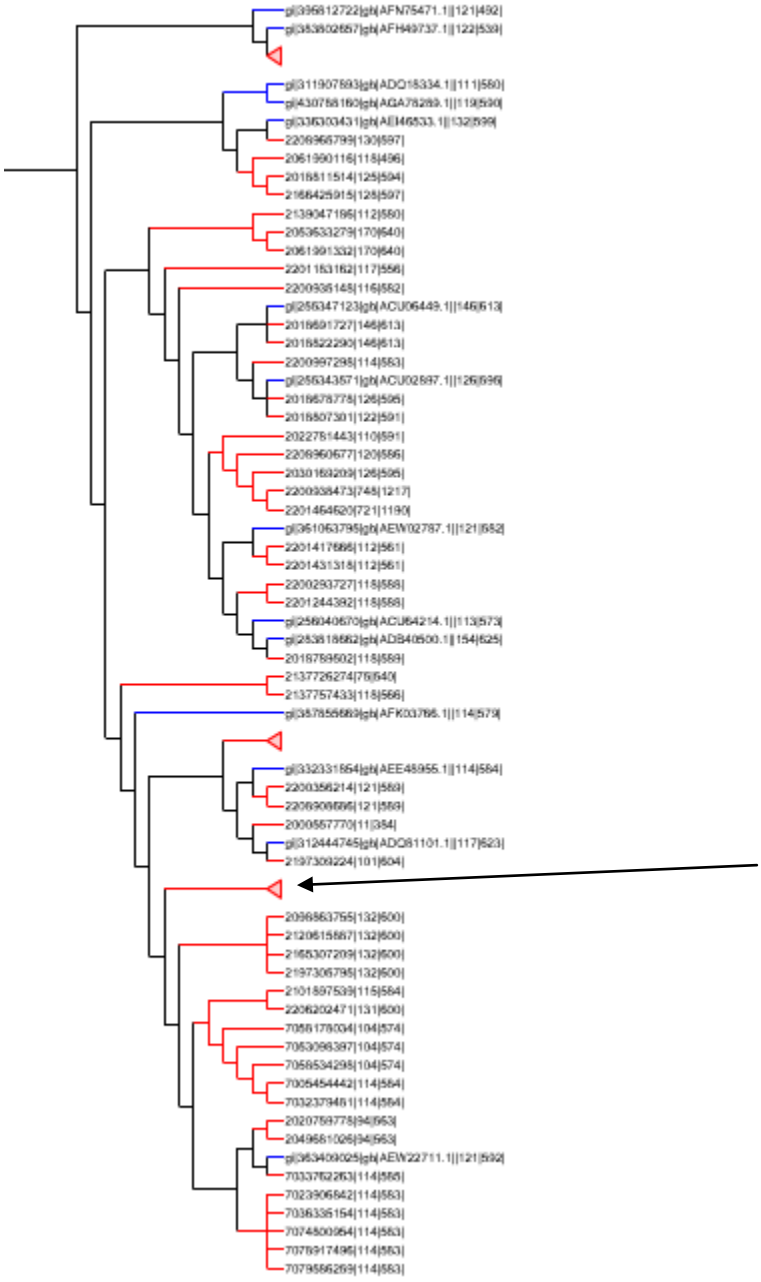


Figure 7: A cladogram that includes potentially novel protein subfamily 4 (Labeled on Figure 1). Several clades has been collapsed and the largest is labeled with an arrow. All of the OTUs with known sources have been identified as bacteria.

References

Finn, R.D., Clements, J., and S.R. Eddy. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*. **Web Server Issue.** 39:W29-W37.

Henrissat B, Davies GJ (1997) Structural and sequence-based classification of glycoside hydrolases. *Curr. Op. Struct. Biol*. **7**:637-644

Katoh, M and Kuma, M. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. **30**:3059-3066.

Letunic I. and Bork P. 2006. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23(1)**:127-128

Price, M.N., Dehal, P.S., and Arkin, A.P. 2009. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26:**1641-1650.

Sukharnikov, L.O., Alahuhta, M., Brunecky, R., Upadhyay, A., Himmel, M.E., Lunin, V.L., and I.B. Zhulin. 2012. Sequence, Structure, and Evolution of Cellulases in Glycoside Hydrolase Family 48. *The Journal of Biological Chemistry*. **287:** 41068-41077.

Wilson, D.B. 2009. Cellulases and biofuels. *Current Opinion in Biotechnology*. **20:** 295-299.