Maestría en Bioinformática y Biología Computacional

Reclasificación de las enzimas glicosil hidrolasas presentes en la base de datos CAZy con el fin de mejorar la anotación de enzimas degradadoras de celulosa y hemicelulosa

JORGE HERNÁN SUÁREZ OSORIO

UNIVERSIDAD CATÓLICA DE MANIZALES
MAESTRÍA EN BIOINFOMÁTICA Y BIOLOGÍA COMPUTACIONAL
MANIZALES
2018

Reclasificación de las enzimas glicosil hidrolasas presentes en la base de datos CAZy con el fin de mejorar la anotación de enzimas degradadoras de celulosa y hemicelulosa

Tesis de Maestría para optar al título de Magister en Bioinformática y Biología Computacional del estudiante:

JORGE HERNÁN SUÁREZ OSORIO

Directora:
ANDREA GARAVITO ESPEJO. Ph.D.

Co-directora:
GLORIA MARÍA RESTREPO FRANCO. Ph.D.

UNIVERSIDAD CATÓLICA DE MANIZALES
MAESTRÍA EN BIOINFOMÁTICA Y BIOLOGÍA COMPUTACIONAL
MANIZALES
2018

Tabla de contenido

Agradecimientos 6
Resumen 7
Abstract 8
Lista de tablas 9
Lista de figuras 10
Lista de anexos 11
1. Introducción 12
1.1. Campo temático 12
1.2. Planteamiento del problema 14
1.3. Justificación 15
1.4. Objetivos 17
1.4.1. Objetivo general 17
1.4.2. Objetivos específicos 18
2. Referente teórico y antecedentes 19
2.1. Antecedentes 19
2.1.1. CAT (CAZymes Analysis Toolkit) 19
2.1.2. dbCAN 20
2.1.3. HotPEP 21
2.1.4. MycoCLAP 22
2.1.5. Macroproyecto “Construcción de un dataset de familias de enzimas lignocelulolíticas para su anotación y minería de datos de proyectos de las ómicas” 23
2.2. Referente teórico 24
2.2.1. Complejo lignocelulósico 24
2.2.2. Celulosa 25
2.2.3. Hemicelulosa 27
2.2.4. Lignina 30
2.3. Degradación de lignocelulosa 31
2.3.1. Degradación de lignocelulosa y las glicosil hidrolasas 33
2.3.2. Degradación de la celulosa 35
2.3.3. Degradación de hemicelulosa 38
2.3.4. Degradación de hemicelulosa tipo xilano 38

2.3.5. Degradación de hemicelulosa tipo xiloglucano 41
2.3.6. Degradación de hemicelulosa tipo manano 42
2.3.7. Bases de datos en Bioinformática 43
2.3.8. CAZy 44
2.3.9. Homología de secuencias y OrthoMCL 46
2.3.10. Métodos de comparación entre secuencias de proteínas 48
2.3.11. Modelos Ocultos de Markov (HMM) 49
3. Materiales y métodos 52
3.1. Identificación de nuevos grupos de secuencias de enzimas glicosil hidrolasas correspondientes a familias y subfamilias de acuerdo con su similitud, a partir de la base de datos pública CAZy 53
3.1.1. Obtención de las secuencias de enzimas del CAZy 53
3.1.2. Alineamientos pareados con el software BLAST del NCBI 53
3.1.3. Agrupación de las subfamilias utilizando OrthoMCL 55
3.1.4. Generación de subfamilias a partir de los resultados del OrthoMCL .. 56
3.2. Validación *in silico* de los nuevos grupos de familias y subfamilias obtenidos, a través de la generación de perfiles proteicos, para su uso posterior en la anotación de secuencias de enzimas celulolíticas y hemicelulolíticas. 57
3.2.1. Construcción de la Base de datos de enzimas celulolíticas y hemicelulolíticas 57
3.2.2. Generación de perfiles proteicos utilizando modelos ocultos de Markov con HMMER3 59
3.2.3. Validación de los resultados de las nuevas familias 59
3.3. Desarrollo de una herramienta para la anotación de secuencias de enzimas degradadoras de celulosa y hemicelulosa 60
3.3.1. Programación de un servicio para anotación de enzimas degradadoras de celulosa y hemicelulosa 60
4. Resultados y discusión 61
4.1. Identificación de los nuevos grupos de secuencias de enzimas glicosil hidrolasas correspondientes a familias y subfamilias de acuerdo con su similitud a partir de la base de datos pública CAZy 61
4.1.1. Obtención de las secuencias de enzimas del CAZy 61
4.1.2. Alineamientos pareados con el software BLAST del NCBI 64
4.1.3. Agrupación de las subfamilias utilizando OrthoMCL 68

4.1.4. Generación de subfamilias a partir de los resultados del OrthoMCL .. 69
4.2. Validación *in silico* de los nuevos grupos de familias y subfamilias obtenidos, a través de la generación de perfiles proteicos, para su uso posterior en la anotación de secuencias de enzimas celulolíticas y hemicelulolíticas. 70
4.2.1. Construcción de la base de datos de enzimas celulolíticas y hemicelulolíticas 70
4.2.2. Generación de perfiles proteicos utilizando modelos ocultos de Markov con HMMER3 71
4.2.3. Validación de los resultados de las nuevas familias 72
4.3. Desarrollo de una herramienta para la anotación de secuencias de enzimas degradadoras de celulosa y hemicelulosa 79
4.3.1. Programación de un servicio para anotación de enzimas degradadoras de celulosa y hemicelulosa 79
4.4. Resultados extras de aplicación general 81
5. Conclusiones generales 90
5.1. Contribuciones de la Tesis 90
5.2. Impactos Potenciales de la Tesis 90
5.3. Recomendaciones y trabajos futuros 91
6. Referencias bibliográficas 93
7. Anexos 98

Agradecimientos

Instituciones y oficinas

- A la *Colección de Microorganismos* de la Universidad Católica de Manizales por su apoyo en la realización de la presente tesis.
- Al *Centro de Bioinformática y Biología Computacional* BIOS por el uso del clúster para procesar los alineamientos requeridos para el proyecto.
- Al proyecto de Regalías, Caldas Bioregión, por la beca suministrada para la realización de esta tesis.
- Al *Grupo de Investigaciones Biológicas (GIBI)*, por el apoyo permanente durante el desarrollo de la investigación.

Doctores e investigadores

- A la Dra. Andrea Garavito Espejo, Directora de la tesis, por su apoyo en la revisión y desarrollo de la tesis.
- A la Dra. Gloria María Restrepo Franco, Co-directora, por su apoyo como guía de desarrollo de este trabajo.
- A mi compañero Narmer Fernando Galeano Vanegas, por su apoyo metodológico y teórico durante todo el proyecto.

Otras personas

- A Mónica Quintero quien fue de gran ayuda en el inicio del planteamiento y desarrollo del proyecto como tutora de la tesis, mientras formó parte de la Universidad Católica de Manizales.
- A el doctor Raúl Ramos Pollán y la doctora Sandra Montoya Barreto por sus aportes y recomendaciones para la organización de la tesis.

Agradecimientos personales

- A mi familia quien me ha apoyado incondicionalmente para alcanzar los más importantes logros de mi vida, especialmente a mi madre.

Resumen

La lignocelulosa es un conjunto de polímeros contenidos en la pared celular de las plantas, conformada por tres polisacáridos principales: celulosa, hemicelulosa y lignina. Dicho complejo ha sido estudiado en épocas recientes por su gran potencial en cuanto al contenido energético y su utilización para la creación de biocombustibles, como también en metodologías ambientalmente amigables para su degradación efectiva. Actualmente existe una gran cantidad de secuencias de enzimas con potencial para hacer degradación de celulosa y hemicelulosa en las bases de datos públicas como *Genbank*, cuya adquisición puede ser más económica que las disponibles comercialmente, pero hay pocas herramientas investigativas que las haya tenido en cuenta para facilitar su búsqueda. También existen clasificaciones de estas enzimas como las disponibles en la base de datos CAZy (*Carbohydrate-Active enZYmes Database*), pero su uso de cara al investigador es difícil por la accesibilidad de los datos disponibles, además de que falta una mayor especificidad de estas clasificaciones para investigaciones más especializadas, pero ningún estudio las ha abordado para su fácil investigación. La búsqueda de este tipo de enzimas es un reto para los investigadores.

De acuerdo con lo anterior surgió la necesidad de reclasificar las enzimas glicosil hidrolasas (GH) presentes en la base de datos CAZy, clase que tiene entre otras funciones la degradación de celulosa y hemicelulosa. Para esto, en el desarrollo metodológico se obtuvieron las secuencias publicadas en CAZy las cuales se dividieron según la base de datos (PDB, *Uniprot* y *Genbank*) y la taxonomía de los organismos de procedencia (bacterias, virus, eucariotas, arqueas y no clasificados). Posteriormente, se realizaron alineamientos de éstas contra la base de datos no redundante (NR) del NCBI, con el fin de aumentar la cantidad de enzimas a investigar. Se realizó luego la agrupación por familias y subfamilias con el *software* OrthoMCL, y se crearon datos de referencia de alineamiento con las herramientas Blast y HMMER3.

Se generó una base de datos de referencia que permite la anotación de enzimas degradadoras de celulosa y hemicelulosa. En la cual para la base de datos PDB se pasó de 63 familias a 497 *clusters*, para *Uniprot* de 197 a 10.822 y para *GenBank* de 231 a 6.874. Estos resultados fueron validados para la familia GH6 contra la base de datos RefSeq donde se evaluaron falsos negativos y falsos positivos. Se desarrolló un *script* para ejecutar búsquedas de enzimas teniendo como referencia los archivos de alineamientos, que da como resultado un reporte.

Como resultados del proyecto se relacionan dos bases de datos de referencia de alineamientos, un *script* para realizar búsquedas, y un servicio *web* para realizar búsquedas de estas enzimas. Los *clusters* finales tras el uso de OrthoMCL requieren de una mayor curación para que los modelos queden correctamente hechos.

Palabras clave: Enzimas degradadoras de celulosa y hemicelulosa, CAZy, perfiles proteicos.

Abstract

Lignocellulose is a group of polymers contained in the cellular wall of plants, consisting of three main polysaccharides: cellulose, hemicellulose and lignin. This complex has been studied in recent times for its great potential in terms of energy content and its use for the creation of biofuels, as well as in environmentally friendly methodologies for its effective degradation. Currently there is a large number of enzyme sequences with potential to make cellulose and hemicellulose degradation in public databases such as Genbank, whose acquisition may be cheaper than those available commercially, but there are few research tools that have taken them into account for facilitate its research. There are also classifications of these enzymes, such as those available in the CAZy database (Carbohydrate-Active enZYmes Database), but their use for researchers is difficult because of the accessibility of the available data, as well as the lack of specificity of these classifications for more specialized research, but no study has addressed them for easy research. The search for this type of enzyme is a challenge for researchers.

According to the above, there was a need to reclassify the glycosyl hydrolases (GH) enzymes present in the CAZy database (*Carbohydrate-Active enZYmes Database*), a class which has the function of degrading cellulose and hemicellulose, among other functions. For this, in the methodological development, the sequences published in CAZy were obtained, which were divided according to data base (PDB, Uniprot and Genbank) and taxonomy (bacteria, virus, eukaryotes, archaea and unclassified), to then align these against the non-redundant database (NR) of the NCBI, in order to increase the number of enzymes to be researched. The clustering by families and subfamilies was then performed with the OrthoMCL software, and alignment reference data was created with the Blast and HMMER3 tools.

A reference database was generated allowing the annotation of degrading enzymes of cellulose and hemicellulose. For the PDB database it went from 63 families to 497 clusters, for Uniprot from 197 to 10,822 and for GenBank from 231 to 6,874. These results were validated for the GH6 family against the RefSeq database, where false negatives and false positives were evaluated. A script was developed to execute enzyme searches based on the alignment files that result in a report.

The results of the project are related to two reference databases of alignments, a search script, and a web service to search these enzymes. The final clusters after the use of OrthoMCL require a better curation so that the models are correctly done.

Keywords: Cellulose and hemicellulose degrading enzymes, CAZy, protein profiles.

Lista de tablas

Tabla 1. Porcentaje de composición de lignocelulosa. 25
Tabla 2 Tipos de hemicelulosa. 27
Tabla 3. Proporción de azúcares para las hemicelulosas de madera blanda. . 29
Tabla 4. Enzimas que intervienen en la degradación de lignina. 31
Tabla 5. Características principales de los tipos de degradación de la lignocelulosa, mediada por hongos. 32
Tabla 6. Enzimas implicadas en la degradación de lignocelulosa. 33
Tabla 7. Relación de enzimas degradadoras de celulosa y hemicelulosa con familias GH. 34
Tabla 8. Enzimas requeridas para hidrólisis de hemicelulosa. 38
Tabla 9. Clasificación de enzimas presentes en CAZy. 44
Tabla 10. Métodos para determinación de secuencias ortólogas. 47
Tabla 11. Distribución de las familias de las secuencias iniciales, por taxonomía y por base de datos. 53
Tabla 12. Familias filtradas para el proyecto. 58
Tabla 13. Estadísticas generales de las secuencias de acuerdo con la base de datos. 61
Tabla 14. Distribución de familias por taxonomía y por base de datos según CAZy (Fecha de consulta: octubre 12 de 2017). 62
Tabla 15. Estadísticas iniciales base de datos PDB. 63
Tabla 16. Variación de las longitudes de las secuencias por base de datos. .. 63
Tabla 17. Tiempos de ejecución del alineamiento pareado con Blast. 64
Tabla 18. Resultado de la ejecución del filtro (3). 65
Tabla 19. Estadísticas del filtro de los alineamientos para la base de datos PDB. 66
Tabla 20. Estadísticas de las secuencias filtradas base de datos PDB. 66
Tabla 21. Listado de familias descartadas por el filtro. 67
Tabla 22. Variación de las longitudes de las secuencias por base de datos. .. 68
Tabla 23. Resultado general de la ejecución del MCL. 69
Tabla 24. Estadísticas finales base de datos PDB. 69
Tabla 25. Análisis de subfamilias de bacterias GH6 para *Genbank*. 74
Tabla 26. Análisis de subfamilias de eucariotas GH6 del PDB. 76
Tabla 27. Actividades familia GH6. 77
Tabla 28. Número de secuencias filtradas por el HMMSearch. 78
Tabla 29. Matriz de conjunción subgrupo GH6 – Bacteria. 78
Tabla 30. Distribución los falsos positivos. 78
Tabla 31. Parámetros de ejecución del *script* searcher.py. 80
Tabla 32. Muestra de secuencias que presentan anormalidades. 82
Tabla 33. Enzimas degradadoras de lignina no tenidas en cuenta. 89

Lista de figuras

Figura 1. Servicio web del proyecto dbCAN. 21
Figura 2 Búsqueda del término *Xylanase* en la base de datos MycoCLAP. 22
Figura 3. Estructura química de la celulosa, las líneas punteadas son enlaces de hidrogeno. 26
Figura 4 Mananos tipo galactomanano y galactoglucomanano. 28
Figura 5. TIpos de xilano, glucuronoxilano y glucuronoarabinoxilano. 29
Figura 6. Tipos de xiloglucano, fucogalactoxiloglucano y arabinogalactoxiloglucano. 30
Figura 7. Esquema de degradación de celulosa. 35
Figura 8. Esquema de degradación de hemicelulosa tipo xilano. 39
Figura 9. Esquema de degradación de hemicelulosa tipo xiloglucano. 41
Figura 10. Esquema de degradación de hemicelulosa tipo galactoglucomanano. 42
Figura 11. Vista de la parte de la familia GH1 publicada en CAZy 45
Figura 12. Homología de secuencias, ortólogos y parálogos. 46
Figura 13. Ejemplo de funcionamiento de un *Hidden Markov Model*. 51
Figura 14. Flujo de trabajo general para la construcción de la base de datos de los nuevos grupos de enzimas degradadoras de celulosa y hemicelulosa. 52
Figura 15. Salida del MCL para PDB 57
Figura 16. Ejemplo de estadística de subfamilias para PDB 57
Figura 17. Histograma de distribución de longitudes de las secuencias para la base de datos *GenBank*. 64
Figura 18. Histograma de longitudes de secuencias para la base de datos *GenBank* después del alineamiento. 68
Figura 19. Subfamilias asociadas a la familia GH6 de Eukaryota 70
Figura 20. Funciones enzimáticas para degradación de celulosa y hemicelulosa con familias GH 71
Figura 21. Árbol filogenético familia GH6 PFam. 73
Figura 22. Árbol filogenético de la subfamilia GH6 Bacteria Genbank 233. 75
Figura 23. Test de falsos positivos para la familia GH6. 77
Figura 24. Herramienta de consulta visual LignoSearch. 79
Figura 25. Vista de los taxones de una subfamilia específica en la herramienta LignoSearch. 80
Figura 26. Prueba de la existencia de secuencias irregulares en CAZy. 82
Figura 27. Existencia de la enzima ACX49739.1 en CAZy. 86
Figura 28. Alineamiento de la enzima ACX49739.1 contra WP_0675936.1. ... 86
Figura 29. Ejecución de InterPro Scan para ACX49739.1. 87
Figura 30. Ejecución de InterPro Scan para WP_067593936.1. 88

Lista de anexos

Anexo I. Familia Glicosil Hidrolasas del CAZy. 98
Anexo II. Archivos iniciales CAZy 99
Anexo III. Script de Python para filtrado de secuencias de alineamientos pareados en formato tabla..... 100
Anexo IV. Pasos siguientes de OrthoMCL. 101
Anexo V. Estadística de las familias iniciales. 102
Anexo VI. Listado de casos anormales de las familias iniciales del CAZy. 103
Anexo VII. Histogramas de longitudes de las secuencias iniciales. 104
Anexo VIII. Estadística de los alineamientos. 105
Anexo IX. Estadística de las secuencias filtradas. 106
Anexo X. Secuencias pertenecientes a múltiples familias. 107
Anexo XI. Resultados y estadística de la ejecución del MCL. 108
Anexo XII. Asociación de las secuencia a los identificadores por cluster. 109
Anexo XIII. Estadística de las secuencias finales. 110
Anexo XIV. Estadística de los clusters del OrthoMCL 111
Anexo XV. Script para busqueda de enzimas degradadoras de celulosa y hemicelulosa. 112

1. Introducción

Este trabajo está dividido en cuatro segmentos principales, introducción, referente teórico, metodología y resultados. En la primera parte se establece la problemática alrededor del proyecto de una manera rigurosa planteando formalmente las causas que llevaron a su desarrollo, pasando posteriormente por la justificación y los objetivos planteados. En la segunda parte se relaciona una recolección de datos y referentes teóricos necesarios para explicar el desarrollo del trabajo, donde inicialmente se analizan las tres herramientas más importantes relacionadas con el proyecto: Hotpep (P. K. Busk et al., 2017), CAT (CAZymes *Analysis Toolkit*) (Park et al., 2010) y dbCAN (*database for Carbohydrate active enzymes ANnotation*), para luego abordar el concepto de lignocelulosa, haciendo énfasis en la celulosa y hemicelulosa, y las clasificaciones de enzimas que existen para su degradación., Por último se explican los métodos de comparación que se utilizan para hacer búsqueda de secuencias. En el tercer segmento se describe la metodología para el desarrollo del flujo de trabajo en la búsqueda de enzimas degradadoras de celulosa y hemicelulosa, y se explica por partes cada uno de los pasos que se utilizaron en la extracción final de grupos de familias y perfiles proteicos, que son los principales insumos para el análisis con el *script* final de búsqueda de enzimas que se desarrolló. Finalmente, en la cuarta parte se plasman todos los resultados finales logrados con el proyecto.

1.1. Campo temático

Millones de toneladas de desechos son eliminadas al año producto del desarrollo de consumibles a base de insumos agrícolas, causando un problema ambiental que no parece tener una solución a corto plazo, desaprovechando además todo el beneficio potencial que está en la degradación de residuos en compuestos más útiles, como se ha demostrado en el caso de obtención de bioetanol a partir de la biomasa vegetal. Adicionalmente la reducción cada vez más avanzada de combustibles fósiles como el petróleo, plantea un problema de tipo energético, que puede ser solucionado con la generación de biocombustibles. La degradación del material lignocelulósico por medio de enzimas es un campo prometedor que genera grandes beneficios económicos, sociales y ambientales.

Esta investigación trata sobre las enzimas degradadoras de lignocelulosa, específicamente de las hidrolasas que degradan celulosa y hemicelulosa. La lignocelulosa

es el principal compuesto presente en la pared celular de las plantas y es conocida por ser el mayor componente de los desechos vegetales de cualquier tipo. Los investigadores en este campo de búsqueda de enzimas carecen de herramientas suficientes para desarrollar su respectiva investigación, lo que genera dificultades en tiempo y dinero en los desarrollos de los proyectos de investigación en el área temática de la degradación de enzimas.

Existen bancos de enzimas comerciales que brindan información sobre los materiales lignocelulósicos que éstas degradan, sin embargo, para su acceso se debe pagar un valor económico significativo que dificulta su accesibilidad, además no tienen una estructura organizada que permita explotar mejor dicha información.

Es por esto que el Grupo de Investigaciones Biológicas – GIBI, en el marco de sus tres líneas de investigación: Estudio y conservación de la diversidad microbiana, Bioinsumos, y Aprovechamiento biológico de residuos agroindustriales, ha estado interesado en orientar las acciones institucionales hacia las áreas de la biotecnología agrícola, bioinsumos y la tecnología de enzimas, con el objetivo de contribuir al desarrollo científico y tecnológico, mediante la producción de conocimiento, formación investigativa y el aporte de soluciones a las necesidades del sector productivo, comunitario e industrial.

El grupo está conformado por investigadores experimentados en diversas áreas del conocimiento como microbiología, bacteriología, biología, ingeniería química, ingeniería biológica y ciencias biológicas, principalmente.

En el año 2014 se presentó la propuesta "*Construcción de un dataset de familias de enzimas lignocelulolíticas para su anotación y minería de datos de proyectos de las ómicas*", el cual tiene por objetivo construir un *dataset* de familias de enzimas involucradas en el proceso de degradación de la lignocelulosa para facilitar la anotación y minería de datos provenientes de proyectos de –ómicas. En el marco de este proyecto la presente tesis aporta los *datasets* de enzimas que degradan celulosa y hemicelulosa, dos de los tres componentes principales de la lignocelulosa, faltando únicamente por desarrollar el relacionado con la lignina.

El estudiante investigador de este proyecto de tesis es beneficiario de las becas de la convocatoria "Caldas Bioregión". Por lo tanto, el cumplimiento de este proyecto fue supervisado por la entidad administradora del proyecto de regalías que ejecutó la asignación de los recursos de la convocatoria, el Centro de Bioinformática y Biología Computacional BIOS.

1.2. Planteamiento del problema

Uno de los inconvenientes de la agroindustria es la generación de grandes cantidades de biomasa lignocelulósica; solo de celulosa y hemicelulosa se generan un total de desechos de $7,2 \times 10^{10}$ y 6×10^{10} toneladas anuales respectivamente (Kubicek, 2013). China, por ejemplo, tiene abundantes recursos de biomasa lignocelulósica, se reporta que se producen casi 0,73 mil millones de toneladas de residuos agrícolas por año de sus industrias agroindustriales y forestales (Cao G. et al, 2016). A pesar de que existen leyes para el manejo adecuado de los residuos, el tratamiento de algunos de ellos es de alto costo y requiere de maquinaria especializada para llevarse a cabo. En el informe de la Superservicios (2013) “Situación de la disposición final de residuos sólidos en Colombia”, se enuncia que el departamento de Caldas produce en promedio 794,2 toneladas/día de residuos totales (incluidos lignocelulósicos), por lo que en el plan de acción estratégico del sector BIO de Manizales se prioriza la utilización de conocimientos científicos y de la ingeniería para la obtención de procesos a escala industrial, que permitan su reutilización.

Los desechos de origen vegetal, deben ser pre-tratados mediante hidrólisis, para facilitar su uso en la obtención de compuestos de importancia económica. La hidrólisis se puede realizar por vía ácida o enzimática, siendo la primera la más utilizada. Sin embargo, la hidrólisis ácida resuelve temporalmente el problema de la disminución de desechos agroindustriales debido a que a largo plazo genera un problema mayor de contaminación ambiental. La segunda opción basada en el uso de enzimas, mitiga el problema ambiental, pero es mucho más costosa dado a que el aislamiento y caracterización de las enzimas se hace en laboratorios especializados, y los requerimientos de producción a gran escala limitan su implementación en la industria. Por lo tanto, aún existe la necesidad de continuar investigando en el desarrollo de enzimas y su producción, para superar las anteriores dificultades, debido al enorme potencial que se ha reconocido en éstas.

De acuerdo con lo anterior, el estudio de enzimas degradadoras de lignocelulosa constituye una necesidad investigativa, dada la gran cantidad de material lignocelulósico que se genera en la agroindustria, el cual tiende a incrementar a través de los años. La investigación biotecnológica de diversos mecanismos enzimáticos que contribuyen al proceso de degradación de la lignocelulosa es relevante, permitiendo a las empresas agroindustriales solucionar el problema de la disminución de costos en la producción enzimática y obtener subproductos, como los biocombustibles.

Para realizar la investigación en este campo, existe información de una gran cantidad de enzimas, a las cuales se les conocen sus actividades biológicas (xilanas, mananasas, etc), y las fuentes de obtención. En los grupos de investigación que tienen como objeto el aprovechamiento de materiales lignocelulósicos, surge la necesidad de determinar el mejor microorganismo o enzima a utilizar en los diseños experimentales de las investigaciones, con el fin de facilitar la toma de decisiones que finalmente evita una inversión excesiva en tiempo y dinero. Como se aprecia en estas investigaciones, es crítica la fase de selección de la enzima a emplear, de acuerdo con sus características y especificaciones.

Las bases de datos bioinformáticas como el NCBI (*National Center for Biotechnology Information*), PDB (*Protein Data Bank*) (Berman et al., 2000), UNIPROT (*Universal Protein Resource*) (Apweiler et al., 2004) o CAZy (*Carbohydrate Active Enzymes*) (Cantarel et al., 2009) se convierten en una herramienta indispensable al momento de desarrollar las investigaciones en laboratorio, pero hasta el momento no existe una base de datos de referencia depurada y consolidada, que permita: (i) la anotación de datos de enzimas celulolíticas y hemicelulolíticas de una manera más fácil e intuitiva que el proceso actual, (ii) la reducción de la incertidumbre y la comprobación de los perfiles proteicos de las enzimas que se quieran investigar, con el fin de fortalecer las investigaciones en el campo.

Pregunta de investigación: ¿La información de las enzimas celulolíticas y hemicelulolíticas puede ser organizada a través de un flujo de trabajo que permita su anotación?

1.3. Justificación

La creciente disminución de las fuentes petrolíferas ha promovido la investigación en la búsqueda de otras fuentes de energía equivalentes, que cumplan la misma función que los combustibles derivados del petróleo, y que en un futuro puedan ser el reemplazo renovable de los mismos. Se ha encontrado una opción en la lignocelulosa, ya que, al ser un complejo polisacárido de alto contenido energético, produce en su degradación mediante hidrólisis enzimática moléculas de alcohol, como el etanol y el butanol, que a su vez son combustibles. Para lograr esto, se han realizado investigaciones en diferentes tipos de microorganismos que faciliten el proceso de adquisición de estos derivados.

En las últimas décadas, y con el avance en el conocimiento de los microorganismos y sus metabolitos, se han utilizado enzimas microbianas para degradar los sustratos de difícil manejo. Esta metodología es en consecuencia aplicable a las enzimas lignocelulolíticas. Recientemente la vinculación de herramientas bioinformáticas que ayudan a los investigadores en el diseño o hallazgo de posibles enzimas degradadoras han contribuido a optimizar el proceso. La definición de perfiles proteicos de las diferentes familias y subfamilias puede dar indicios mucho más claros de la estructura de las enzimas, y facilitar el filtrado y selección de los materiales de investigación, debido a que la función enzimática está estrechamente ligada a la estructura terciaria de las moléculas.

La investigación sobre este tipo de enzimas que degradan el material lignocelulósico es importante dado que los productos resultantes de la degradación son útiles para la industria, por ejemplo, los xilo-oligosacáridos y la xilobiosa son de interés para la industria alimenticia debido a la aplicación como prebióticos y endulzantes. Por otra parte, la producción de xilosa directamente de estos sustratos se usa en la industria de pulpa de papel o en la subproducción de xilitol, un edulcorante alternativo (Alvarez et al., 2013). Finalmente, estas enzimas son importantes porque los productos resultantes de la degradación de celulosa y hemicelulosa como la manosa, glucosa, galactosa, xilosa o arabinosa se pueden convertir en combustibles renovables tipo etanol (Gírio et al., 2010) por medio de un proceso de fermentación alcohólica, siendo considerada como una tecnología prometedora para sustituir a los combustibles fósiles y para asistir a la necesidad mundial de energía limpia. Sin embargo, a pesar del reciente crecimiento en la producción de biocombustibles a partir de biomasa vegetal, todavía existen varios cuellos de botella tecnológicos que hacen que el proceso de bioconversión aún no sea rentable.

La adquisición de enzimas degradadores de celulosa y hemicelulosa en laboratorios comerciales puede ser muy costosa, por lo que la creación y la búsqueda con este tipo de herramienta presentada para este proyecto puede permitir obtener y ofrecer al mercado enzimas más económicas de microorganismos diferentes a los comerciales.

Según Brink et al. (2011), con el aumento de la información bioquímica, y del establecimiento de familias más definidas (como las incluidas en el *CAZy*), a través de la división en subfamilias más pequeñas y mejor definidas de acuerdo con su función hidrolítica, permitirá un mejor valor predictivo para las anotaciones futuras de secuencias, que finalmente permitan la consolidación de mejores herramientas para apoyar los

propósitos industriales. Los anteriores planteamientos se incluyen en buena medida en la intención de este proyecto.

De acuerdo con lo anterior, este trabajo pretende presentar a la comunidad científica un nuevo enfoque para hacer análisis y búsqueda de enzimas celulolíticas y hemicelulolíticas, en el cual se destaca la inclusión de las bases de datos disponibles actualmente. Este análisis se logra haciendo una división de las secuencias por taxonomía (Eukaria, Bacteria, Archaea, Virus), y subdividiéndolas además con las definiciones de las familias de enzimas de carbohidratos presentes en CAZy (Cantarel et al., 2009) para con ellas obtener un conjunto de definiciones de grupos y perfiles proteicos más detallado que los disponibles en el PFAM (*Protein Families Database*) (Finn et al., 2016). Con el resultado de hallar estas subfamilias se pueden realizar búsquedas con diferentes estrategias como lo son el alineamiento local – Blast (*Basic Local Alignment Search Tool*) y los modelos ocultos de Markov HMMer (*Hidden Markov Models*) (S. R. Eddy, 2009) y hacer hallazgos de manera más precisa de enzimas degradadoras de celulosa y hemicelulosa, las cuales van a permitir hacer degradación de desechos de una manera más económicamente efectiva y ágil en tiempo.

Localmente, los beneficios de realizar esta investigación para el Grupo GIBI en la línea de investigación Aprovechamiento biológico de residuos agroindustriales, serán el de contar con una base de un flujo de trabajo que al hacerse pública a través de artículos y servicios *web,* fortalecerá la producción del grupo y su visibilidad, además de proyectar las investigaciones futuras del grupo, derivadas de este desarrollo investigativo para la ejecución de una segunda versión y también para otros *spin-offs* que potencialmente pueden surgir. Finalmente, otro beneficio es la formación de un profesional a nivel de Maestría en el tema de investigación sobre enzimas celulolíticas y hemicelulolíticas.

1.4. Objetivos

1.4.1. Objetivo general

Reclasificar las enzimas glicosil hidrolasas presentes en la base de datos CAZy con el fin de mejorar la anotación de enzimas degradadoras de celulosa y hemicelulosa.

1.4.2. Objetivos específicos

- Identificar nuevos grupos de secuencias de enzimas glicosil hidrolasas correspondientes a familias y subfamilias de acuerdo con su similitud, a partir de la base de datos pública CAZy.

- Validar *in silico* los nuevos grupos de familias y subfamilias obtenidos, a través de la generación de perfiles proteicos, para su uso posterior en la anotación de secuencias de enzimas celulolíticas y hemicelulolíticas.

- Desarrollar una herramienta para la anotación de secuencias de enzimas degradadoras de celulosa y hemicelulosa.

2. Referente teórico y antecedentes

2.1. Antecedentes

En la necesidad de desarrollar mecanismos que facilitarán la investigación en enzimas funcionales en compuestos orgánicos, surgió la clasificación con los EC *numbers* (*Enzyme Commission Numbers*). Estos corresponden a un identificador de las reacciones que catalizan cada familia de enzimas, el cual se basa en cuatro números separados por puntos en donde cada nivel numérico representa una subclasificación específica funcionalmente (Bairoch, 1994). Más adelante surgió CAZy como un mecanismo de clasificación de enzimas que se limitaba únicamente a los carbohidratos activos, también describe las familias estructural y funcionalmente que degradan, modifican o crean enlaces glicosídicos. El CAZy establece familias y subfamilias que pueden tener múltiples EC *numbers*, reconociendo 5 macrofamilias principales: (i) glicosil hidrolasas implicadas en la degradación de polisacáridos; (ii) glicosil transferasas que tienen función de transporte de metabolitos; (iii) liasas de polisacáridos para la creación de polisacáridos; (iv) esterasas de carbohidratos que sirven para romper enlaces alcohólicos y ácidos y por último (v) las enzimas de actividades auxiliares, siendo las primeras (glicosil hidrolasas) las que ocupan nuestra investigación por su función degradadora.

Alrededor de las clasificaciones establecidas en el CAZy se han desarrollado múltiples herramientas que ayudan a hacer búsqueda sobre familias utilizando diferentes estrategias, siendo las principales herramientas CAT, dbCAN y hotPEP, MycoCLAP cada una abordando el problema de una manera particular. A continuación, se muestra el desarrollo evolutivo de estos proyectos.

2.1.1. CAT (CAZymes Analysis Toolkit)

Publicada en el año 2010 (Park et al., 2010); CAT actualmente en desuso, prestaba un servicio *web* que hacía asignación de familias de CAZy a una secuencia de entrada haciendo uso de algoritmos de búsqueda propios. El modo de uso era mediante una interfaz *web* en donde un usuario subía una secuencia o un conjunto de secuencias, en la que se podían seleccionar algoritmos de anotación y un *threshold* o valor umbral para darle la

sensibilidad al algoritmo. Además, el usuario podía hacer comparaciones con las definiciones de PFAM y las longitudes de las secuencias que se obtenían en el CAZy. El usuario podía también ingresar el correo electrónico para el envío de resultados, además de que el sistema generaba un identificador para almacenar el resultado durante 2 días.

CAT tenía dos modos de uso principalmente:

- Búsqueda de enzimas por grupo taxonómico, tipo de organismo, tipo de enzima y familias del CAZy.
- Búsqueda por asociaciones entre familias, de utilidad para esas proteínas que comparten múltiples definiciones de familias de CAZy, dando por resultado una lista de asociaciones, número de proteínas encontradas y las reglas que derivan dichas asociaciones.

2.1.2. dbCAN

La base de datos dbCAN (*database for Carbohydrate active enzyme ANnotation*) (Yin et al., 2012), hecha pública en el año 2012 y aún vigente, proporciona una clasificación automática de las familias del CAZy, pero a diferencia del CAT, solo se enfoca en la macrofamilia de Glicosil Hidrolasas (en adelante GH, siendo también el principal objetivo de esta investigación), donde para cada familia GH de CAZy, dbCAN extrae dominios principales y crea un modelo HMM para representarlos.

dbCAN no contiene todas las familias de GH, y nunca se ha realizado una evaluación general del rendimiento del método HMM en la recuperación de la clasificación estándar de GH implementada en CAZy (Rossi, Mello, & Schrago, 2017).

Para su utilización al igual que la mayoría de las bases de datos de proteínas públicas, como Pfam (*Protein families*) y CDD (*Conserved Domain Database*), un usuario de dbCAN tiene disponible su servicio a través de una página *web* pública (Figura 1) y permite enviar ejecuciones, también se pueden descargar los archivos HMM y ejecutar el comando *hmmsearch* contra un *set* de proteínas o genomas de manera local.

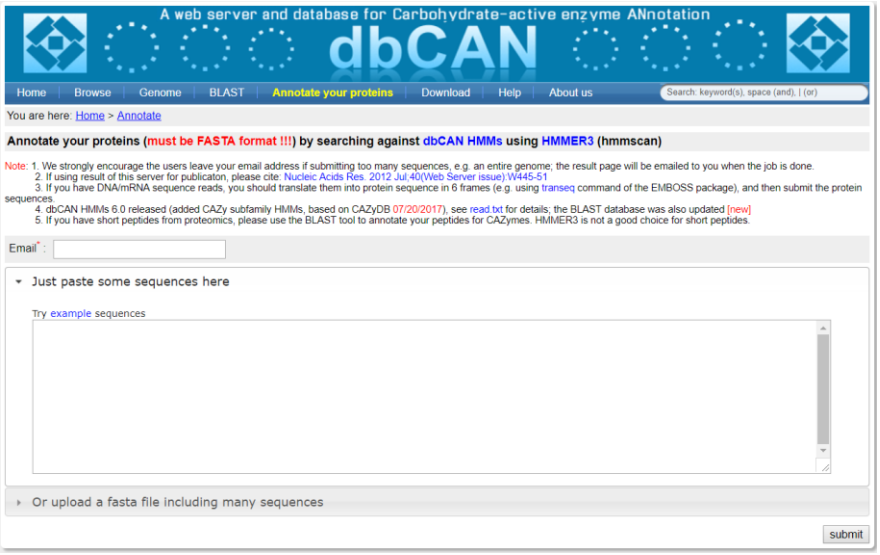


Figura 1. Servicio web del proyecto dbCAN.

Datos importantes de dbCAN:

- Por resultado muestra información de las ubicaciones de los dominios CAZy identificados y un diagrama de la arquitectura de los dominios (Yin et al., 2012).
- Proporciona alineaciones de secuencias, HMM y filogenias de los dominios representativos en todas las familias de CAenZimas
- Tiene otras utilidades como navegación basada en familias CAenZimas, exploración basada en genoma, palabra clave de búsqueda, búsqueda Blast así como una anotación funcional detallada para cada secuencia incluida en dbCAN.

2.1.3. HotPEP

La herramienta Hotpep (abreviatura de *Homology to Peptide*) (P. K. Busk et al., 2017) realiza anotaciones automáticas de enzimas activas en carbohidratos con una alta tasa de éxito. El resultado de la anotación con Hotpep es comparable a la anotación automática con HMM. Además, HotPEP proporciona una predicción de la función directamente a partir de la secuencia de aminoácidos. HotPEP está disponible como un *script* independiente que se ejecuta en el sistema operativo *Windows*.

Al buscar homología con patrones de péptidos con HotPEP, se puede investigar si una secuencia de proteína contiene un número suficiente de péptidos conservados de una

subfamilia específica generada por PPR (*Peptide Pattern Recognition*) para ser considerada como un miembro de esta subfamilia. Además, si se ha caracterizado experimentalmente un número suficiente de las proteínas utilizadas para generar la subfamilia de PPR y tienen la misma función, se puede predecir que la nueva proteína tiene la misma función (Peter K. Busk et al., 2014). Este método para predecir la función proteica ha demostrado que predice correctamente la función de aproximadamente 80-90% de las proteínas en familias de glicosil hidrolasa seleccionadas.

2.1.4. MycoCLAP

MycoCLAP es una base de datos que facilita la anotación funcional de enzimas degradadoras de biomasa de plantas. Contiene una selección de genes caracterizados que codifican enzimas de glicosil hidrolasas, liasas de polisacáridos, esterasas de carbohidratos y enzimas con actividades auxiliares para hongos. Se trata de una base de datos que es manualmente curada y se basa en la evidencia reportada en la literatura. Su principal objetivo es proveer de un listado de proteínas caracterizadas para facilitar la anotación funcional de nuevas proteínas degradadoras de lignocelulosa. Las últimas actualizaciones también incluyen genes de bacterias.

***myco*CLAP - Characterized Lignocellulose-Active Proteins of Fungal Origin**

Home | Search | Downloads | Data Summary | Correction | New Entry | BLAST | Useful Links | Internal User (Login) | About Us | Help | Ter

Search results of **'Xylanase on all'**

Entry Name	Species	Taxonomic Domain	Enzyme Name	CAZy Family	Host For Recombinant Expression	Substrates
ABF51A_PLEOS	Pleurotus ostreatus	Eukarya	alpha-arabinofuranosidase	GH51	Pichia pastoris X33	1,5-alpha-arabinotriose
ABF51C_CHRLU	Chrysosporium lucknowense	Eukarya	alpha-arabinofuranosidase	GH51	native	Single and double substituted arabinoxylans.
ABF54A_AURPU	Aureobasidium pullulans	Eukarya	alpha-arabinofuranosidase	GH54	Saccharomyces cerevisiae Y294 (ATCC 201160)	Corn fiber arabinoxylan.
AGU67A_ASPTU	Aspergillus tubingensis	Eukarya	alpha-glucuronidase	GH67	native	Aldotriouronic acid-aldobiuronic acid.
AXE1A_ASPOR	Aspergillus oryzae	Eukarya	acetylxylan esterase	CE1	Pichia pastoris SMD1168H	alpha-naphthyl propionate
AXE6A_ORPSP	Orpinomyces sp. PC-2	Eukarya	acetylxylan esterase	CE6	Escherichia coli XL-1 Blue	glucose pentaacetate
AXH43G_CHRLU	Chrysosporium lucknowense	Eukarya	arabinoxylan arabinofuranohydrolase	GH43	native	Singly and doubly substituted arabinoxylans.
CBH7A_PENFN	Penicillium funiculosum	Eukarya	cellobiohydrolase	GH7	native	carboxymethyl cellulose
EGL5D_NEOPA	Neocallimastix patriciarum	Eukarya	endoglucanase	GH5	Escherichia coli XL1-Blue	carboxymethyl cellulose

Figura 2 Búsqueda del término *Xylanase* en la base de datos MycoCLAP.
Fuente: (https://mycoclap.fungalgenomics.ca/mycoCLAP/).

MycoCLAP usa un algoritmo de inteligencia artificial para hacer búsquedas en bases de datos de artículos científicos como *Pubmed* y *Google scholar*, y de otros recursos *online* como la base de datos del *CAZy* y *Brenda*. Tras la recolección, los artículos son revisados por un grupo de curadores (Strasser et al., 2015). Cada entrada de MycoCLAP tiene relacionada el *Enzyme Commission Number* (EC), la función molecular, el proceso biológico y la ubicación subcelular basado en la literatura.

La interfaz *web* tiene tres funciones principales que incluye la búsqueda de enzimas caracterizadas, la obtención de datos o secuencias y el uso de Blast para comparar una secuencia de interés con las secuencias publicadas en MycoCLAP (Strasser et al., 2015).

2.1.5. Macroproyecto "Construcción de un dataset de familias de enzimas lignocelulolíticas para su anotación y minería de datos de proyectos de las ómicas"

El macroproyecto titulado "Construcción de un *dataset* de familias de enzimas lignocelulolíticas para su anotación y minería de datos de proyectos de las ómicas" de la Universidad Católica de Manizales al cual pertenece este proyecto, tiene por objetivo construir un *dataset* de familias de enzimas involucradas en el proceso de degradación de la lignocelulosa para facilitar la anotación y minería de datos provenientes de proyectos de ómicas. Solo se consideraron las enzimas de la clase Glicosil Hidrolasa (GH), donde los datos de los identificadores de las secuencias fueron recolectados manualmente por estudiantes de semillero de investigación. Los datos fueron tomados directamente de la página *web* del CAZy en archivos individuales segmentados por familia y por grupo taxonómico según está dividido en dicho servicio (por ejemplo: GH1_Bacteria), y posteriormente se obtuvieron las secuencias correspondientes a estos identificadores haciendo uso de la herramienta *Batch Entrez*, puestos luego los archivos en tres carpetas diferentes para las tres bases de datos a trabajar PDB, Uniprot y Genbank.

Las subclasificaciones descritas por el CAZy no fueron tenidas en cuenta al momento de descargar los datos, al igual que no todas las familias, ni todos los grupos taxonómicos para cada base de datos, ni tampoco se recolectaron todas las secuencias por familia.

Las herramientas dbCAN y HotPEP (anteriormente mencionadas y que continúan funcionando actualmente), realizan el análisis sobre el tipo de enzima GH. La GH como su nombre lo indica cumple la función de hidrolizar enlaces glicosídicos para descomponer

compuestos en moléculas más pequeñas. Compuestos tales como la celulosa y hemicelulosa, son degradados por las GH, siendo las enzimas de mayor interés para esta investigación.

Esta tesis comparte similitudes con los cuatro antecedentes mencionados, debido a que también se basa en la familia GH del CAZy y permite hacer búsquedas de enzimas a partir de ellas. Sin embargo, el enfoque diferencial de este proyecto consiste principalmente en la inclusión de secuencias de base de datos públicas conjuntamente con las secuencias del CAZy, el procesamiento segmentado por subfamilia del CAZy - especie, la salida de nuevos subgrupos basados en clusterización del OrthoMCL, y finalmente su uso para búsqueda de enzimas celulolíticas y hemicelulolíticas.

Como características principales del proyecto se tienen:

- Búsqueda de enzimas por grupo taxonómico, tipo de enzima y familias del CAZy.
- *Script* de búsqueda *standalone* a partir de alineamientos con *blast* y modelos ocultos de Markov con *hmmer.*
- Herramienta de consulta *web* para investigación del CAZy y de las nuevas subfamilias del proyecto.
- Definiciones de cada familia en donde puede encontrar la el árbol filogenético y archivos de secuencias y búsqueda relacionados.
- Muestra las relaciones entre las familias del CAZy y las nuevas subfamilias.
- Filtro para búsqueda y anotación de enzimas degradadoras de celulosa y hemicelulosa específicamente.

Del resultado final del proyecto derivarán más investigaciones que incluyan análisis por dominios y *motifs* que mejorarán la caracterización de cada familia. Todos estos conceptos necesarios para el proyecto, se describen a continuación para su mayor profundización.

2.2. Referente teórico

2.2.1. Complejo lignocelulósico

La lignocelulosa es una estructura compleja que contiene celulosa (un polímero homólogo de glucosas conectado por enlaces β-1,4), hemicelulosa (un polímero heterólogo

de pentosas y hexosas) y lignina (un polímero aromático complejo). La composición varía en porcentaje según sea el tipo de vegetal (Tabla 1). Este componente de las plantas sirve para dar estructura y está ubicado mayoritariamente en la pared celular. Para su conversión energética, estos polisacáridos deben ser degradados a moléculas mucho más pequeñas, pero las plantas no cuentan con el material enzimático para esta labor. Existen microorganismos como bacterias u hongos que poseen enzimas con la capacidad de tomar la celulosa y degradarla a celobiosa o glucosa, y al mismo tiempo, tomar la hemicelulosa e hidrolizarla en pentosas como la xilobiosa o xilosa, entre otras (Agustini et al., 2012).

Tabla 1. Porcentaje de composición de lignocelulosa.

Fuente	Celulosa (%)	Hemicelulosa (%)	Lignina (%)
Madera dura	40-55	24-40	18-25
Madera blanda	45-50	25-35	25-35
Cáscaras de nuez	25-30	25-30	30-40
Mazorcas de maíz	45	35	15
Papel	85-99	0	0-15
Paja de trigo	30	50	15
Paja de arroz	32.1	24	18
Basura clasificada	60	20	20
Hojas	15-20	80-85	0
Pelos de semillas de algodón	80-95	-	0
Periódico	40-55	25-40	18-30
Papel usado de pulpas químicas	60-70	-	-
Bagazo fresco	33.4	30	18.9
Residuos porcinos	6	28	NA
Estiércol de ganado sólido	1.6-4.7	1.4-3.3	2.7-5.7
Pasto alto	45	31.4	12

Fuente: Cao & Sheng (2016).

Cada uno de los polímeros que componen el complejo lignocelulósico tienen características químicas y estructurales distintas, por lo que serán tratadas con un mayor detalle, a continuación.

2.2.2. Celulosa

La celulosa es el compuesto orgánico más abundante en la Tierra. Cada año, las plantas producen más de 10^{11} toneladas métricas de celulosa (Glaser & Nikaido, 2007).

Químicamente se define como una cadena de glucosa unidas por enlaces β-(1-4)–glicosídicos (Figura 3). Su fórmula química es $(C_6H_{10}O_5)_n$, donde n es el grado de polimerización, que puede ser de 10.000 para la cadena de celulosa en la naturaleza y de 15.000 para celulosa nativa de algodón (Ummartyotin et al., 2015). La unidad repetitiva básica es la celobiosa, un disacárido unido por enlaces β-(1-4) de glucosa. El grupo hidroxilo de una cadena puede formar puentes de hidrógeno con el oxígeno de otra cadena, lo que dota de rigidez a la molécula para formar microfibras (Kubicek et al., 2012). La cadena del polímero de celulosa tiene una estructura plana tipo cinta estabilizada por enlaces de hidrógeno internos. Sus propiedades químicas son hidrofilia, biodegradabilidad, insolubilidad dado a sus puentes de hidrógeno y es más resistente a la degradación que otros polímeros de glucosa, como el almidón (Gupta et al., 2016). Se caracteriza porque el terminal del carbono 1 tiene propiedades reductoras y el grupo hidroxilo del extremo del carbono 4, tiene muestra propiedades no-reductoras.

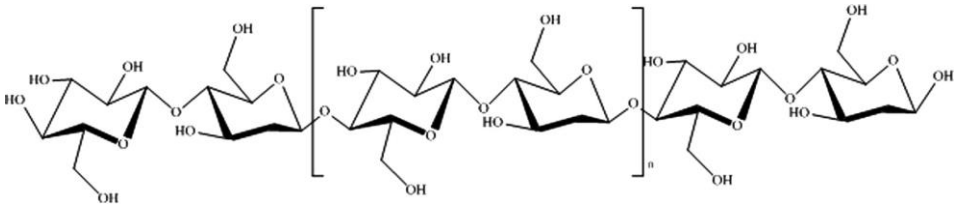


Figura 3. Estructura química de la celulosa, las líneas punteadas son enlaces de hidrogeno.
Fuente: Ummartyotin et al. (2015).

Existen 7 tipos de polimorfos de celulosa (I_a, I_b, II, III_1, III_{11}, IV_1, IV_{11}) donde la celulosa I es la forma encontrada en la naturaleza, y el resto son modificaciones para el pretratamiento de la biomasa (Kubicek et al., 2012).

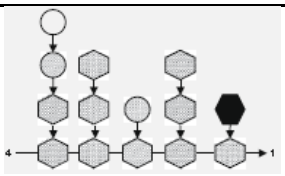
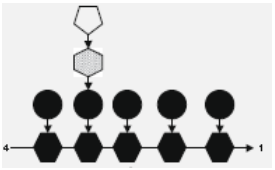
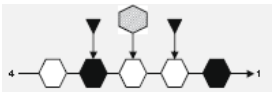
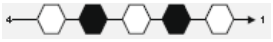
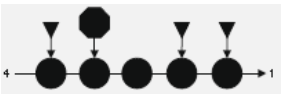
La celulosa presenta dos formas, una que es organizada llamada cristalina que compone la mayor parte de la celulosa y otra pequeña parte desorganizada llamada amorfa. La celulosa amorfa es más susceptible a ser degradada enzimáticamente que la cristalina, por lo que la celulosa cristalina necesita ser transformada en celulosa amorfa para que pueda ser hidrolizada efectivamente por celulasas (Cao et al., 2016).

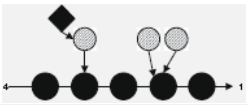
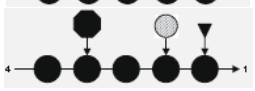
2.2.3. Hemicelulosa

La hemicelulosa es un complejo parecido a la celulosa, pero con un nivel medio de complejidad. Está conformada por diferentes hexosas y pentosas que generalmente están acetiladas y en forma de cadenas (Martínez et al., 2005). A diferencia de la celulosa que presenta una estructura homóloga casi simétrica, las hemicelulosas son heteropolisacáridos altamente ramificados, generalmente no cristalinos, que muestran un grado menor de polimerización (<200 unidades de azúcar) (Glaser & Nikaido, 2007). La hemicelulosa puede tener diferentes tipos de azúcares, sean pentosas (D-xilosa, L-arabinosa), hexosas (D-galactosa, L-galactosa, D-manosa, L-fucosa) o ácidos urónicos (ácido D-glucurónico) (Peng et al., 2011).

Los tipos de hemicelulosa (Tabla 2) varían en su composición química, tipos de enlaces y configuración estructural según sea el tipo de material lignocelulósico, el cual puede ser de madera blanda, madera dura, pastos, cereales o algas.

Tabla 2 Tipos de hemicelulosa.

Tipo de polisacárido	Origen	Cadena principal	Cadenas laterales	Enlaces	Representación
Arabinogalactano	Madera blanda	β-D-galactosa	β-D-galactosa, α-L-arabinosa, β-D-arabinosa	β-(1-6) α-(1-6) β-(1-3)	
Xiloglucano	Madera dura, pastos	β-D-glucosa, β-D-xilosa	β-D-xilosa, β-D-galactosa, α-L-arabinosa, α-L-fucosa, acetil	β-(1-4) α-(1-3) β-(1-2) α-(1-2) α-(1-2)	
Galactoglucomanano	Madera blanda	β-D-manosa, β-D-glucosa	β-D-galactosa, acetil	α-(1-6)	
Glucomanano	Madera dura y blanda	β-D-manosa, β-D-glucosa	-		
Glucuronoxilano	Madera dura	β-D-xilosa	4-O-Metil-α-D-gluco-β-acetil	α-(1-2)	
Arabinoglucuronoxilano	Pastos, cereales y	β-D-xilosa	4-O-Metil-α-D-gluco-β-arabinosa	α-(1-2) α-(1-3)	

	madera blanda				
Arabinoxilano	Cereales	β-D-xilosa	α-L-arabinosa	α-(1-2) α-(1-3)	
Glucuronoarabinoxilano	Pastos y cereales	β-D-xilosa	α-L-arabinosa, 4-O-Metil-α-D-glucosa	α-(1-2) α-(1-3)	
Homoxilano	Algas	β-D-xilosa	-		

Modificado de Gírio et al. (2010).
Convenciones: β-D-galactosa, β-D-glucosa, β-D-manosa, β-D-xilosa, β-L-arabinosa, β-L-arabinosa, α-L-fucosa, 4-O-Metil- α-D-gluco, acetil.

Las hemicelulosas se pueden agrupar en tres grandes grupos: mananos, xilanos y xiloglucanos, los cuales se describen a continuación.

2.2.3.1. Manano

Los mananos se pueden agrupar en cuatro clases distintas: manano, glucomanano, galactomanano y galactoglucomanano. Donde manano y galactomanano tienen en la cadena principal solo enlaces β-1-4 de manosa, mientras que para glucomanano y galactoglucomanano contienen unidades de manosa y glucosa enlazadas en una configuración β-1-4. Para el galactomanano y el galactoglucomanano el residuo de manosa se cambia por un residuo de galactosa (Pauly et al., 2013).

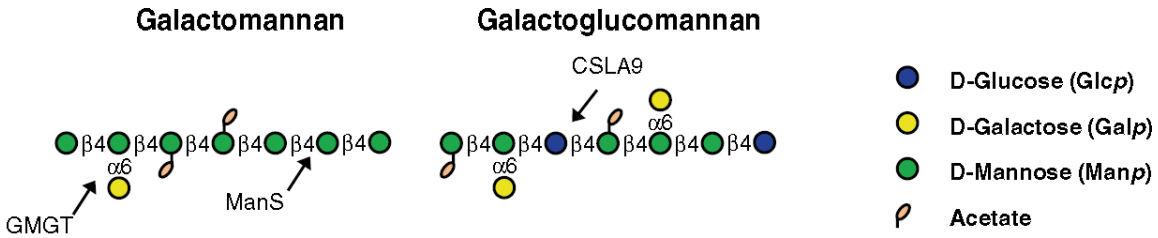


Figura 4 Mananos tipo galactomanano y galactoglucomanano.
Fuente: Pauly et al. (2013).

Los galactoglucomananos, son las hemicelulosas principales en madera blanda que ocupan el 20 al 25% de su masa (Gírio et al., 2010). Tiene una columna principal de D-glucosa y D-manosa unidos por enlaces β-1-4, y residuos laterales de acetil y galactosil unidos por enlaces α-(1,6) ya sea a la manosa o la glucosa. Algunos pueden ser solubles en agua dado a que tienen una mayor cantidad de galactosa, unidad hidrofílica (Pauly et

al., 2013). Tiene una mayor cantidad de manosa que de glucosa. El galactomanano posee una configuración similar, pero con la diferencia de que su cadena principal no tiene glucosa.

Las distribuciones de azúcares para glucomanano, galactoglucomanano cambian según sea el tipo de manano, para el glucomanano la proporción de manosa es mayor a que en el galactoglucomanano, como se muestra en la Tabla 3.

Tabla 3. Proporción de azúcares para las hemicelulosas de madera blanda.

Fuente	Glucomanano	Galactoglucomanano
Galactosa	0,1	1,0
Glucosa	1,0	1,0
Manosa	4,0	3,0

Fuente: Glaser & Nikaido (2007).

2.2.3.2. Xilano

Los xilanos son abundantes en cereales. Consiste en una cadena principal de xilosa unida por enlaces β-1-4, que puede ser cambiada dependiendo del tipo de material lignocelulósico (Pauly et al., 2013).

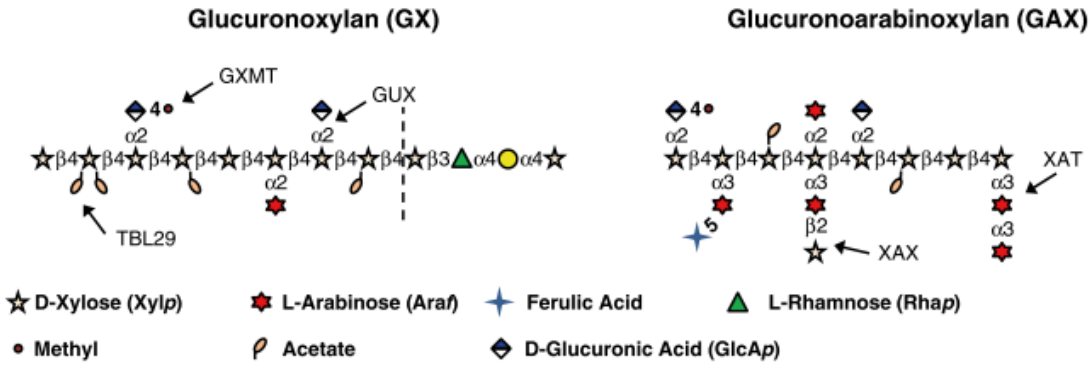


Figura 5. TIpos de xilano, glucuronoxilano y glucuronoarabinoxilano.
Fuente: Pauly et al. (2013).

La mayor hemicelulosa de madera dura es el glucuronoxilano, el cual puede estar representado entre el 15% y 30% dependiendo de la especie de planta (Peng et al., 2011), con una cadena principal de unidades de β-d-xilosa unidas con enlaces 1-4, la mayoría de las cuales están acetiladas. Hay residuos de ácido glucurónico o su equivalente metilado 4-O-metil-α-d-ácido glucurónico unidos a la base, que están presente cada 10 unidades de xilosa (Glazer et al., 2007).

Los arabinoxilanos son similares a los glucuronoxilanos pero la cantidad de L-arabinosa es mucho mayor. La α-L-arabinosa se une lateralmente al oxígeno de la posición 2 o 3 de la xilosa y el ácido glucurónico también puede estar unido al oxígeno 2 (Gírio et al., 2010).

2.2.3.3. Xiloglucano

El xiloglucano es el tipo de hemicelulosa más típicamente encontrada en madera dura. Su cadena principal consiste enlaces β-1-4 de D-glucosa (Figura 6), donde el 75% de los residuos laterales que se unen a la glucosa en la posición del oxígeno 6, de su estructura hexagonal son xilosa, de donde a su vez, se le pueden añadir D-galactosa o L-arabinosa (Peng et al., 2011).

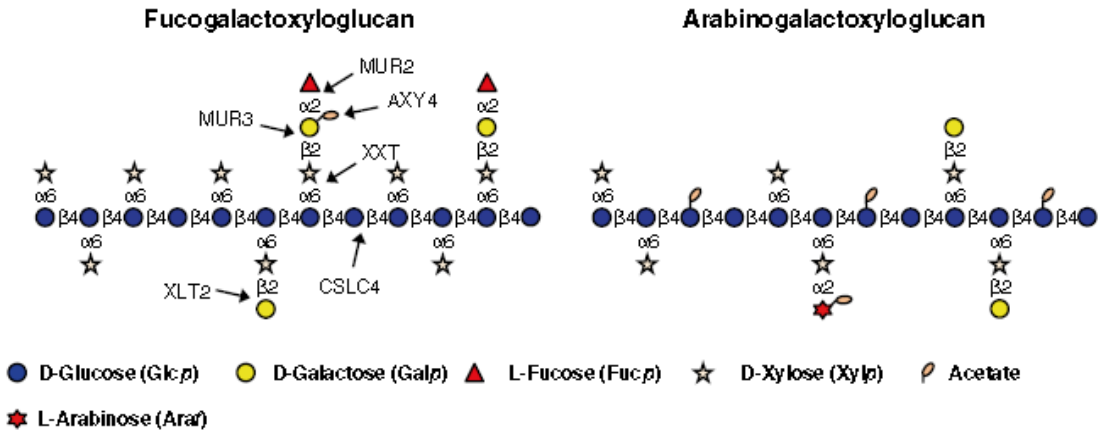


Figura 6. Tipos de xiloglucano, fucogalactoxiloglucano y arabinogalactoxiloglucano.
Fuente: Pauly et al. (2013).

En ocasiones se puede añadir una unidad de fucosa a una galactosa (o a su equivalente acetilado) lateral en la posición de su oxígeno 2 de su estructura hexagonal. A esta formación se le llama fucogalactoxiloglucano (Pauly et al., 2013). También existen formaciones donde la cadena principal tiene combinaciones de D-glucosa y su versión acetilada, y donde lateralmente se adhiere L-arabinosa a los residuos de xilosa, a esta forma se le llama arabinogalactoxiloglucano.

2.2.4. Lignina

La lignina se encuentra en las paredes celulares de las plantas espermatocitas (gimnospermas y angiospermas), helechos y musgos, predominantemente en los tejidos vasculares especializados para el transporte de líquidos. No se encuentra en los musgos, líquenes y algas que no tienen traqueidas (células tubulares largas propias de la xilema).

La lignina peroxidasa y la manganeso peroxidasa, son las enzimas degradadoras de lignina, que son más antiguamente conocidas (Kubicek et al., 2012). La Tabla 4 muestra todos los tipos de enzimas que intervienen en la degradación de lignina.

Tabla 4. Enzimas que intervienen en la degradación de lignina.

Fuente	EC Number	Sustratos
Lignina peroxidasa	1.11.1.14	Aromáticos no-fenólicos
Manganeso peroxidasa	1.11.1.13	Iones de manganeso
Coprinopsis cinerea peroxidasa	1.11.1.7	Compuestos fenólicos
Cloroperoxidasa	1.11.1.10	Halogenuros y compuestos fenólicos
Peroxigenasa aromática	1.11.-	Aromáticos y n-alcanos
Peroxidasa decolorante	1.11.-	Colorantes y fenol

Fuente: Kubicek et al., (2012).

2.3. Degradación de lignocelulosa

La degradación de lignocelulosa es realizada por diferentes tipos de microorganismos, principalmente hongos y bacterias. En el reino fungi, este proceso es realizado por los llamados hongos de la pudrición blanca, que son mayormente de la clase *basidiomiceto* caracterizados por su habilidad de degradar lignina, hemicelulosa y celulosa (Martínez et al., 2005). El nombre de pudrición blanca está asociado con la forma en que la madera toma su coloración tras la degradación. Es el único mecanismo que puede degradar completamente la estructura de la madera porque es capaz de romper los enlaces presentes en la lignina.

La pudrición café es el tipo de degradación que se da mayormente en madera blanda. Según Martínez et al. (2005), solo el 7% de los basidiomicetos de pudrición realizan este proceso. En este mecanismo de pudrición se degrada la celulosa y hemicelulosa de la madera, y mínimamente la lignina, cuya acumulación finalmente les da la coloración café a los residuos de madera (Kubicek et al., 2012).

La última es la pudrición blanda que se da por ascomicetos y basidiomicetos, esta actúa sobre madera de bajo contenido de lignina. Requiere mayor humedad y contacto directo con el suelo. La Tabla 5 muestra un resumen de las características de los tipos de degradación por hongos.

Tabla 5. Características principales de los tipos de degradación de la lignocelulosa, mediada por hongos.

	Pudrición blanca	**Pudrición café**	**Pudrición blanda**
Apariencia	Blanqueado, esponjoso	Café, frágil, polvoriento.	Suave, mojado, café y desmoronizado
Tipo de madera	Madera dura	Madera blanda	Mayormente madera dura
Tipo de polímero que ataca	Celulosa, hemicelulosa y lignina	Celulosa, hemicelulosa y en poco grado lignina	Celulosa, hemicelulosa y en poco grado lignina
Hongos	Basidiomicetos (e.g. *Trametes versicolor, Irpex lacteus, Phanerochaete chrysosporium y Heterobasidium annosum*) y algunos Ascomicetos (e.g. *Xylaria hypoxylon*). Basidiomicetos (e.g. *Ganoderma australe, Phlebia tremellosa, C. subvermispora, Pleurotus spp. y Phellinus pini*).	Basidiomicetos exclusivamente (e.g. *C. puteana, Gloeophyllum trabeum Laetiporus sulphureus, Piptoporus betulinus, Postia placenta* y *Serpula lacrimans*).	Ascomicetos (*Chaetomium globosum, Ustulina deusta*) y Deuteromicetos (*Alternaria alternata, Thielavia terrestris, Paecilomyces spp*.), y algunas bacterias. Algunos basidiomicetos de pudrición blanca (*Inonotus hispidus*) y pudrición café (*Rigidoporus crocatus*) causan pudrición blanda.

Fuente: Modificado de Martínez et al. (2005).

En bacterias la degradación de celulosa cristalina se realiza por bacterias aeróbicas y anaeróbicas. La endoglucanasa y exoglucanasa extracelular en bacterias muestran la misma estructura de la vista en los hongos Ascomicetos como el *Trichoderma reseei* (Glaser & Nikaido, 2007). Las celulasas son secretadas como enzimas solubles extracelularmente y en bacterias anaerobias son ensambladas en complejos llamados celulosomas que se añaden a su superficie celular. Estos celulosomas también contienen enzimas para degradar hemicelulasas como xilanasas, mananasas, arabinofuranosidasas (Glaser & Nikaido, 2007).

Estos microorganismos contienen el material enzimático para la degradación de la celulosa, hemicelulosa y lignina, la Tabla 6 resume de manera general las funciones

enzimáticas necesarias que intervienen en la descomposición de la lignocelulosa. En el siguiente aparte se tocan cada una de estas en mayor detalle.

Tabla 6. Enzimas implicadas en la degradación de lignocelulosa.

Enzima	Tipo	Sustratos sobre los que actúa
Endoglucanasas	Celulolíticas	Carboximetilcelulosa, celobiosa amorfa y celo-oligosacáridos
Exoglucanasas: Exocelobiohidrolasas Celobiohidrolasas	Celulolíticas	Celulosa
Celobiasas: β-glucosidasas	Celulolíticas	Celobiosa
Xilanasas: Endoxilanasas Xilosidasas	hemicelulolíticas	Xilano
Mananasas	hemicelulolíticas	Mananos
Ligninperoxidasa (LiP)	Ligninolíticas	Sustratos aromáticos fenólicos y no fenólicos
Manganesoperoxidasa (MnP)	Ligninolíticas	Sustratos aromáticos fenólicos y no fenólicos
Lacasa	Ligninolíticas	Sustratos aromáticos fenólicos y no fenólicos
Peroxidasa versátil (PV)	Ligninolíticas	Compuestos no fenólicos
Peroxidasa manganeso independiente	Ligninolíticas	

Fuente: Recolectado de Brink & Vries (2011).

2.3.1. Degradación de lignocelulosa y las glicosil hidrolasas

Las glicosil hidrolasas (GH) son un amplio grupo de enzimas que hidrolizan el enlace glicosídico entre dos o más carbohidratos o entre un carbohidrato y una fracción no carbohidratada. Son estudiados extensivamente debido a su papel en la degradación de celulosa, hemicelulosa y lignina (Rossi et al., 2017).

La función enzimática de las glicosil hidrolasas se basa en su especificidad de sustrato y ocasionalmente en su mecanismo molecular; dicha clasificación no refleja las características estructurales de estas enzimas. Se propuso una clasificación de glucósidos de hidrolasas en familias, basada en similitudes de secuencias de aminoácidos propuesta por Henrissat et al., (1991), debido a que existe una relación directa entre secuencia y similitudes de plegado, y una relación evolutiva que la anterior clasificación no mostraba.

Genes codificantes de GH abundan en la mayoría de genomas, este tipo de enzimas está organizado en 113 familias, las cuales son responsables de la hidrólisis y

transglicosilación de enlaces glicosídicos (Cantarel et al., 2009). Las descripciones del total de familias de enzimas glicosil hidrolasas se encuentran en el Anexo I.

Este trabajo incluye la descripción de las glicosil hidrolasas por su papel de hidrolizar enlaces glicosídicos como los presentes en la celulosa y los presentes en la hemicelulosa donde estas degradan la mayoría de enlaces a excepción de dos tipos los cuales son degradados por esterasas en la hemicelulosa de tipo xilano, siempre siendo estas GH las más representativas de todas. La Tabla 7 relaciona las funciones enzimáticas con los grupos de GH necesarios en la degradación de celulosa y hemicelulosa.

Tabla 7. Relación de enzimas degradadoras de celulosa y hemicelulosa con familias GH.

Sustrato	Actividad enzimática	Familia
Celulosa	β-1,4-endoglucanasa (EC 3.2.1.4)	GH5, GH7, GH12, GH45
	Celobiohidrolasa (EC 3.2.1.91)	GH6, GH7
	β-1,4-glucosidasa (EC 3.2.1.21)	GH1, GH3
Xiloglucano (Hemicelulosa)	Xiloglucano β-1,4-endoglucanasa (Xiloglucanasa) (EC 3.2.1.151)	GH12, GH74
	α-arabinofuranosidasa (EC 3.2.1.55)	GH51, GH54
	α-xilosidasa (EC 3.2.1.177)	GH31
	α-fucosidasa (EC 3.2.1.51)	GH29, GH95
	α-1,4-galactosidasa (EC 3.2.1.22)	GH27, GH36
Xilano (Hemicelulosa)	β-1,4-endoxilanasa (EC 3.2.1.8)	GH10, GH11
	β-1,4-xilosidasa (EC 3.2.1.37)	GH3, GH43
	α-arabinofuranosidasa (EC 3.2.1.55)	GH51, GH54
	Arabinoxilano arabinofuranohidrolasa (EC 3.2.1.55)	GH62
	α-glucuronidasa (EC 3.2.1.131)	GH67, GH115
	α-1,4-galactosidasa (EC 3.2.1.22)	GH27, GH36
Galactomanano (Hemicelulosa)	β-1,4-endomananasa (EC 3.2.1.78)	GH5, GH26
	β-1,4-manosidasa (EC 3.2.1.25)	GH2
	β-1,4-galactosidasa (EC 3.2.1.23)	GH2, GH35
	α-1,4-galactosidasa (EC 3.2.1.22)	GH27, GH36

Fuente: Modificado de Brink & Vries (2011).

Las enzimas fúngicas involucradas en la degradación de lignocelulosa incluyen 35 familias de glicosil hidrolasas, 3 esterasas y 6 liasas de polisacáridos (Brink & Vries, 2011).

2.3.2. Degradación de la celulosa

Los tres tipos de enzimas que degradan celulosa son la endoglucanasa (EGL), la celobiohidrolasa (CBH) (exoglucanasa) y la β-glucosidasa (BGL) (Ver Figura 7). La celobiohidrolasa (CBH) efectúa la hidrólisis de las partes cristalinas de la celulosa y la endoglucanasa (EGL) se dedica a las partes internas más amorfas, después de que la endo y exo glucanasa hayan actuado, las β-glucosidasa (BGL) actúan sobre los oligosacáridos restantes quedando los monómeros de glucosa (Brink & Vries, 2011).

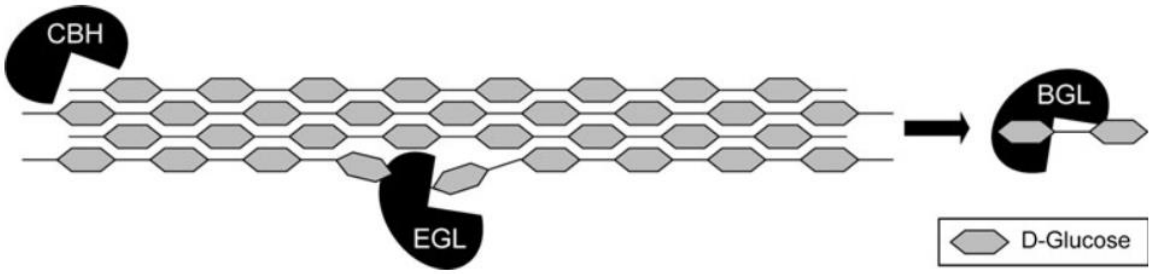


Figura 7. Esquema de degradación de celulosa.
Fuente: Brink & Vries (2011).

2.3.2.1. *Exoglucanasa o Celobiohidrolasa (CBH) (EC 3.2.1.91)*

La celobiohidrolasa es, en los cultivos de hongos, la responsable de la mayor parte de la hidrólisis de celulosa (Yao et al., 2018). La celobiohidrolasa libera celobiosa ya sea por los terminales reductores o no reductores de la cadena de celulosa (Manavalan et al., 2015), y degrada la celulosa cristalina, presumiblemente por los extremos de la cadena. La mayoría de CBH liberan celobiosa y una pequeña cantidad de glucosa a partir de la celulosa. De acuerdo con su función se denominan exo-1,4-β-glucanasa (EC 3.2.1.91). Existen dos tipos de celobiohidrolasas de tipo fúngico, que están categorizadas en las familias glicosil hidrolasas GH6 y GH7 (Peter K. Busk et al., 2014).

La primera celobiohidrolasa GH7 caracterizada y con estructura 3D fue la del hongo *Trichoderma reesei*. Se han encontrado proteínas ortólogas en todos los ascomicetos y en los genomas de basidiomicetos en pudrición blanca. El dominio catalítico de este tipo de celobiohidrolasa consiste en una formación de rollo, en donde dos hojas beta antiparalelas se empaquetan cara a cara para formar un β-sandwich curveado. Contiene entonces un túnel donde caben aproximadamente 10 subsitios para degradación de celulosa (Kubicek

et al., 2012). Una vez insertada, la celulosa es fragmentada, generando celobiosa. Estas actúan en los extremos reductores de la cadena de celulosa.

Las celobiohidrolasas de la familia GH6 pueden ser encontradas en todos los hongos (excepto los de pudrición café) y bacterias. Actúan removiendo la celobiosa de los terminales no-reductores e invierte la anomería química. Los enlaces β-1-4-glicosidicos son rotos por catálisis ácida usando ácido aspártico como el donador de protón (Kubicek et al., 2012). Como la formación del sitio activo de este tipo de enzima es flexible, la enzima puede cambiar entre exo y endo, dependiendo de la forma de este.

2.3.2.2. *Endoglucanasas (EC 3.2.1.4)*

La endoglucanasa separa aleatoriamente los enlaces internos de las regiones amorfas de la celulasa y las convierte en disacárido celobiosa directamente o en oligosacáridos de menor tamaño, donde posteriormente la celobiohidrolasa sigue degradando. De acuerdo con su función se denomina endo-1,4-β-glucanasa (EC 3.2.1.4).

Solo una pequeña cantidad de endoglucanasas han sido aisladas y caracterizadas parcialmente en hongos como el *Agaricus bisporus, Agaricus brasilensis, Armillaria gemina, Dichomitus squalens, Irpex lacteus, Phanerochaete chrysosporium, Polyporus arcularius, Polyporus arcularius, Polyporus schweinitzii, Trametes versicolor, Volvariella volvacea,* y *Schizophyllum commune* (Manavalan et al., 2015). Las endo-β-1,4-glucanasas de interés están categorizadas en las familias GH5, GH7, GH12, GH45.

La familia GH5 actúa sobre múltiples tipos de sustratos. La mayoría son endo-β-1,4-glucanasas y endo-β-1,4-mananasas, pero también se incluyen otras actividades. La estructura 3D de esta celulasa solo ha sido resuelta para bacterias (*Clostridium thermocellum endoglucanase CelC y Bacillus agaradhaerens Cel5A*), tienen un plegado tipo barril TIM (8 hélices alfa y 8 láminas beta intercaladas) (Kubicek et al., 2012). Son enzimas retenedoras.

Las endo-β-1,4-glucanasas de la familia GH7 son producidas mayormente por hongos. Son similares a las celobiohidrolasas GH7 en estructura (mostrando una estructura de rollo) y en que atacan los terminales reductores por un mecanismo de retención, la mayor diferencia es que la última tiene un hoyo para enlace de sustrato en vez de un túnel, lo que habilita un ataque desde el medio a la celulosa (Kubicek et al., 2012).

Las endo-β-1,4-glucanasas de la familia GH12 tienen miembros de bacterias y hongos con proteínas relativamente pequeñas que no tienen *dominio de unión de celulosa* porque

lo que no son capaces de hacer unión con la celulosa cristalina y solo hidrolizan celulosa amorfa. Se subdivide en 5 subgrupos. Enzimas de *Aspergillus niger* y *Malbrachea cinnamomea* del subgrupo GH12-2 tienen función xiloglucanasa (Rawat et al., 2015). Exhiben una estructura de β-sandwich que se curva para crear un sitio de enlace de celulosa (Kubicek et al., 2012).

La primera endoglucanasa GH45 fue de *H. insolens* y *Trichoderma Reesei* EGV. Son las celulasas más pequeñas (242 aminoácidos), a diferencia con la GH12, contiene *dominio de unión de celulosa* (CBD). Consisten en un dominio en forma de barril beta (Kubicek et al., 2012). Las endo-β-1,4-Glucanasas GH45 se conocen en diferentes microorganismos como bacteria, hongos, insectos y mariscos. Son capaces de hidrolizar enlaces β-1,3/1,4 por el mecanismo de inversión. Su pH optimo se acerca al neutro, y por esto se investigan por su aplicación en la industrial textil y de detergentes (Cha et al., 2018).

2.3.2.3. *Celobiasas o β-glucosidasas (GH) (EC 3.2.1.21)*

Tienen la capacidad de dividir los disacáridos (celobiosa) o los gluco-oligosacáridos resultantes de la glucosa. De acuerdo con su función se nominan β-1,4-glucosidasa (EC 3.2.1.21). Las β-1,4-Glucosidasas, son producidas por todos los hongos y están presentes en las familias GH1 y GH3, con menor representación en familias GH 2, 5, 9, 30, 39 y 116 (Salgado et al., 2018).

La familia GH1 (EC 3.2.1.21) consiste de β-glucosidasas bien caracterizadas, que son usualmente implementadas en los procesos de valorización de biomasa (Jin et al., 2018). La mayoría de β-Glucosidasas GH1 son enzimas intracelulares. Su estructura 3D consiste en una estructura de barril TIM α/β, donde cada pliegue contiene un sitio activo tipo hendidura (Kubicek et al., 2012). Son enzimas retenedoras y son capaces de romper enlaces beta de oligosacáridos de cadenas de hasta 9 glucosas. La mayoría tiene función β-Glucosidasa y β-galactosidasa. Se inhiben por la misma glucosa resultante.

Las β-D-glucosidasas GH3 son exo hidrolasas que eliminan residuos de terminales no reductoras de polisacáridos. Comprenden la mayoría de β-Glucosidasas extracelulares de tipo fúngica.

2.3.3. Degradación de hemicelulosa

La degradación de hemicelulosa requiere de una cantidad de enzimas más amplia que la degradación de la celulosa, la Tabla 8 relaciona los tipos de enzimas requeridos con los grupos de glicosil hidrolasas en los cuales se agrupan y el tipo de hemicelulosa que puede hidrolizar.

La hemicelulosa de tipo xilano requiere también de acetilxilano esterasa y 4-O-metil-glucuronol metilesterasa, que se clasifican dentro de las esterasas de carbohidratos, pero este trabajo se enfocó en las glicosil hidrolasas que son las más representativas.

Tabla 8. Enzimas requeridas para hidrólisis de hemicelulosa.

Función	Grupos CAZy	Xiloglucano	Xilano	Galactomanano
Acetilxilano/feruloil esterasa	CE1, CE2, CE3, CE5		x	
4-O-metil-glucuronol metilesterasa	CE15		x	
Xiloglucano β-1,4-endoglucanasa (EC 3.2.1.151)	GH12, GH74	x		
α-1,4-galactosidasa (EC 3.2.1.22)	GH27,GH36	x	x	x
α-arabinofuranosidasa (EC 3.2.1.55)	GH51,GH54	x	x	
α-glucuronidasa (EC 3.2.1.131)	GH67,GH115		x	
α-xilosidasa (EC 3.2.1.177)	GH31	x		
β-1,4-endomananasa (EC 3.2.1.78)	GH5,GH26			x
β-1,4-endoxilanasa (EC 3.2.1.8)	GH10,GH11		x	
β-1,4-galactosidasa (EC 3.2.1.23)	GH2,GH35	x	x	x
β-1,4-manosidasa (EC 3.2.1.25)	GH2			x
β-1,4-xilosidasa (EC 3.2.1.37)	GH3,GH43		x	

Fuente: Kubicek et al. (2012)

2.3.4. Degradación de hemicelulosa tipo xilano

La hidrólisis completa de xilano (Figura 8) involucra la acción colectiva de α-glucuronidasa (AGU), la β-1,4-endoxilanasa (XLN), Arabinoxilano α-arabinofuranohidrolasa (AXH), Ferulol esterasa (FAE), Acetil (xilano) esterasa (AXE) y β-1,4-xilosidasa (BXL). La β-1,4-endoxilanasa (XLN) actúa sobre la cadena principal de xilano liberando xilo-oligosacaridos de tamaños más pequeños. Para las ramificaciones laterales actúan la α-glucuronidasa (AGU) removiendo el ácido glucurónico del xilano, la arabinoxilano α-arabinofuranohidrolasa (AXH) liberando L-arabinosa de la cadena principal, ferulol esterasa

(FAE) y acetil (xilano) esterasa (AXE) remueven ácido ferúlico de los residuos de arabinosa y acetil de la cadena de xilano respectivamente, y, por último, la β-1,4-xilosidasa (BXL) actúa sobre los xilo-oligosacaridos libres obteniendo monosacáridos tipo xilosa.

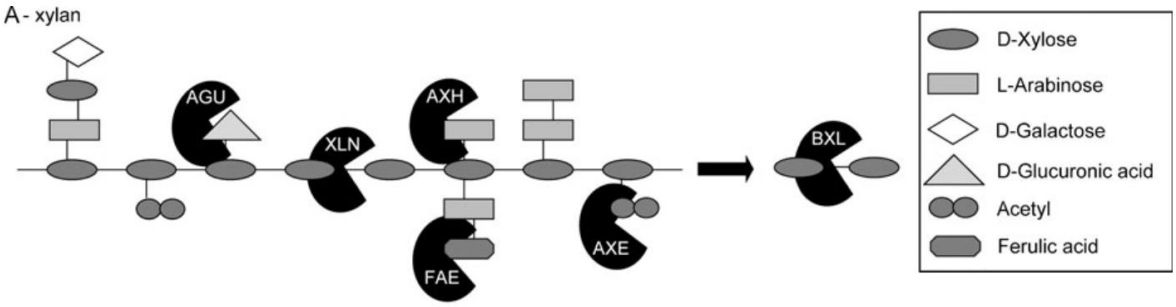


Figura 8. Esquema de degradación de hemicelulosa tipo xilano.
Fuente: Modificado de Brink & Vries (2011).

2.3.4.1. *Endo-β-1,4-xilanasa (EC 3.2.1.8)*

La endo-1,4-β-xilanasa rompe los enlaces glicosídicos de la cadena principal de xilano y por eso reduce el grado de polimerización en el sustrato, resultando en cadenas cortas de xilo-oligosacaridos de tamaño variable (Geetha et al., 2017). La mayoría de xilanasas de tipo fúngica han sido identificadas en las familias GH10 y GH11 (Alvarez et al., 2013). Los grupos GH5, GH7 y GH8 contienen dominios catalíticos para endo-1,4-β-xilanasa.

La primera xilanasa GH10 fue caracterizada para *Trichoderma reesei*. Su estructura muestra un barril TIM α/β de triosefosfato isomerasa. La estructura cristalina de GH10 de *P. simplicissimum* revela que pequeñas cadenas de xilo-oligómeros como xilobiosa y xilotriosa se unen a la hendidura del sitio activo con un extremo reductor de hidrógeno unido a la región catalítica (Brink et al., 2011). Son capaces de atacar versiones mínimamente cambiadas de xilano, pero a diferencia de GH11, solo degradan unidades cortas de β-1,4-xilosa y muestran gran afinidad por β-1,4-xilooligosacaridos pequeños (Kubicek et al., 2012).

Las endoxilanasas GH11 son las xilanasas más caracterizadas para hongos. Su estructura 3D consiste en un rollo donde la cadena principal se pliega para formar dos hojas beta antiparalelas, estas se pliegan para formar una gran hendidura que tiene el sitio activo. Son enzimas retenedoras.

2.3.4.2. β-xilosidasas (EC 3.2.1.37)

Las xilosidasas hidrolizan pequeños restos de xilooligosacaridos y el disacárido xilobiosa liberando residuos de xilosa por el terminal no reductor (Gírio et al., 2010).

La mayoría de hongos no contienen GH30 exo-β-1-4-xilanasa y por eso requiere de otra enzima exo para la eficiente hidrólisis de xilano, por lo que β-xilosidasas han sido identificadas en numerosos hongos y están distribuidas en las familias GH3, GH43 y GH54. Las β-xilosidasas GH3 muestran una especificidad del sustrato para residuos de monosacáridos, posición de enlace, y longitud del sustrato. El mecanismo de acción ocurre al remover residuos glicosídicos de los terminales no reductores de su sustrato.

Las β-xilosidasas GH43 existe en hongos y muestra actividad α-L-arabinofuranosidasa, endo- α-L-arabinanasa y β-D-xilosidasa (Bastien et al., 2013).

2.3.4.3. α-galactosidasa (EC 3.2.1.22)

La α-galactosidasa están distribuidas en las familias GH27 y GH36. La familia GH27 hidrolizan enlaces con α-galactosa por un mecanismo de retención (Brink & Vries, 2011). La estructura 3D en *Trichoderma reseei* muestra una configuración de barril (α/β). En hongos se pueden encontrar en las especies *A. niger, P. purpurogenum, P. simplicissimum, T. reesei, A. nidulans* (Kubicek et al., 2012).

Son parecidas a la GH27 en cuanto a estructura y mecanismo de acción, con la diferencia de que tienen un dominio β-*sandwich* que sirve para el anclaje del sustrato con terminal de aminoácido al sitio activo. En hongos se ha encontrado en *Mycocladus corymbifera, Penicillium, A. niger, A. nidulans, Gibberella sp., Rhizopus sp. F78, T. reesei* (Kubicek et al., 2012).

2.3.4.4. β-galactosidasa (EC 3.2.1.23)

Las β-galactosidasa de tipo fúngica se encuentran en las familias GH2 y GH35, donde la primera es intracelular y la segunda es secretada.

La GH35 cataliza el extremo no reductor de β-D-galactosa. Son enzimas retenedoras. Estructuralmente tienen una forma 3D de barril (α/β).

2.3.4.5. α-glucuronidasa (EC 3.2.1.131)

Remueven el ácido glucurónico de los residuos laterales de la cadena de xilano, se encuentran en las familias GH67 y GH115 (Yan et al., 2017).

Las que son de tipo GH67 remueven el ácido glucurónico que está ligado al carbono 2 de la xilosa y actúan por el mecanismo de inversión. Están presentes en todos los hongos. Estructuralmente tienen una configuración de barril (α/β) (Shallom et al., 2003). Las GH115 también cumplen la misma función, pero a diferencia de la GH67 que solo rompe enlaces glucuronicos de xilooligosacaridos laterales, también rompen enlaces internos en xilooligosacaridos y xilanos (Yan et al., 2017).

2.3.5. Degradación de hemicelulosa tipo xiloglucano

La degradación de hemicelulosa tipo xiloglucano (Figura 9) implica que la Xiloglucano β-1,4-endoglucanase (XEG) degrade los enlaces β-1,4 de glucosa, α-fucosidasa (AFC) remueva fucosa de las cadenas laterales, α-xilosidasa (AXL) remueva los residuos de xilosa de la cadena principal de glucano, α-arabinofuranosidasa (ABF) elimine L-arabinosa de los residuos laterales de xilosa y finalmente β-1,4-galactosidasa (LAC) hidrolice el enlace de galactosa en el residuo lateral.

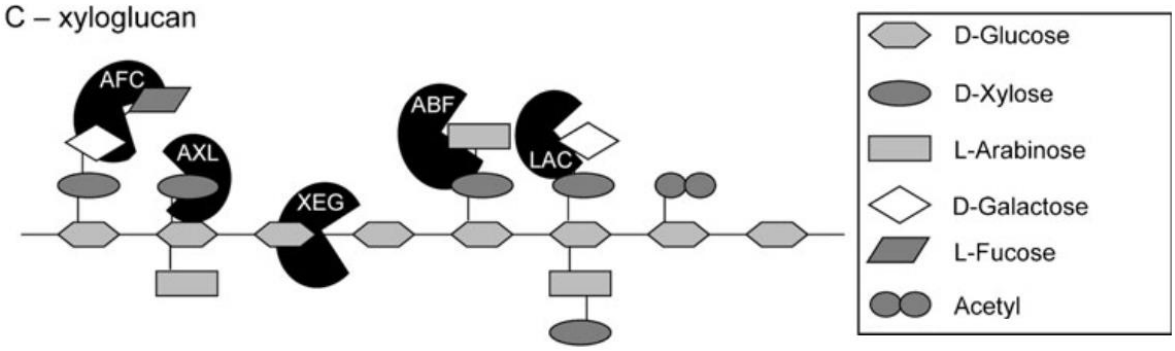


Figura 9. Esquema de degradación de hemicelulosa tipo xiloglucano.
Fuente: Modificado de Brink & Vries (2011).

2.3.5.1. *Glucuronoxilano xilanohidrolasas*

Escasas en hongos. Las enzimas del grupo GH30 tipicamente contiene enzimas con actividades β-glucosilceramidasa, β-1-6-glucanasa y β-xilosidasa, que son encontradas en procariotas y eucariotas. Recientemente también se han descubierto xilanasas tipo glucuronoxilano xilanohidrolasas que eran anteriormente del grupo GH5. Las enzimas GH30 son retenedoras. Posee la típica estructura de barril TIM α/β.

2.3.5.2. α-arabinofuranosidasa (EC 3.2.1.55)

Hidrolizan hemicelulosas que contienen arabinofuranosa y se pueden encontrar en las familias GH3, GH43, GH51, GH54 y GH62 (Murphy et al., 2011). Muestran una amplia especificidad del sustrato actuando sobre enlaces O5, O2 y O3 de la arabinofuranosidasa. En los organismos *Meripilus giganteus* y *H. insolens,* catalizan la degradación de residuos de α-1,2 y α-1,3- arabinofuranosa de los residuos de xilosa del arabinoxilano (Kubicek et al., 2012).

La primera estructura 3D para GH43 fue determinada para *C. japonicus.* Tiene una estructura de pliegue de hélices β (Shallom et al., 2003). La GH51 se encuentran abundantemente en *Aspergillus* y *Penicillium spp* (Kubicek et al., 2012).

2.3.6. Degradación de hemicelulosa tipo manano

Para la completa degradación de la cadena principal requiere la acción simultanea de endo-β-1,4-mananasas (MAN) y β-manosidasas (MND), donde las primeras hidrolizan el manano hasta llevarlo a oligosacáridos más pequeños, y los segundos degradan estos hasta volverlos monosacáridos tipo manosa. Adicionalmente, para remover las cadenas laterales que se adhieren a los mananos son requeridas las enzimas β-glucosidasas (EC 3.2.1.21), α-galactosidasas (AGL) y acetil manano esterasas.

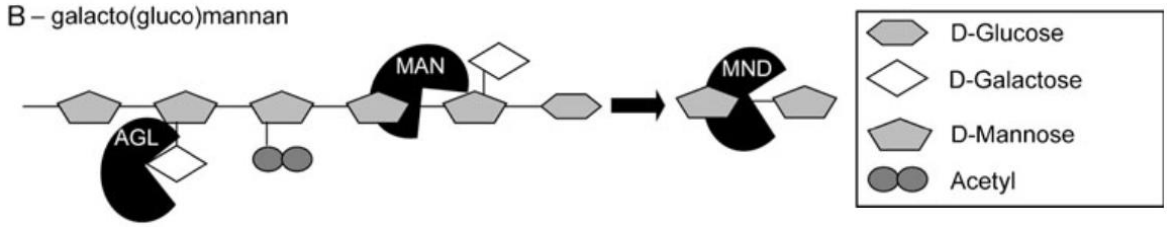


Figura 10. Esquema de degradación de hemicelulosa tipo galactoglucomanano.
Fuente: Modificado de Brink & Vries (2011).

2.3.6.1. β-1,4-endomananasa (EC 3.2.1.78)

Hidrolizan la cadena de manano de las hemicelulosas y liberan β-1,4-oligomananos. Es responsable de la degradación de la cadena principal del manano. Se encuentran presentes en las familias GH5 y GH26. La primera estructura 3D para GH5 fue determinada para el organismo *Cellvibrio japonicus* tiene una estructura de barril (α/β) (Shallom et al., 2003). Se han encontrado en eucariotas como *B. circulans, Clostridium cellulolytic, T. reesei, A.*

aculeatus, y Agaricus bisporus. Para la familia GH26, las mananasas se han encontrado exclusivamente en hongos como *Aspergillus, Humicola, N. crassa, P. chrysogenum,* y *Orpinomyces sp* (Kubicek et al., 2012).

Las mananasas están propagadas en múltiples tipos de bacterias, pero comercialmente solo se consiguen en un corto rango, entre ellos los más importantes son: *Bacillus sp., Streptomyces sp., Caldibacillus cellulovorans, Caldicellulosiruptor Rt8B, Caldocellum saccharolyticum* (Gírio et al., 2010).

2.3.6.2. β-1,4-manosidasa (EC 3.2.1.25)

Hidroliza los enlaces β-1,4 de los oligomananos resultantes de la degradación efectuada por la endomananasa en la cadena de manano por su extremo no reductor. Está presente en las familias GH1, GH2 y GH5 (Shallom et al., 2003).

2.3.7. Bases de datos en Bioinformática

Una base de datos biológica es una colección de información sobre ciencias de la vida, recogida de experimentos científicos, literatura publicada, tecnología de experimentación de alto rendimiento, y análisis computacional. Las bases de datos más relevantes en biología incluyen datos de secuencias de nucleótidos, proteínas, estructura de proteínas, genomas, expresión genética, bibliografía, taxonomía, metabolismo, factores de transcripción, etc. (Baxevanis & Bateman, 2015).

Las principales bases de datos para la investigación en bioinformática son:

- *GenBank*, una base de datos pública de secuencias de ADN, es la más actualizada, junto *DNA Databank of Japan (DDBJ) y European Nucleotide Archive (ENA)*, hacen parte del grupo *National Institutes of Health (NIH)*, que concentra secuencias de más de 100.000 organismos, el acceso a datos se puede realizar vía *FTP* (*File Transfer Protocol*).
- ENA (Stoesser et al., 2002) (*European Nucleotide Archive*) Es la base de datos de secuencias de nucleótidos que pertenece al *EMBL* (*The European Molecular Biology Laboratory*), financiada por la unión europea, contiene datos crudos y anotación funcional.

- *Uniprot* (Apweiler et al., 2004), (*Universal Protein*), es una base de datos que concentra información de *Swiss-Prot, TrEMBL y PIRt*, la que la convierta en el mayor contenedor de secuencias de proteínas.
- PDB (Berman et al., 2000), (*Protein DataBank*), es la base de datos hecha específicamente para proteínas, es la que tiene más curación de todas.

Existen grupos de investigación dedicados a realizar grupos de datos para ramas más específicas de la investigación como lo son las bases de datos en enzimas.

2.3.8. CAZy

Existen bases de datos más específicas que abarcan los subespacios de los problemas biológicos, tal es el caso de la base de datos CAZy (*Carbohydrate Active enZymes*) (Cantarel et al., 2009). CAZy, funciona desde 1998 de manera *online*, es una base de datos especializada dedicada a la visualización y al análisis de la información genómica, estructural y bioquímica de las enzimas activas en carbohidratos (CAZymes).

La información funcional y la estructura en 3D es agregada y curada regularmente, basada en la literatura publicada. Nuevos genomas se añaden regularmente, poco después de que aparecen en las versiones diarias del GenBank. Las nuevas familias se crean sobre la base de la evidencia publicada para la actividad de al menos un miembro de la familia y todas las familias se actualizan regularmente, tanto en contenido como en la descripción. En el servicio en línea de CAZy se describen cinco clases de enzimas de carbohidratos (Tabla 9). CAZy presenta la clasificación de glicosil hidrolasas que contiene entre otras las enzimas que degradan celulosa y hemicelulosa (a excepción de las enzimas de tipo esterasa) que se utilizan en esta investigación.

Tabla 9. Clasificación de enzimas presentes en CAZy.

Clasificación	Descripción
Glicosil hidrolasas (GH)	Hidrólisis o reordenamiento de enlaces glicosídicos
Glicosil transferasas (GT)	Formación de enlaces glicosídicos
Liasas de polisacáridos (PL)	División no hidrolítica de enlaces glicosídicos
Esterasas de Carbohidratos (CE)	Hidrólisis de esteres de carbohidratos

Actividades Auxiliares (AA)	Enzimas de óxido-reducción que actúan en conjunto con enzimas de carbohidratos.

Fuente: cazy.org

En el sitio *web* de *CAZy* (Figura 11) la definición de cada familia está compuesta por un encabezado donde están descritas las actividades, mecanismo de acción, estructura 3D, entre otros datos, y un cuerpo que contiene todos los identificadores de enzimas asociadas a la clasificación de la familia. Estos detalles se componen de ocho pestañas: 'All', que contiene todas los identificadores de secuencias, 'Archaea' que contiene solo las enzimas que pertenecen a la taxonomía arquea, 'Bacteria' que contiene todos los de bacterias, 'Eukaryota' que contiene todos los identificadores de eucariota, 'Viruses' que contiene todos los identificadores para virus, 'Unclassified' que contiene secuencias que han sido publicadas pero que no se conoce su clasificación tales como las secuencias de patentes, 'structure' que son las secuencias cuya estructura 3D ya ha sido determinada y por último las 'characterized' que ya han sido totalmente caracterizadas.

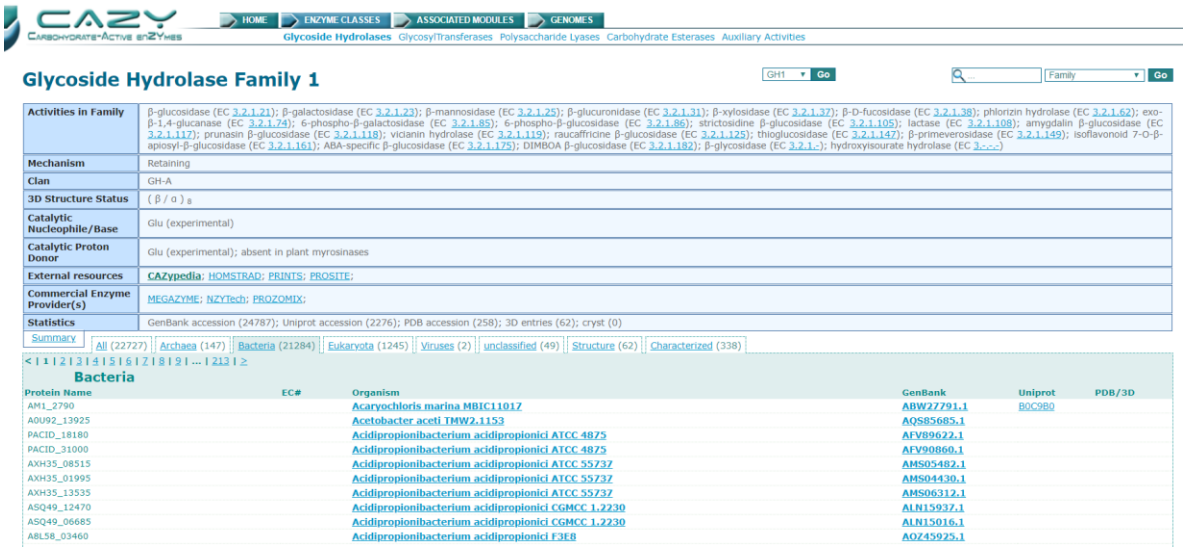


Figura 11. Vista de la parte de la familia GH1 publicada en CAZy
Fuente: cazy.org/GH1.html

Para el desarrollo de este proyecto se contó con datos de identificadores extraídos manualmente en archivos separados de las pestañas 'Archaea', 'Bacteria', 'Eukaryota', 'Viruses' y 'Unclassified' de cada una de las familias glicosil hidrolasas separadas por familia y taxonomía así, GH1_Archaea, GH1_Bacteria, etc. Este proceso fue realizado por

estudiantes del pregrado en Bacteriología de la Universidad Católica de Manizales que formaron parte del macroproyecto al cual está adscrito este proyecto.

2.3.9. Homología de secuencias y OrthoMCL

El rápido crecimiento de los datos genómicos, ha generado la necesidad de hacer su anotación funcional a partir de una gama cada vez mayor de especies. Los métodos comparativos basados en la identificación automatizada de secuencias ortólogas parece ser la solución, dado a que facilita la anotación funcional para realizar estudios de genómica comparativa y evolutiva, para lo cual se han desarrollado técnicas basadas en homología de secuencias.

Un gen es homólogo a otro, si ambos descienden de un mismo gen ancestro común, pero que por condiciones evolutivas se han desarrollado con diferentes mutaciones. A nivel de datos el hecho de que dos proteínas sean homólogas, significa que existe una correspondencia de aminoácidos entre sus secuencias. Las proteínas homólogas pueden ser de dos tipos: ortólogas (formados por especiación) o parálogas (surgen por la duplicación de genes).

Los genes ortólogos normalmente conservan la estructura de dominios similares y tienen la misma funcionalidad tras la especiación, mientras que los parálogos son propensos a divergir con nuevas funciones a través de mutaciones puntuales y recombinaciones de dominio. La Figura 12, muestra los eventos de duplicación y especiación correspondientes a la paralogía y ortología, respectivamente.

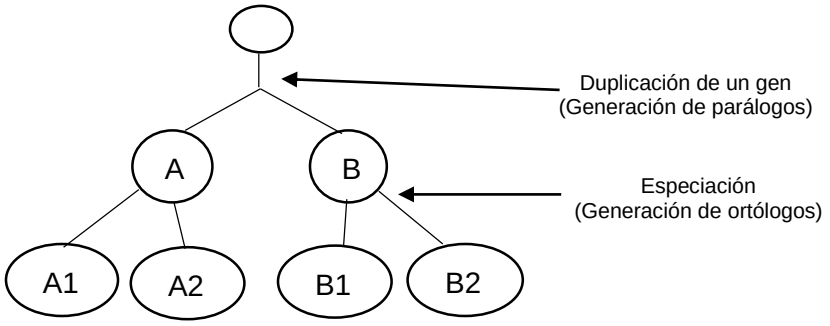


Figura 12. Homología de secuencias, ortólogos y parálogos.

Según Coutinho et al., (2015), algunos beneficios del uso de métodos de comparación basados en homología de secuencias son:

- Anotación de nuevas secuencias basadas en inferencia de ortologías.
- Estimación de los radios de evolución de familias de proteínas.
- Búsqueda de patrones genómicos comunes
- Análisis de genomas a partir de mutaciones como deleciones o SNPs.
- Búsqueda secciones genómicas persistentes durante la evolución del organismo.

Las ortólogas son el tipo de homología requerida para el presente proyecto por su condición de tener identidad funcional y por ayudar a mantener la homología por especiación de los diferentes grupos taxonómicos. Para distinguir la ortología de secuencias se han empleado varias estrategias, dada una secuencia de entrada (Tabla 10).

Tabla 10. Métodos para determinación de secuencias ortólogas.

Por filogenia	**Por distancia**	**Usan BlastP**
RIO (*Resampled Inference of Orthology*)	RSD (*Reciprocal Smallest Distance*)	*Reciprocal Best Hit* (RBH)
Orthostrapper/HOPS (*Hierarchical grouping of Orthologous and Paralogous Sequences*)		COG (*Cluster of Orthologous Groups*)
		KOG (*euKaryotic Orthologous Groups*)
		Inparanoid
		OrthoMCL (Chen, Mackey, Vermunt, & Roos, 2007)

Fuente: Chen et al., (2007).

OrthoMCL provee un método para construir grupos ortólogos usando clusterizado de Markov para agrupar ortólogos y parálogos. Puede ser usado para agrupar ortólogos de múltiples especies. El algoritmo de clusterizado de *Markov* (MCL) incluye un parámetro importante el cual es el valor de inflación que sirve para regular que tan relajado o estrecho va a quedar cada clúster, al aumentar el valor de inflación, aumenta la estrechez de los clústeres (L. Li, et al., 2003).

2.3.10. Métodos de comparación entre secuencias de proteínas

La comparación entre proteínas permite establecer las relaciones funcionales dada la identidad entre secuencias, entre más parecidas sean un par de secuencias en cuanto a la distribución de sus aminoácidos, mayor es su parecido funcional, pudiendo así describir una secuencia objetivo o *subject* en términos de la secuencia de búsqueda o *query*. La existencia de esta relación estructura-función permiten incluso definir los llamados perfiles proteicos para el caso de las proteínas. Un perfil proteico es de una manera más clara, una caracterización de un conjunto de proteínas con unas cualidades funcionales y estructurales específicas que determina la forma en que estas proteínas se configuran en cuanto a su secuencia de aminoácidos. Esta caracterización puede describir dominios específicos para este conjunto de proteínas, lo cual permite establecer sus funcionalidades dado las estructuras genómicas que contiene.

Para este trabajo se utilizan para la búsqueda y anotación de secuencias de proteínas los métodos de alineamiento pareado con el software *Blast* y modelos ocultos de Markov o *HMM* (*Hidden Markov Models*) con el software *HMMer*. El primero es una forma rápida de comparar dos secuencias, en este caso de proteínas, que resalta los aminoácidos en común cuando se alinean ambas frente a frente mediante un algoritmo de alineamiento local, cuando las zonas de coincidencia son muy extensas se puede establecer una relación funcional y estructural entre ambas secuencias. El segundo sirve para identificar patrones estructurales de una familia de secuencias (perfil protéico). Mientras que el primero compara secuencias una a una, el segundo compara en base a la definición de una familia completa, lo que permite caracterizar una secuencia objetivo dentro del perfil de una familia. Así, la intención de involucrar ambas metodologías de búsqueda consiste en tener la mayor cantidad de candidatos que coincidan con las secuencias y familias que se tienen para este proyecto que son familias de enzimas relacionadas con la degradación de celulosa y hemicelulosa, donde con la metodología de alineamiento se encuentran unos candidatos y con la de *HMM* se hallan otros porque usa un modelo estocástico para hallar nuevos resultados que pueden pertenecer a los grupos de enzimas investigados al tener en cuenta otras combinaciones que se pueden dar por probabilidad. En ambos algoritmos se calcula automáticamente un valor de Score que entre más alto sea, mayor es la coincidencia entre las secuencias o perfiles proteicos, este es el criterio de aceptación de las secuencias como parte de una familia que se utiliza en esta investigación.

2.3.11. Modelos Ocultos de Markov (HMM)

Un modelo oculto de Markov es un modelo probabilístico que se usa para describir la evolución de eventos observables que dependen de eventos desconocidos. En un HMM los eventos observables se llaman símbolos y los no conocidos se llaman estados. Tiene dos procesos estocásticos, un proceso no visible de estados ocultos y otro visible de símbolos observables. Los estados ocultos forman una cadena de Markov y la distribución de probabilidad del símbolo observado depende del estado subyacente (S. Eddy, 1998).

Este enfoque de evaluar procesos donde existen eventos conocidos y desconocidos, es útil en el modelado de secuencias proteicas. Una proteína consiste en una cantidad de subestructuras o dominios funcionales diferentes que muestra comportamientos estadísticamente distintos. Entonces dada una nueva proteína es interesante predecir los dominios que la constituyen, que en un HMM correspondería a uno o más estados, y sus posiciones en la secuencia de aminoácidos (eventos observables en el HMM). Además, también se puede encontrar la familia a la que pertenece la proteína. Los HMM han demostrado ser muy efectivos al momento de representar secuencias y han mostrado buen desempeño en la inferencia de homología (Rossi et al., 2017).

En un Modelo Oculto de Markov que trata secuencias biológicas, existe una entrada controlable (un alineamiento múltiple) y una salida no controlable la cual debe ser validada manualmente.

2.3.11.1. *Construccion de un perfil HMM.*

Los perfiles HMM se contruyen a partir de un alineamiento múltiple de secuencias. En el ejemplo descrito en Yoon (2009), hay un alineamiento múltiple de secuencias de 5 aminoácidos como se muestra en la parte A de la Figura 13. Como se observa, las frecuencias de aminoacidos son diferentes en cada una de las columnas.

El k-ésimo estado M_k en el perfil HMM se denomina estado ‘match’ (ver parte B de la Figura 13), ya que se usa para representar el caso cuando un símbolo en una nueva secuencia de observación coincide con el k-ésimo símbolo en la secuencia de consenso de la alineación original. Como resultado, el número de estados de coincidencia en el perfil HMM resultante es idéntico a la longitud de la secuencia de consenso. Al interconectar los estados de coincidencia $M_1, M_2, \ldots, M_5$, obtenemos un HMM sin *gaps* (un *gap* es una mutación por inserción o borrado) como se muestra en la parte B de la Figura 13. Este HMM

sin *gaps* puede representar secuencias que coinciden con la secuencia consenso del alineación sin ningún *gap*, y funciona como la linea principal del perfil-HMM que está por construirse.

Tras la construcción del HMM sin *gaps*, se agregan los estados de inserción I_k y los estados de deleción D_k al modelo para establecer las inserciones y deleciones en las nuevas secuencias de observación. Para el caso en que la secuencia observada es más larga que la secuencia de consenso de la alineación original, si alineamos estas secuencias, habrá una o más letras en la secuencia observada que no están presentes en la secuencia consenso, estos símbolos adicionales están modelados por los estados de inserción. El estado de inserción I_k se usa para manejar los símbolos que se insertan entre las posiciones k-ésima y la (k+1)-ésima en la secuencia de consenso.

Ahora, se considerará el caso cuando la nueva secuencia observada es más corta que la secuencia de consenso. En este caso, habrá una o más bases en la secuencia consenso que no están presentes en la secuencia observada. El k-ésimo estado de deleción D_k se usa para manejar la eliminación del k-ésimo símbolo en la secuencia original de consenso. Después de agregar los estados de inserción y los estados de deleción al HMM sin *gaps*, obtenemos el perfil final-HMM que se muestra en la parte C de la Figura 13.

Para estimar los parámetros de un perfil HMM basado en un alineamiento múltiple de secuencias primero se tiene que decidir qué columnas deben representarse por estados de coincidencia (*Match*) y qué columnas deben modelarse mediante estados de inserción. Cuando una columna contiene uno o más *gaps*, se debe comparar el número de símbolos y el número de *gaps*. Si la columna tiene más símbolos que *gaps*, tratamos los espacios como eliminaciones de símbolos. Por lo tanto, se modela la columna usando un estado de coincidencia M_k (para los símbolos en la columna dada) y un estado de eliminación D_k (para los *gaps* en la misma columna), y para el caso en que haya más *gaps* que símbolos, los símbolos se convierten inserciones, por lo tanto, se usa un estado de inserción I_k para representar la columna.

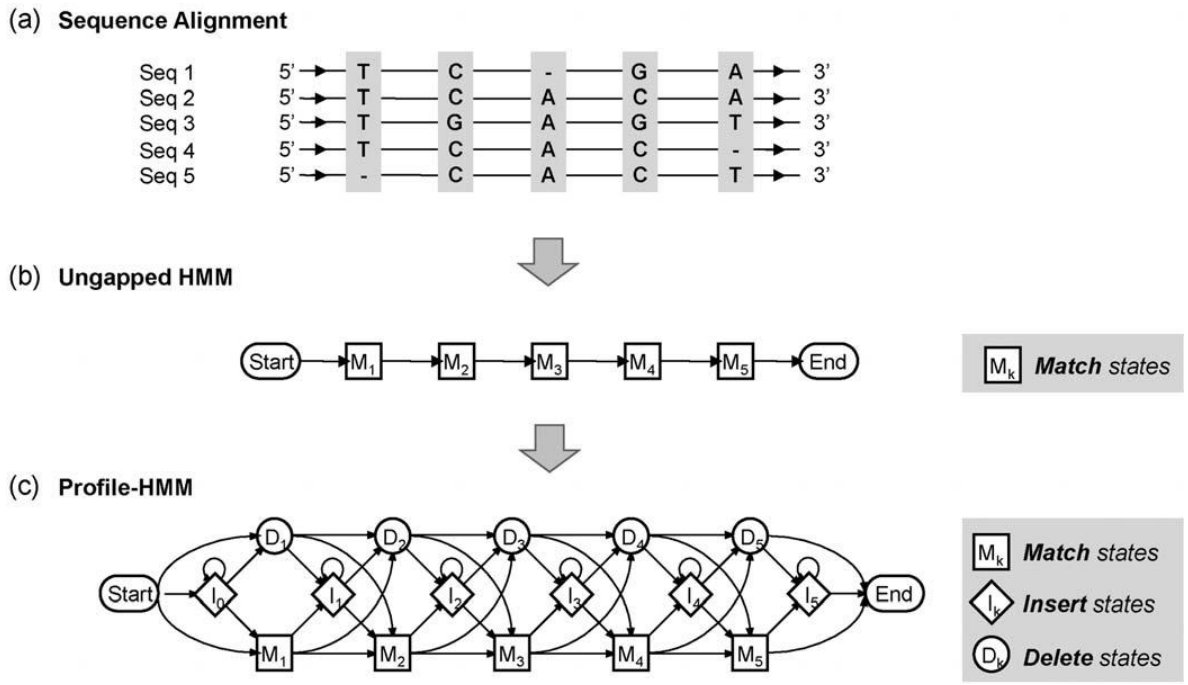


Figura 13. Ejemplo de funcionamiento de un *Hidden Markov Model*.
Fuente: Yoon (2009).

Una vez se ha decidido qué columnas deben representarse mediante estados de coincidencia y cuáles deben representarse mediante estados de inserción, se conoce la secuencia de estado subyacente para cada secuencia de símbolo en la alineación. Por lo tanto, se pueden estimar las probabilidades de transición y las probabilidades de emisión del perfil HMM al contar el número de cada transición de estado o emisión de símbolo y calcular sus frecuencias relativas.

Actualmente paquetes de *software* como HMMer, que pueden ser usados para construir perfiles HMM y que además tienen herramientas para hacer análisis de secuencias. Dado a su capacidad para describir familias, ya existen bases de datos como *PFAM (Protein FAMilies*) que ya tienen perfiles HMM para familias conocidas de secuencias. Finalmente, si existe un perfil HMM que representa una familias de secuencias biológicas, se pueden usar para encontrar homólogos adicionales que pertenecen a la misma familia (Yoon, 2009), lo que permite la clasificación y anotación de la secuencia dada.

3. Materiales y métodos

El proceso de construcción de un flujo de trabajo para la anotación de enzimas degradadoras de celulosa y hemicelulosa implicó la realización de tres etapas principales: (i) La identificación de nuevos grupos de secuencias de enzimas glicosil hidrolasas correspondientes a familias y subfamilias de acuerdo con su similitud, a partir de la base de datos pública CAZy, donde se utilizaron alineamientos pareados, y filtros para determinar la similitud de las secuencias y ortología de secuencias. (ii) La validación *in silico* de los nuevos grupos de familias y subfamilias obtenidos, a través de la generación de perfiles proteicos, para su uso posterior en la anotación de secuencias de enzimas celulolíticas y hemicelulolíticas. (iii) Finalmente, se desarrolló una herramienta para la anotación de secuencias de enzimas degradadoras de celulosa y hemicelulosa.

La Figura 14 muestra el esquema del flujo metodológico que se realizó durante el proyecto.

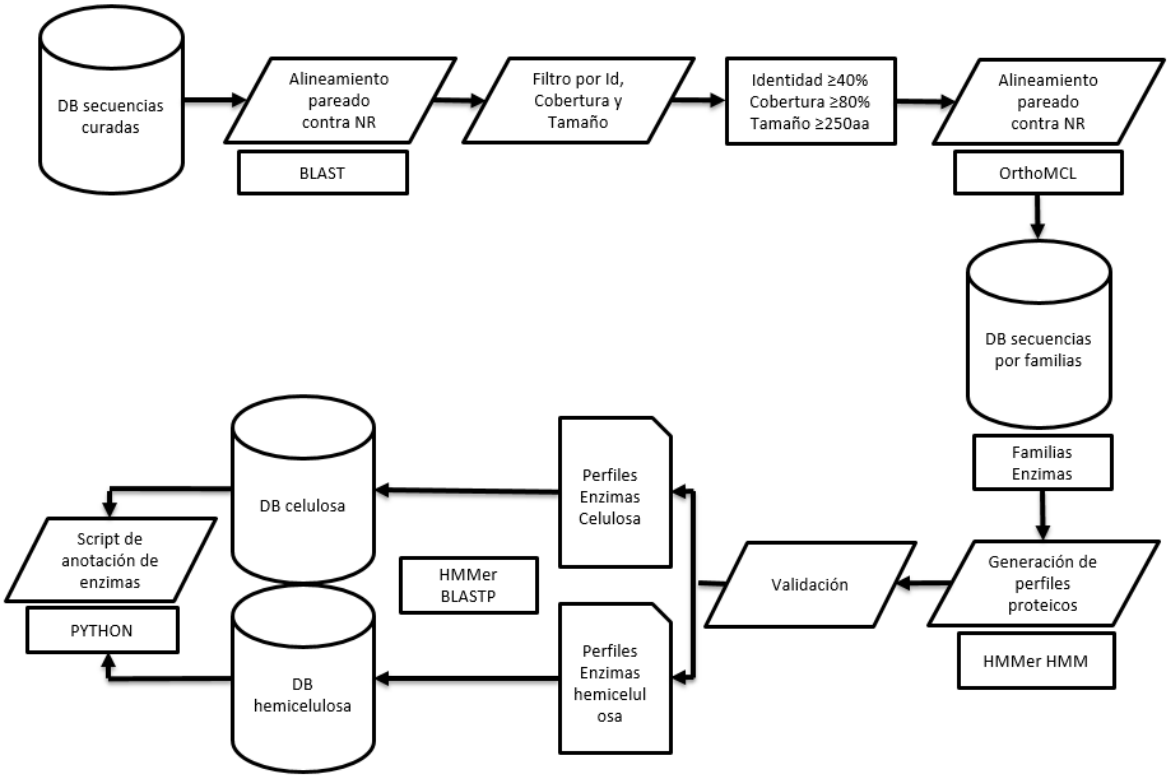


Figura 14. Flujo de trabajo general para la construcción de la base de datos de los nuevos grupos de enzimas degradadoras de celulosa y hemicelulosa.

3.1. Identificación de nuevos grupos de secuencias de enzimas glicosil hidrolasas correspondientes a familias y subfamilias de acuerdo con su similitud, a partir de la base de datos pública CAZy

3.1.1. Obtención de las secuencias de enzimas del CAZy

Los datos de entrada utilizados en el presente trabajo correspondieron a archivos de secuencias provenientes del CAZy (Anexo II), que fueron recolectados por los estudiantes del pregrado de Bacteriología de la Universidad Católica de Manizales durante el segundo semestre de 2015, en el marco del macroproyecto "Construcción de un *dataset* de familias de enzimas degradadoras de celulosa y hemicelulosa para su anotación y minería de datos de proyectos de las ómicas".

En total fueron suministradas las secuencias de 491 grupos de enzimas distribuidos de acuerdo al grupo taxonómico y las bases de datos donde se encontraban depositadas (Tabla 11).

Tabla 11. Distribución de las familias de las secuencias iniciales, por taxonomía y por base de datos.

Taxonomía	PDB	Uniprot	GenBank
Archaea	-	52	56
Bacteria	-	118	60
Eukaryota	50	27	11
Virus	10	-	28
Unclassified	3	-	76
Total	63	197	231

Posteriormente, se identificaron las características generales de los datos como: (i) las estadísticas de los rangos mínimos y máximos de la cantidad de secuencias existentes en las diferentes familias, por base de datos; (ii) la distribución de familias por taxonomía y por base de datos y (iii) las longitudes de las secuencias.

3.1.2. Alineamientos pareados con el software BLAST del NCBI

Inicialmente se hizo necesario aumentar la cantidad de secuencias de partida disponibles, con el fin de buscar secuencias funcionalmente similares para cada uno de los

grupos de enzimas disponibles en el CAZy. Lo anterior se realizó haciendo un alineamiento pareado con la herramienta BLAST v2.5.0 a partir de cada una de las familias de secuencias del CAZy contra la base de datos de secuencias no redundante del NCBI (NR), descargada el día 1 de enero de 2017.

La base de datos de referencia NR se formateó para proteínas utilizando el comando mkblastdb (1):

```
1 $ makeblastdb -in nr -dbtype prot -parse_seqids          (1)
```

Posteriormente, se ejecutó el blastp v2.5.0 para hacer alineamientos de cada uno de los archivos contenidos en las respectivas carpetas correspondientes a las tres bases de datos PDB, *Uniprot* y *GenBank*, usando como parámetro un e-*value* de 1e-5. Este es un valor de referencia que se usa para garantizar que el alineamiento se dé por evolución no un evento aleatorio. La salida de este alineamiento se obtuvo en formato tabular de 20 columnas:

```
1 $ for blastin in *_GenBank.fasta; \
2 $ do blastp -db /BIOS-Share/home/jhsuarezo/proyecto/nr_formatted_db/nr -
query $blastin -outfmt "6 qseqid qacc qlen sseqid sacc slen qstart qend
sstart send qseq sseq evalue length pident positive staxids sskingdoms
sstrand qcovs" -out $blastin.bp -evalue 1e-5 -num_threads 32; \
3 $ done                                                   (2)
```

Seguidamente, se filtraron los *hits* resultantes de ejecutar la línea de comandos 2. Los parámetros del filtro se definieron según lo referenciado por Pearson (2014) donde se establece que dos secuencias son homólogas si tienen un 30% de identidad sobre el total de sus longitudes, en este caso se eligió un valor de 10% más para un total de 40% de identidad, esto para que las secuencias filtradas fueran lo más idénticas posibles. Además, se estableció una cobertura del 80% de la secuencia *query* sobre la secuencia *subject* para darle un poco de holgura a la búsqueda (por si la secuencia objetivo era más pequeña por la evolución, o por si no se encontraba la secuencia completa), y no se excluyeron las que no fueron totalmente iguales en tamaño. El criterio para elegir la longitud de secuencia que se filtró se hizo con base en la mediana del histograma de frecuencia de longitudes de las secuencias de entrada que fue aproximadamente de 250 aminoácidos. El filtro final empleado incluyó, la cobertura del *query* (Columna 20, qcovs) mayor o igual a 80%, la

identidad (columna 15, pident) mayor o igual al 40%, y la longitud de la secuencia *query* (columna 3, qlen) mayor o igual a 250 aminoácidos (3):

```
1 $ for blastin in *_GenBank.fasta.bp; \
2 $ do python /BIOS-Share/home/jhsuarezo/proyecto/filter_tag_finder.py -f
$blastin -c 80 -i 40 -l 250 -o $blastin.filter -e _GenBank.fasta.bp -n 4 -
s ASC -u True -t False -p True \
3 $ done                                                          (3)
```

En este punto se recuperaron las secuencias asociadas a los identificadores filtrados, y se optó por aplicar el comando blastdbcmd (4):

```
1 $ for blastin in *.filter; \
2          $          do          blastdbcmd          -db          /BIOS-
Share/home/jhsuarezo/proyecto/nr_formatted_db/nr  -entry_batch  $blastin  -
out $blastin.fasta;\
3 $ done                                                          (4)
```

3.1.3. Agrupación de las subfamilias utilizando OrthoMCL

Con el fin de agrupar las secuencias filtradas y aumentadas por *clusters* funcionalmente similares y homólogos como describe Li et al. (2003), se ejecutaron los 13 pasos del OrthoMCL v2.2. Estos pasos se pueden encontrar en la guía de usuario disponible directamente en el sitio web del OrthoMCL (http://orthomcl.org/common/downloads/software/v2.0/UserGuide.txt).

Se inició con un alineamiento de todos contra todos con blastp de las secuencias obtenidas con un valor de umbral de e-*value* =1e-5. Se unificaron las secuencias en tres archivos, uno por cada base de datos (PDB, *Genbank*, *Uniprot*). Se eliminó la redundancia a cada archivo de secuencias utilizando el *software* CD-HIT v4.6.5 (W. Li & Godzik, 2006). El *script* completo de la ejecución se relaciona a continuación (5):

```
1 $ for index in *.fasta \
2 $ do cat $index >> mydb_combined.db \
3 $ done \
4 $ cd-hit -i mydb_combined.db -o mydb_unique.db -c 1.00 -n 5 \
5 $ makeblastdb -in mydb_unique.db -dbtype prot -parse_seqids    (5)
```

Posteriormente se ejecutó un segundo blastp con el archivo de secuencias no redundantes contra cada uno de los archivos de secuencias individuales no agrupadas del

paso previo, los parámetros utilizados fueron: e-*value* de 1e-5, corriendo en un *cluster* con 32 hilos (6):

```
1 $ for blastin in *.fasta; \
2 $ do blastp -db mydb_unique.db -query $blastin -out $blastin.bp2 -evalue
1e-5 -num_threads 32; \
3 $ done                                                    (6)
```

Se ajustaron los archivos fasta al formato requerido por OrthoMCL en el cual los identificadores deben tener la estructura >FAMILIA|IDENTIFICADOR, siendo FAMILIA el nombre de la familia a la cual pertenece la secuencia. Para este paso se creó un *script* personalizado en el que el código de familia consistió en una cadena de 4 caracteres donde la primera letra fue el grupo taxonómico (Ejemplo E para eucariota) y los otros tres caracteres correspondieron a las últimas 3 letras del nombre de la familia glicosil hidrolasa por tratar, es decir, para GH1 de taxonomía eucariota, el código de familia quedó EGH1.

Se formateó el resultado de la salida del OrthoMCL con el comando *orthomclBlastParser*, luego se crearon las tres bases de datos orthomcl en el motor de base de datos MySQL, correspondientes a PDB, *GenBank* y *Uniprot*. Antes de realizar la carga de los datos se agregó una llave única para todos los campos de la tabla *similarSequences* del orthomcl para evitar redundancias en los registros de la base de datos y en el agrupamiento del OrthoMCL (Anexo IV).

Con el fin de agrupar las secuencias se usaron dos valores de inflación (1 y 1.5) para armar clústeres con el *software* MCL (Enright et al., 2002), sin limitar el número de secuencias que pudieran ser incluidas en el grupo. Se pasó el formato MCL a *groups* de OrthoMCL (7):

```
1 $ orthomclMclToGroups PDB 0 < mclOutput_1.5 > groups_1.5.txt   (7)
```

3.1.4. Generación de subfamilias a partir de los resultados del OrthoMCL

Se generaron las nuevas subfamilias a partir de los resultados de los grupos obtenidos del archivo groups.txt resultante del OrthoMCL (Figura 15). Este archivo consiste en múltiples líneas donde cada una tiene un identificador de *cluster* (Ejemplo, PDB0) y los identificadores que se agruparon en cada uno.

Familia	Miembros de la familia
PDB0:	EH10|WP_052872423.1 EH10|WP_030198505.1 EH10|WP_051105978.1 EH10|WP_053788188.
PDB1:	EGH1|XP_018899386.1 EGH1|XP_018321393.1 EGH1|3AHZ_A EGH1|3AI0_A EGH1|3VIJ_A EG
PDB2:	VH34|ABB88351.1 VH34|ADD64125.1 VH34|AET74376.1 VH34|AGO87041.1 VH34|AHJ57677.
PDB3:	EGH1|3PTK_A EGH1|B8AVF0.1 EGH1|BAG13451.1 EGH1|BAJ93012.1 EGH1|BAK01899.1 EGH1
PDB4:	EH10|1B30_A EH10|1GOK_A EH10|1I1W_A EH10|1TUX_A EH10|2BNJ_A EH10|3NYD_A EH10|3
PDB5:	EH28|AAA85280.1 EH28|AAF03895.1 EH28|ACS44814.1 EH28|AFH77948.1 EH28|AGV40780.
PDB6:	EH55|3EQN_A EH55|ADX07322.1 EH55|ADX07323.1 EH55|AFS68742.1 EH55|BAE20245.1 EH
PDB7:	EH31|XP_008529406.1 EH31|XP_010358137.1 EH31|AAC39568.2 EH31|AAI20873.1 EH31|A

Figura 15. Salida del MCL para PDB

Cada uno de estos clústeres constituye una nueva subfamilia, para formarla se dividió cada línea del archivo en documentos separados donde el identificador del clúster es el nuevo nombre de la subfamilia, y para cada identificador se obtuvieron las secuencias asociadas dando por resultado nuevos archivos fasta de secuencias de familias.

Para asociar cada nueva subfamilia a las familias iniciales se generó una estadística de cada subfamilia teniendo en cuenta que en los identificadores para el procesamiento por el OrthoMCL tenía asociado el código de la familia, cada subfamilia se asoció a su familia padre por medio de esos identificadores en común determinados. Para el ejemplo de la Figura 16, la subfamilia PDB1 se asoció a la familia EGH1 (Eucariota GH1).

PDB1	EGH1:488
PDB2	VH34:477
PDB3	EGH1:475
PDB4	EH10:471

Figura 16. Ejemplo de estadística de subfamilias para PDB

3.2. Validación *in silico* de los nuevos grupos de familias y subfamilias obtenidos, a través de la generación de perfiles proteicos, para su uso posterior en la anotación de secuencias de enzimas celulolíticas y hemicelulolíticas.

3.2.1. Construcción de la Base de datos de enzimas celulolíticas y hemicelulolíticas

Para el filtro final se escogieron únicamente las familias de glicosil hidrolasas acordes a las funciones enzimáticas que corresponden a la degradación de celulosa y hemicelulosa.

En la Tabla 12 se relacionan las familias que se seleccionaron para realizar los filtros de enzimas degradadoras que se extrajeron tras la revisión bibliográfica.

Para cada función enzimática se escogieron las familias que son más representativas, por ejemplo, la función β-1,4-endoglucanasa (EC 3.2.1.4) según el CAZy está en presente en las familias glicosil hidrolasas con identificadores GH5, GH6, GH7, GH8, GH9, GH10, GH12, GH26, GH44, GH45, GH48, GH51, GH74, GH124 y GHNC, pero según la revisión bibliográfica para el interés de degradar celulosa y hemicelulosa del proyecto, estas enzimas están mayormente representadas en las GH5, GH7, GH12 y GH45, en las demás su contenido es mínimo por lo que su representación en las subfamilias no es significativa y los modelos *HMM* generados para estos grupos no son de ayuda para el objetivo de encontrar y anotar las enzimas para el proyecto.

Tabla 12. Familias filtradas para el proyecto.

Actividad enzimática	**Familia**
β-1,4-endoglucanasa (EC 3.2.1.4)	GH5, GH7, GH12, GH45
Celobiohidrolasa (EC 3.2.1.91)	GH6, GH7
β-1,4-glucosidasa (EC 3.2.1.21)	GH1, GH3
Xiloglucano β-1,4-endoglucanasa (Xiloglucanasa) (EC 3.2.1.151)	GH12, GH74
α-arabinofuranosidasa (EC 3.2.1.55)	GH51, GH54
α-xilosidasa (EC 3.2.1.177)	GH31
α-fucosidasa (EC 3.2.1.51)	GH29, GH95
α-1,4-galactosidasa (EC 3.2.1.22)	GH27, GH36
β-1,4-endoxilanasa (EC 3.2.1.8)	GH10, GH11
β-1,4-xilosidasa (EC 3.2.1.37)	GH3, GH43
Arabinoxilano arabinofuranohidrolasa (EC 3.2.1.55)	GH62
α-glucuronidasa (EC 3.2.1.131)	GH67, GH115
β-1,4-endomananasa (EC 3.2.1.78)	GH5, GH26
β-1,4-manosidasa (EC 3.2.1.25)	GH2
β-1,4-galactosidasa (EC 3.2.1.23)	GH2, GH35

3.2.2. Generación de perfiles proteicos utilizando modelos ocultos de Markov con HMMER3

Con las secuencias obtenidas de cada clúster se realizó un alineamiento múltiple con el *software clustal omega* (Sievers et al., 2011) en formato *stockholm* para posteriormente realizar los perfiles proteicos *HMM* con *HMMer3* (S. Eddy, 1998), utilizando los parámetros por defecto, para cada una de las subfamilias finales. Estos resultados de archivos *hmm* formaron parte de lo que son los perfiles de búsqueda para anotación de enzimas degradadoras de celulosa y hemicelulosa. Son parte del *dataset* para hacer búsquedas por medio del *script* desarrollado para el proyecto.

Estos archivos *hmm* en conjunto con los archivos de secuencias fasta de cada subgrupo, hicieron parte de la base de datos final para búsqueda de enzimas, los primeros para hacer búsquedas con el comando *hmmsearch* y los segundos para hacer comparaciones a modo de alineamientos con el *software* blastp.

3.2.3. Validación de los resultados de las nuevas familias

Para realizar el ejercicio de validación se utilizó la familia GH6, en ella se efectuaron dos tipos de validaciones, una manual para verificar que los resultados de la búsqueda hayan coincidido con la definición del tipo de familia y otra automática la cual pretendió demostrar la efectividad de un perfil creado para una de las subfamilias.

Para el primero se utilizó la base de datos de Refseq de secuencias no redundantes que cuenta con 50.737.205 secuencias. Tras el resultado de la búsqueda, se realizó el conteo visual de las secuencias que hicieron *match* a partir de la cual se realiza una matriz de conjunción donde se establecen los verdaderos positivos y los falsos positivos en base al conteo de endoglucanasas (EC 3.2.1.4) y celobiohidrolasas (EC 3.2.1.91) que están descritas para la familia GH6.

Para la validación de una de las subfamilias resultantes equivalente a la GH6 para eucariotas del PDB, se creó un modelo que se realizó con el 70% de las secuencias de la subfamilia, el 30% restante sirvió para evaluar el modelo planteado siendo estos los verdaderos positivos y para los verdaderos negativos se tomaron de a 3 secuencias de todas las familias diferentes a la familia objetivo que son secuencias aleatorias del resto de subfamilias. Para la realización de la prueba se utilizó el *script* createHMMFromDataSplit

que crea y genera un reporte automático de la ejecución (https://github.com/diriano/TCDB_HMM/blob/master/scripts/createHMMFromDataSplit.pl), el cual está disponible públicamente en internet en github.com. Del resultado gráfico de la prueba se pudo determinar un valor umbral para esta subfamilia.

3.3. Desarrollo de una herramienta para la anotación de secuencias de enzimas degradadoras de celulosa y hemicelulosa

3.3.1. Programación de un servicio para anotación de enzimas degradadoras de celulosa y hemicelulosa

Se desarrolló un *script* en el lenguaje de programación Python el cual usa como insumo los *datasets* de archivos fasta y perfiles *HMM* resultantes del proyecto investigativo para realizar búsqueda y anotación de enzimas degradadoras de celulosa y hemicelulosa. Este usa opcionalmente dos tipos de *software* que pueden ser HMMer3, Blast para anotación con *HMM* y búsqueda con alineamiento local respectivamente. Además, tiene por entrada un archivo de secuencias que se desea investigar.

Como valor agregado al proyecto se desarrolló una aplicación de consulta *web* que permite hacer búsquedas a los usuarios de las familias del proyecto para facilitar la investigación. Se utilizó el lenguaje de programación PHP v5.4, base de datos MySQL v15.1 y servidor de aplicaciones Apache. Se debe correr en un servidor Linux idealmente Biolinux v8.

4. Resultados y discusión

4.1. Identificación de los nuevos grupos de secuencias de enzimas glicosil hidrolasas correspondientes a familias y subfamilias de acuerdo con su similitud a partir de la base de datos pública CAZy

4.1.1. Obtención de las secuencias de enzimas del CAZy

Con el fin de entender la distribución de los datos de secuencia de entrada, se analizaron los archivos iniciales de CAZy tal como se obtuvieron del macroproyecto al cual pertenece este proyecto.

Tabla 13. Estadísticas generales de las secuencias de acuerdo con la base de datos.

Base de datos	Número de familias	Cantidad Secuencias (Rango mínimo)	Cantidad Secuencias (Rango máximo)
PDB	63	1	719
UNIPROT	197	1	5.198
GENBANK	231	1	4.666

La cantidad de secuencias por familia (Tabla 13) fue muy variable. Se identificaron familias que incluían desde una secuencia (Familia GH25 Archaea de Uniprot) hasta familias con 5.198 secuencias (Familia GH13 Bacteria de *Uniprot*). En la identificación de la proporción del número de familias en las tres bases de datos analizadas, se observó que PDB fue la que presentó el menor número de secuencias publicadas y con la que se obtuvo menos familias (63 para esta investigación), seguida por la base de datos *UNIPROT* con 197 familias y la base de datos del *GenBank* con 231 familias. Se encontraron secuencias que variaban según su taxonomía entre Archaea, Bacteria, Eukariota, Virus y *Unclassified* (clasificaciones correspondientes a las presentes en el sitio *web* de CAZy) (Lombard, et al., 2014).

Tabla 14. Distribución de familias por taxonomía y por base de datos según CAZy (Fecha de consulta: octubre 12 de 2017).

Taxonomía	PDB	Uniprot	GenBank
Archaea	9	51	61
Bacteria	105	122	132
Eukaryota	55	82	92
Virus	10	16	32
Unclassified	6	31	74

De acuerdo con los datos suministrados (Tabla 11), se observó que si bien la relación en la cantidad de familias se mantuvo según la base de datos (PDB, *GenBank* o *Uniprot*), esta proporción no persistió al momento de analizarlas por taxonomía. Por ejemplo, en la base de datos del *GenBank* para la clasificación 'Bacteria' solo se relacionaron 60 familias, mientras que para *Uniprot* se evidenciaron 118, cuando la proporción real de familias reportada directamente del CAZy mostró lo contrario (Tabla 14), debido a que en el *GenBank* (132) se tuvo la mayor cantidad de registros en comparación con *Uniprot* (122). Lo mismo ocurrió con la clasificación Eukaryota donde la proporción fue totalmente contraria a la realidad, siendo PDB la base de datos que tuvo la mayor cantidad y *GenBank* la que tuvo la menor. Lo anterior implica que un investigador al hacer uso de esta herramienta debe decidir qué base de datos utilizar dada la necesidad de su trabajo investigativo, es decir, si requiere hacer la búsqueda de enzimas en bacterias, probablemente sea más conveniente utilizar la base de datos *Uniprot* que la de *Genbank* dada la cantidad de familias utilizadas, y si desea buscar enzimas en virus, será mejor buscarlas en el *Genbank* que en las otras dos bases de datos. Esto se debió probablemente a un error en la adquisición de los datos iniciales del CAZy, que al ser capturados manualmente dada la no existencia de otro método más efectivo en el momento, no permitió que se obtuvieran correctamente. Adicionalmente, los datos fueron tomados el año 2014 lo que implica que los datos estaban desactualizados con respecto a la fecha de realización del proyecto.

Algunos representantes de las familias que se utilizaron como *input* del proyecto se pueden observar en la Tabla 15 provenientes de la base de datos PDB. Para efectos de ilustración solo se muestran las 10 primeras caracterizaciones de las familias; una vista más completa de todos los datos se puede obtener en el ***Anexo V***. Se puede apreciar que los rangos de los tamaños de las secuencias, variaron entre las mismas familias, buena parte de ellas con una desviación estándar considerable.

Tabla 15. Estadísticas iniciales base de datos PDB.

Especie	Familia CAZy	Base de datos	Cantidad de secuencias	Mínima Longitud	Máxima Longitud	Media	Desviación Estándar
Eukaryota	GH105	PDB3D	2	373	373	373	0
Eukaryota	GH10	PDB3D	36	274	335	309	11,89
Eukaryota	GH11	PDB3D	62	178	381	201	36,46
Eukaryota	GH12	PDB3D	36	218	413	233	44,7
Eukaryota	GH131	PDB3D	12	245	252	249	3,37
Eukaryota	GH13	PDB3D	158	32	905	492	174,59
Eukaryota	GH14	PDB3D	34	491	535	501	14,34
Eukaryota	GH15	PDB3D	12	470	599	500	47,69
Eukaryota	GH16	PDB3D	38	267	298	290	11,05
Eukaryota	GH17	PDB3D	20	306	323	315	6,5

Las longitudes de las secuencias también presentaron una variación diferente en cada base de datos (ver Tabla 16); éstas se encontraron entre 10 y 4.914 aminoácidos de longitud para el caso de la base de datos *GenBank*. La familia GH18 de Eucariotas de la base de datos PDB tuvo errores en algunas secuencias, con un total de 20 secuencias, cuyos identificadores no coincidieron correctamente con los datos disponibles en PDB, dichas secuencias tuvieron una variación de longitud entre 3 y 5 aminoácidos de longitud. El Anexo VI muestra una tabla en detalle de algunos casos no corrientes de secuencias cuya longitud fue menor o igual a 15 o mayor a 1.500 aminoácidos de longitud.

Tabla 16. Variación de las longitudes de las secuencias por base de datos.

Taxonomía	Longitud mínima	Longitud máxima	Promedio
PDB	3	1.045	346
Uniprot	10	3.081	548
GenBank	10	4.914	520

La mediana de las longitudes fue de aproximadamente 500 aminoácidos, la Figura 17 muestra el histograma de longitudes *versus* la cantidad de secuencias para esa longitud, en la base de datos *GenBank*. En el Anexo VII se presenta el documento completo de los datos en todas las bases de datos (PDB, *Uniprot*, *GenBank*).

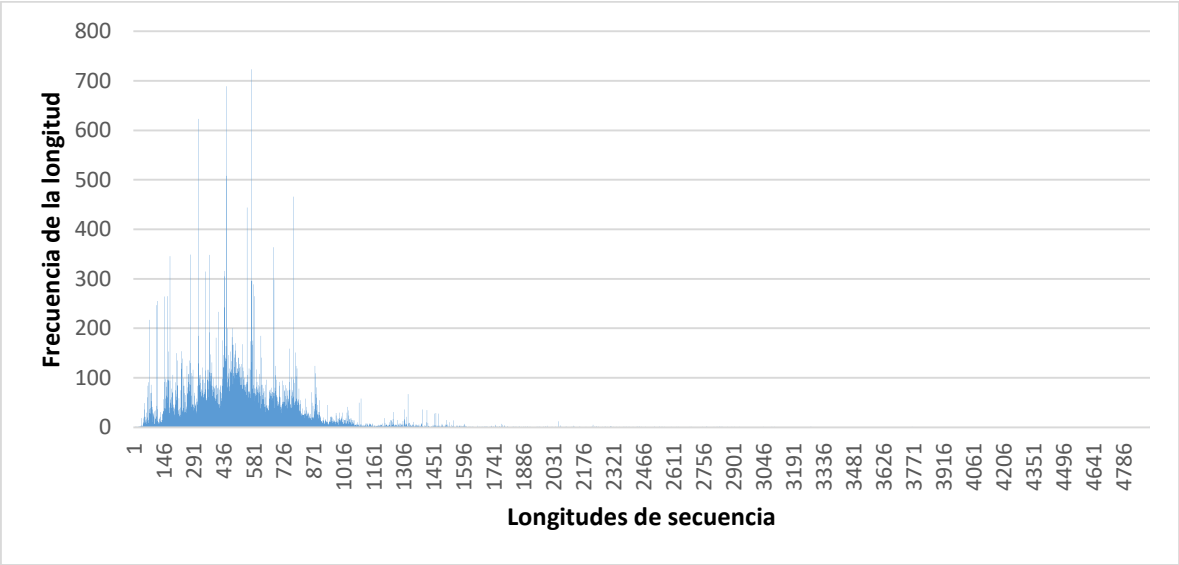


Figura 17. Histograma de distribución de longitudes de las secuencias para la base de datos *GenBank*.

4.1.2. Alineamientos pareados con el software BLAST del NCBI

En la ejecución del alineamiento pareado con blastp, los resultados de los tiempos de ejecución variaron en igual relación a la cantidad de familias que se procesaron por cada base de datos; en la Tabla 17 se muestra el detalle de los tiempos de ejecución.

Tabla 17. Tiempos de ejecución del alineamiento pareado con Blast.

Base de datos	Nodo (*Cores*)	Tiempo
PDB	1 nodo (32 *cores*)	0 días, 13 horas, 41 mins
Uniprot	1 nodo (32 *cores*)	23 días, 0 horas, 36 mins
GenBank	1 nodo (32 *cores*)	29 días, 17 horas, 56 mins

En este punto se efectuó el filtrado de los *hits* de los alineamientos provenientes del BLAST para lo que se creó el *script filter_tag_finder.py* en el lenguaje de programación Python v3.5.2 (Ver Anexo III), cabe anotar que la elección de tamaño de 250 como filtro para todos los alineamientos causó un sesgo para las secuencias pequeñas como lo son algunas correspondientes a virus y la categoría *unclassified*. El resultado de aplicar el filtro (cobertura >= 80, identidad >= 40 y longitud >= 250) es un listado de identificadores únicos de las secuencias en este caso el *sequence id* de la secuencia objetivo.

El resultado general de aplicar el filtro a las alineaciones *versus* los *hits* originales (3) se muestra en la Tabla 18.

Tabla 18. Resultado de la ejecución del filtro (3).

Base de datos	***Hits***	**Secuencias tras aplicar el filtro**	**Porcentaje de secuencias filtradas vs *Hits***
PDB	1.501.012	57.052	3,8%
Uniprot	22.194.362	951.368	4,28%
GenBank	31.980.872	555.150	1,73%

Los datos de las cantidades de secuencias tras aplicar el filtro no siguen una proporción con respecto al tamaño de los *hits* que se obtuvieron tras el alineamiento como inicialmente podría esperarse. Para la base de datos del *GenBank* que desde un inicio fue la que más secuencias tenía, terminó por tener menos que la base de datos de *Uniprot* que fue la que en tamaño le seguía. Como se observa en la Tabla 18 sólo el 1,73% (555.150) de los alineamientos fueron filtrados. Este comportamiento se justifica por la forma en que estaban distribuidas las familias para las tres bases de datos, en los archivos que inicialmente se tenían del CAZy, las familias correspondientes a bacterias del *GenBank* fueron superadas casi al 100 por ciento en cantidad de familias disponibles para *Uniprot* (118 para *Uniprot* y 60 para *GenBank*), siendo bacterias la categoría que más *hits* produjera; otra razón del comportamiento es que de las 231 divisiones iniciales de *GenBank*, 104 son de virus y *unclassified*, que por un lado gran parte tienen un tamaño menor a 250 y por otro lado son categorías que producen la menor cantidad de *hits* en relación con el resto de familias.

Algunas muestras del detalle del filtrado para PDB se presentan en la Tabla 19. Para efectos de visualización solo se muestran las 10 primeras líneas por base de datos, una vista más completa de todos los datos filtrados se puede obtener en el Anexo VIII.

Tabla 19. Estadísticas del filtro de los alineamientos para la base de datos PDB.

Especie	Familia CAZy	Base de datos	Hits	Filtrados
Eukaryota	GH1	PDB3D	81.859	3.362
Eukaryota	GH10	PDB3D	18.039	1.645
Eukaryota	GH105	PDB3D	1.000	500
Eukaryota	GH11	PDB3D	31.151	198
Eukaryota	GH12	PDB3D	18.000	491
Eukaryota	GH13	PDB3D	78.112	4.538
Eukaryota	GH131	PDB3D	4.908	302
Eukaryota	GH14	PDB3D	17.094	582
Eukaryota	GH15	PDB3D	6.005	586
Eukaryota	GH16	PDB3D	19.010	1.939

En la gran mayoría de los casos se logró el objetivo de aumentar la cantidad de secuencias, algunos detalles de los 10 primeros registros se muestran en la Tabla 20 para PDB. Las tablas completas de resultados se relacionan en el Anexo IX. Comparativamente se puede decir que la desviación estándar mantuvo el mismo comportamiento que los datos iniciales, observándose familias con secuencias chicas y al mismo tiempo con secuencias muy grandes.

Tabla 20. Estadísticas de las secuencias filtradas base de datos PDB.

Especie	Familia CAZy	Base de datos	Cantidad de secuencias	Mínima Longitud	Máxima Longitud	Me dia	Desviación Estándar
Eukaryota	GH105	PDB3D	500	337	877	380	32,77
Eukaryota	GH10	PDB3D	645	252	3578	395	132,3
Eukaryota	GH11	PDB3D	198	256	1674	353	129,32
Eukaryota	GH12	PDB3D	491	334	1016	436	31,56
Eukaryota	GH131	PDB3D	302	214	810	327	75,68
Eukaryota	GH13	PDB3D	4536	332	3365	677	279,74
Eukaryota	GH14	PDB3D	582	367	1429	547	76,28
Eukaryota	GH15	PDB3D	585	403	1360	617	94,76
Eukaryota	GH16	PDB3D	1939	227	2225	349	123,43
Eukaryota	GH17	PDB3D	1003	251	2005	353	98,85

Se determinó que hubo un sesgo en algunas familias, lo cual se pudo presentar al aplicar el filtro de longitud de la secuencia de búsqueda (columna 3 del formato tabla, *query length* o *qlen*), debido a que todos los alineamientos con secuencias menores a 250 aminoácidos fueron automáticamente descartados. Cuando se presentaron familias completas cuyas secuencias no superaron el umbral, la familia fue consecuentemente borrada. En la Tabla 21 se muestra el listado completo de las familias que fueron eliminadas por el filtro.

Tabla 21. Listado de familias descartadas por el filtro.

Taxonomy	**PDB**	***Uniprot***	***GenBank***
Archae			
Bacteria			GH7
Eukariota	GH23, GH25, GH45		
Unclassified			GH112, GH12, GH24, GH25
Virus	GH104, GH19		GH13

El valor mínimo en la longitud de la secuencia tras el filtro fue de 146 aminoácidos, mientras que en los datos iniciales la secuencia con el menor tamaño tuvo un valor de longitud de 3 aminoácidos. El corte en tamaño de 250 aminoácidos y las secuencias de gran tamaño que pasaron el filtro, causaron un aumento en el promedio de la secuencia para las tres familias (ver

Tabla 22). El filtro de 250 aminoácidos fue aplicado sobre la secuencia de búsqueda (*query length* o *qlen*), y no sobre la secuencia objetivo (*subject length* o *slen*) lo cual explica el filtrado de algunas secuencias menores de 250. El límite inferior funcionó bien con longitudes de secuencias más cercanas a tamaños normales de proteínas (aproximadamente 180 en adelante) que contrastan con las longitudes de secuencias iniciales en cuyo límite inferior llegaban a los 10 aminoácidos que solo representan péptidos de corta longitud. Para el límite superior se pasó de tener una secuencia de 4.914 aminoácidos, a tener una secuencia de 15.080 aminoácidos.

Tabla 22. Variación de las longitudes de las secuencias por base de datos.

	Longitud Mínima	Longitud Máxima	Promedio
PDB	214	7.023	548
Uniprot	146	15.080	633
GenBank	146	15.080	609

La Figura 18 muestra el histograma de secuencias tras el alineamiento, en el cual se aprecia un comportamiento de distribución uniforme incluso mejor que los datos iniciales, aunque con una cola mucho más pronunciada de 15.080 de longitud.

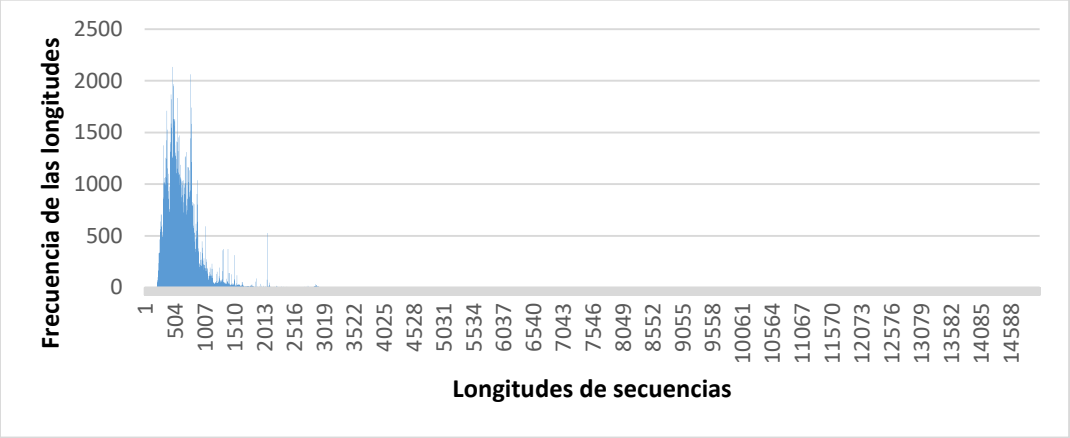


Figura 18. Histograma de longitudes de secuencias para la base de datos *GenBank* después del alineamiento.

4.1.3. Agrupación de las subfamilias utilizando OrthoMCL

Se formatearon los identificadores de las secuencias de los archivos individuales a la forma FAMILIA|IDENTIFICADOR como es requerido para ejecutar la clusterización de nuevas familias. Este ejercicio de recuperar las familias de las secuencias permitió descubrir que había ciertas secuencias de enzimas que dieron un *hit* con más de una familia, algunas incluso tenían hasta 4 posibles familias glicosil hidrolasas asociadas. Se generó un segundo archivo como un reporte de los identificadores que fueron encontrados

en otros archivos de diferentes familias, del tipo IDENTIFICADOR: FAMILIA1|FAMILIA2|…|FAMILIAn (Anexo X).

Posteriormente, se corrió mcl (Enright, et al., 2002) con valores de inflación de 1 y de 1,5. La relación de cantidades de los nuevos grupos se pueden ver en la Tabla 23, para información más detallada ver Anexo XI.

Tabla 23. Resultado general de la ejecución del MCL.

Base de datos	Familias CAZy	MCL 1	MCL 1.5
PDB	63	250	497
UNIPROT	197	5.993	10.822
GENBANK	231	4.191	6.874

El resultado de la ejecución del orthomcl fue un archivo de grupos clusterizados, donde cada línea representó un *cluster* y cada *cluster* mostró los identificadores de las secuencias asociadas.

4.1.4. Generación de subfamilias a partir de los resultados del OrthoMCL

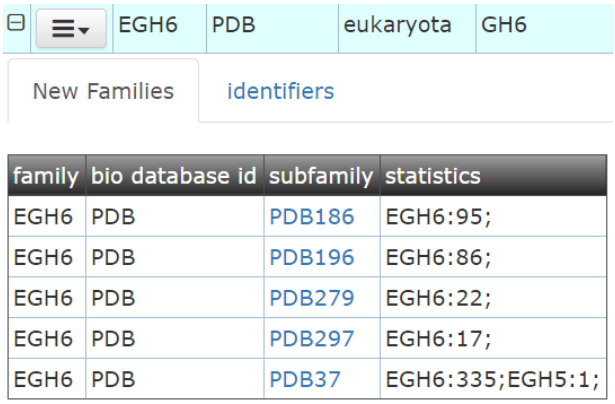
Se construyeron las nuevas subfamilias a partir de la salida del OrthoMCL donde cada *cluster* de identificadores fue una nueva subfamilia en un archivo fasta independiente con secuencias asociadas. Para esto se ejecutaron los comandos del Anexo XII. Algunas muestras de las características de las familias finales después del proceso de reagrupamiento se pueden observar en las Tabla 24, para PDB. Para efectos de reducción de espacio solo se muestran las 10 primeras caracterizaciones de las familias, una vista más completa de todos los datos finales se puede obtener en el Anexo XIII.

Tabla 24. Estadísticas finales base de datos PDB.

Grupo	Cantidad de secuencias	Mínima longitud	Máxima longitud	Media	Desviación estándar
PDB0	497	298	857	472	137,34
PDB1	488	394	1.870	543	135,75
PDB10	436	721	1.195	771	35,54
PDB100	246	421	1.212	489	66,35
PDB101	246	451	471	470	1,69
PDB102	240	323	1.191	439	125,69

PDB103	236	303	792	392	46,97
PDB104	235	305	1.033	414	65,88
PDB105	233	348	3.232	610	345,68
PDB106	233	373	380	379	0,58

Se generaron archivos de estadísticas para la base de datos PDB, *GenBank*, *Uniprot* que contienen la relación del número de secuencias que quedaron para cada familia del CAZy en cada uno de los *clusters, el Anexo XIV* contiene la estadística completa. A partir de esta información se generó una base de datos con las asociaciones de subfamilias finales generadas por el proyecto con las familias CAZy. Un ejemplo de la forma como quedó relacionada la familia GH6 de eucariotas para PDB se puede ver en la Figura 19; allí la columna ´*statistics*´ de la primera fila relaciona a la subfamilia PDB186 con la familia del CAZy GH6 de eucariotas (EGH6) con 95 secuencias asociadas.



family	bio database id	subfamily	statistics
EGH6	PDB	PDB186	EGH6:95;
EGH6	PDB	PDB196	EGH6:86;
EGH6	PDB	PDB279	EGH6:22;
EGH6	PDB	PDB297	EGH6:17;
EGH6	PDB	PDB37	EGH6:335;EGH5:1;

Figura 19. Subfamilias asociadas a la familia GH6 de Eukaryota

4.2. Validación *in silico* de los nuevos grupos de familias y subfamilias obtenidos, a través de la generación de perfiles proteicos, para su uso posterior en la anotación de secuencias de enzimas celulolíticas y hemicelulolíticas.

4.2.1. Construcción de la base de datos de enzimas celulolíticas y hemicelulolíticas

Se construyó en la base de datos la relación entre las funciones enzimáticas relacionadas con la degradación de celulosa y hemicelulosa con las familias del CAZy tal como fueron descritas en la Tabla 12. La Figura 20 muestra la forma como se crearon las asociaciones para la base de datos. De todas las glicosil hidrolasas procesadas en el proyecto, únicamente se filtran las subfamilias asociadas con las GH que efectivamente son celulolíticas y hemicelulolíticas.

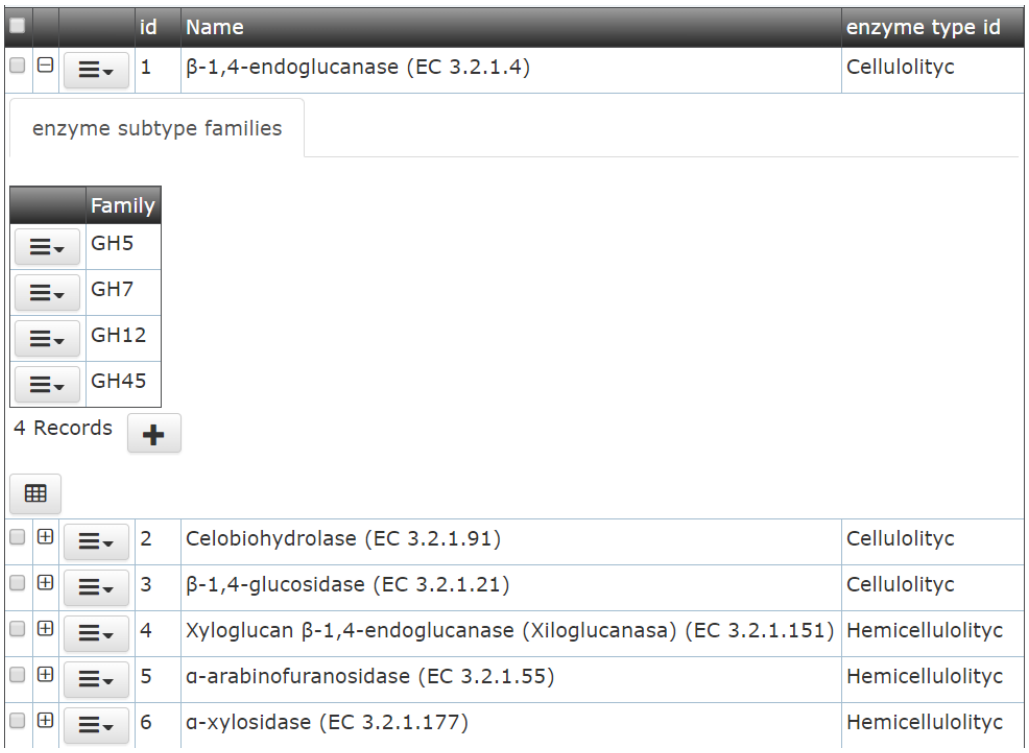


Figura 20. Funciones enzimáticas para degradación de celulosa y hemicelulosa con familias GH

4.2.2. Generación de perfiles proteicos utilizando modelos ocultos de Markov con HMMER3

Se ejecutó el alineamiento múltiple con formato *stockholm* y se crearon los perfiles de HMM (*Hidden Markov Models*) para cada una de las subfamilias finales, que junto con los archivos *fasta* de cada una de las subfamilias hacen parte de lo que es el *dataset* para el análisis de enzimas degradadoras de celulosa y hemicelulosa, siendo los archivos de secuencias *fasta* el insumo para hacer análisis con la herramienta Blastp y los archivos HMM el insumo para la anotación de enzimas con la herramienta HMMer3.

4.2.3. Validación de los resultados de las nuevas familias

A continuación, se describen los resultados para la familia GH6. Esta familia tiene funciones β-1,4-endoglucanasa (EC 3.2.1.4) y celobiohidrolasa (EC 3.2.1.91) que son actividades que corresponden en parte a la degradación de celulosa.

4.2.3.1. *Análisis de resultados para la familia GH6*

Se hizo un análisis básico del contenido filogenético de la familia GH6 y de los subgrupos que quedaron tras el procesamiento de datos. En la Figura 21 se muestra el árbol filogenético para GH6 disponible en el PFAM. La familia está constituida en un 50 % por enzimas de bacterias y el otro 50% por hongos. Los hongos están compuestos en su mayoría, en un 75% aproximadamente por la clase Ascomicetos de la familia Saccharomicetales y el resto por Basidiomicetos mayormente de la familia de los Agaromicetos. Se hayan ambas actividades enzimáticas, sobresaliendo la función endoglucanasa sobre la celobiohidrolasa, ésto se debe a la especificidad del sustrato que tiene la celobiohidrolasa. En cuanto a bacterias la variación de familias fue mucho mayor, se halló un 90% de Actinobacterias de las familias Streptomices, Micronosporales (Micromonosporas, Actinoplanos), Frankiales y Celulomonas, entre otras. El 10% restante se asoció a Proteobacterias tipo Gammabacterias (Xanthomonadales) y Deltabacterias. La distribución de las actividades enzimáticas fue similar a la que tienen los hongos, siendo mayor la actividad endoglucanasa sobre la celobiohidrolasa.

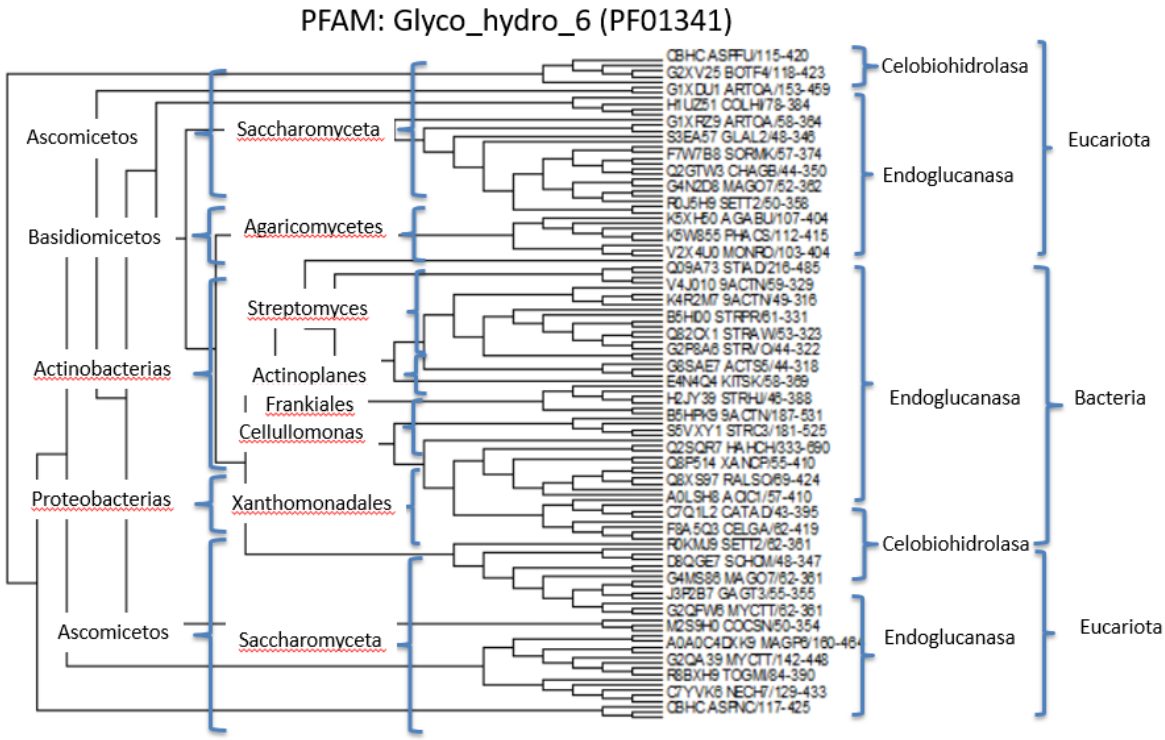


Figura 21. Árbol filogenético familia GH6 PFam.
Fuente: Modificado de xfam.org

Se analizó también la conformación de las subfamilias más representativas para bacterias del Genbank con la familia GH6, que tuvieran relacionadas más de 150 miembros de la familia (Tabla 25). Las subfamilias GENBANK847, GENBANK466, GENBANK233, fueron 100% caracterizadas para anotación de endoglucanasas, donde la primera y la última estuvieron compuestas casi en un 100% por la familia streptomices y la segunda incluyó un 20% de la clasificación de proteobacteria. Las familias GENBANK855 y GENBANK756 fueron subfamilias similares en cuando a variabilidad de los organismos y contienen clasificación de celobiohidrolasa y endoglucanasa. Lo anterior se puede explicar a través de trabajos realizados por Meinke (1995) donde en un experimento con Thermomonospora fusca, un tipo de actinobacteria de la familia Streptosporangiales, esta fue capaz de cambiar de actividad exo a endo solo al recortarle el bucle extendido de la celobiohidrolasa. Estas dos familias son interesantes en cuanto a que muestran ese comportamiento híbrido, lo que permite hacer anotación de enzimas en esta doble función catalítica, pero se requiere de mayor investigación para llegar a esta conclusión.

La GENBANK321 se trató de una subfamilia formada por hongos Ascomicetos y Basidiomicetos en vez de subfamilias bacterias como idealmente era previsto, esto ocurrió

en el segundo paso del proceso metodológico donde se realizó el alineamiento de las secuencias de CAZy contra la base de datos NR de secuencias no redundantes, donde muchas secuencias correspondientes a Actinobacterias de la familia Streptomices hicieron *match* con secuencias de hongos. Lo anterior se pudo presentar debido a que las actinobacterias o actinomicetos tiene una similaridad con los hongos (de ahí que tengan sufijo myces que significa hongo). Otra hipótesis conduce a la posibilidad que en algún punto de la evolución se haya dado un evento de *transferencia horizontal de genes* y a estos hongos se les haya transferido este gen que codifica este tipo de celulasas, y de ahí su similitud. Esta subfamilia además fue la que tuvo mayor cantidad de celobiohidrolasas sobre endoglucanasas, siendo un buen candidato para anotación de celobiohidrolasas específicamente. Sin embargo, estas subfamilias en estas condiciones deben ser reubicada a la familia correspondiente en eucariotas.

Tabla 25. Análisis de subfamilias de bacterias GH6 para *Genbank*.

Familia	Subfamilia	Genero	Función
GH6 – Bacteria - Genbank	GENBANK855	Actinobacteria variadas entre Frankiales, Glicomicetales, Micronosporales (Micromonosporas, Actinoplanos) Pseudonocardiales, Streptosporangiales y Streptomycetales	Celobiohidrolasa (42) 23%, Endoglucanasa (124) 66%, celulasa GH6 no definida (20) 10%
	GENBANK847	98% Actinobacteria –> Streptomycetales -> streptomices	100% endoglucanasa
	GENBANK764	96% Actinobacteria –> Streptomycetales -> streptomices	Endoglucanasa (178), celobiohidrolasa (4), celulasa GH6 no definida (10), Streptomices hipotéticos sin función definida (30)
	GENBANK756	Actinobacteria de genero variado Streptomycetales, Pseudonocardiales,, Micromonosporales, celulomonas entre otras	Endoglucanasas (26), celobiohidrolasa (5), celulasa GH6 (130), celulasa GH5 (6), proteínas hipotéticas (57)
	GENBANK466	80% actinobacterias de genero streptomices, Pseudonocardiales y Micromonosporales, 20% proteobacterias Xanthomonadales y Myxococcales	100% endoglucanasa
	GENBANK321	80% Ascomicetos, 20% basidiomicetos	Endoglucanasa (5), celobiohidrolasa II (108), celulasa GH6 no definida

			(29), proteínas hipotéticas (60)
	GENBANK233	99% Actinobacterias -> Streptomycetales -> streptomices	100% Endoglucanasa

El subgrupo 233 quedó definido para la taxonomía *Bacteria, Actinobacteria, Streptomycetales, Streptomycetaceae, Streptomyces,* lo que indica que el subgrupo tiene una mayor especificidad lo que permite una caracterización más exacta del modelo para anotación de enzimas con esta característica.

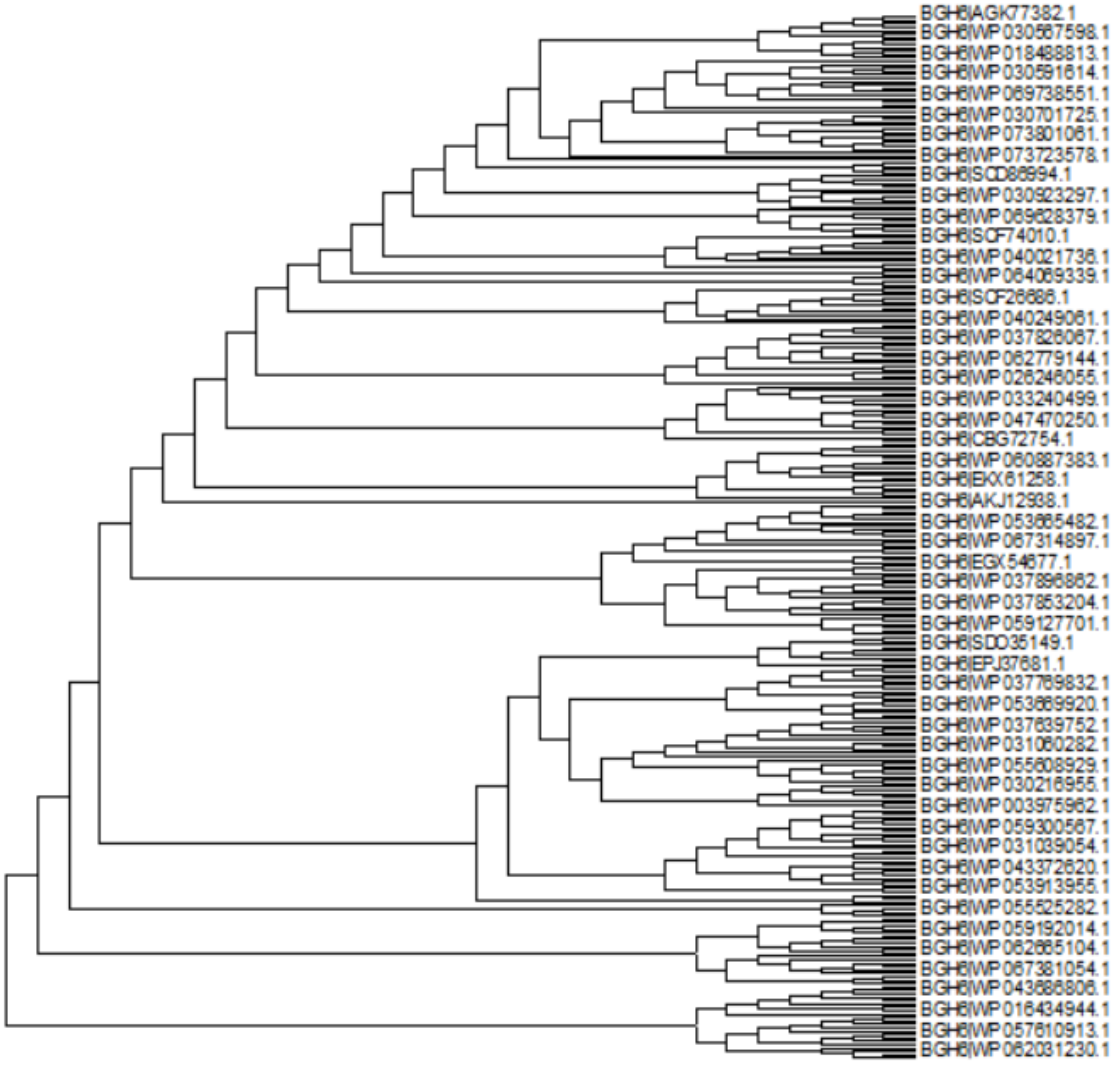


Figura 22. Árbol filogenético de la subfamilia GH6 Bacteria Genbank 233.

Para el análisis de hongos se analizó la familia GH6 de eucariotas del PDB cuya subfamilia más representativa fue la PDB37 (Tabla 26). Los resultados fueron variables

teniendo un mayor filtro de enzimas con función celobiohidrolasa en vez de endoglucanasa. La PDB279 filtró proteobacterias en vez de solo hongos por las mismas razones que en bacterias se filtraron ascomicetos y basidiomicetos.

Tabla 26. Análisis de subfamilias de eucariotas GH6 del PDB.

Familia	Subfamilia	Genero	Función
GH6 – Eukariota - PDB	PDB186	90% ascomicetos, 10% basidiomicetos	Celobiohidrolasa (25), endoglucanasa (13), celulasa GH6 no definida (16), proteína hipotética (37)
	PDB196	90% ascomicetos, 10% basidiomicetos	Celobiohidrolasa (33), celulasa GH6 no definida (22), proteína hipotética (28)
	PDB279	60% Chytridiomicota -> Neocallimastigomicetes, 40% proteobacteria -> Myxococcales	Celobiohidrolasa (4), Celobiosidasa (7), celulasa GH6 no definida (7)
	PDB297	100% basidiomicetos -> agaromicetos	Celobiohidrolasa (2), celobiosidasa (1), celulasa GH6 no definida (7), proteína hipotética (7)
	PDB37	70% Ascomicetos, 25% basidiomicetos, 5% otros	Endoglucanasa (5), celobiohidrolasa (133), celulasa GH6 no definida (75), proteína hipotética (75)

4.2.3.2. Afinación del modelo para la subfamilia PDB0

Se hizo una prueba de afinación del modelo *HMM* para la subfamilia PDB0 utilizando el *script* createHMMFromDataSplit el cual analizó la precisión del modelo automáticamente así: (i) un porcentaje de los datos los utilizó para armar el modelo, (ii) el resto del porcentaje de datos, los empleó para analizar los verdaderos positivos y (iii) unas muestras de los datos no pertenecientes a la familia se usaron para evaluar los verdaderos negativos.

El resultado para este grupo se muestra en la Figura 23. Marcado en rojo están los verdaderos negativos, en azul los datos del modelo y los puntos verdes son los datos pertenecientes al grupo que fueron sometidos a prueba. Se evaluaron los datos con base al *score* del HMMSearch, y finalmente se pudo determinar que dada la separación de la muestra los datos del modelo fueron correctos.

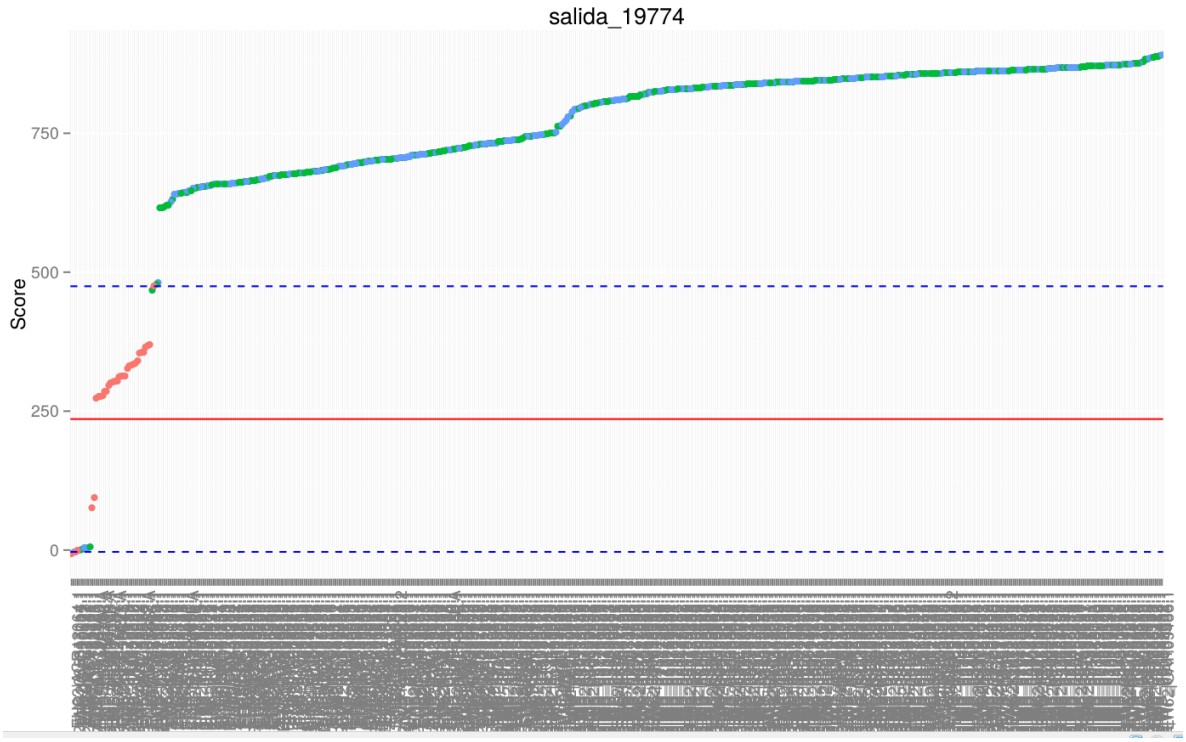


Figura 23. Test de falsos positivos para la familia GH6.

En este punto se estableció un valor umbral para clasificar la subfamilia, que para PDB0 el elegido fue de 550, que corresponde a un valor promedio entre el último *score* clasificado de la agrupación más grande y el *score* para el verdadero negativo más alto. Algunos verdaderos positivos (1,2%) fueron puestos por debajo del umbral. Así valores por encima de 550 del *score* evaluados por este modelo *HMM* pueden ser considerado clasificatorios como candidatos a pertenecer a la subfamilia PDB0 con un error de 1,2%.

4.2.3.3. Resultados del test con la familia GH6

Se realizaron los *test* de los modelos HMM para el grupo GH6, haciendo la comparación de los resultados del *HMM* del PFAM y los *HMM* del subgrupo GH6 perteneciente a bacterias. La Tabla 27 muestra las actividades de la familia GH6.

Tabla 27. Actividades familia GH6.

Familia de prueba	Actividades
Glicosil Hidrolasa 6 (GH6)	Endoglucanasa (3.2.1.4), Celobiohidrolasa (3.2.1.91)

La base de datos de prueba estuvo conformada por las secuencias no redundantes de Refseq la cual contiene un total de 50.737.205 secuencias. La búsqueda de prueba se hizo con la función *hmmsearch* del *software* HHMer3, los resultados se ven en la Tabla 28.

Tabla 28. Número de secuencias filtradas por el HMMSearch.

	Secuencias sobre el umbral
Glyco_Hydro_6 (PFAM - HMM)	2.471
GH6 – Bacteria (Genbank)	9.560

Del subgrupo GH6 – Bacteria se hizo la revisión manualmente. La matriz de conjunción creada se aprecia en la Tabla 29.

Tabla 29. Matriz de conjunción subgrupo GH6 – Bacteria.

	Matriz de conjunción
Verdaderos positivos	6.438
Falsos positivos	627

Los falsos positivos quedaron divididos en tres tipos de enzimas diferentes (*Tabla 30*), donde todas tienen relación por el tipo de actividad con la familia GH5. Se sabe que la familia GH5 tiene gran relación con la familia GH6, hay secuencias que son multi-especies y que están en ambas familias. Por lo que la determinación de los falsos positivos necesita una mayor curación y los resultados presentados son parcialmente concluyentes con respecto a la afinación del modelo.

Tabla 30. Distribución los falsos positivos.

Nombre enzima	**Número EC**	**CAZy**	**Cantidad**
Xiloglucanasa	3.2.1.151 (Hidrolisis de enlaces 1,4-D-glicosidicos en xiloglucanos)	GH5	267
Arabinofuranosidasa	3.2.1.55 (Hidrolisis de terminales de alfa-L-arabinofuranosa)	GH2, GH3, GH5, GH43, GH51	126
Beta-manosidasa	3.2.1.25 (Hidrolisis de beta-D-mananasa)	GH1,GH2, GH5	234

4.3. Desarrollo de una herramienta para la anotación de secuencias de enzimas degradadoras de celulosa y hemicelulosa

4.3.1. Programación de un servicio para anotación de enzimas degradadoras de celulosa y hemicelulosa

A partir de todos los datos y archivos recolectados, se relacionó la información del CAZy y de los resultados finales en una misma base de datos estructurada, de donde se creó la herramienta de consulta que se le dio por nombre LignoSearch (Figura 24). Recibe su nombre así por el hecho de que debe ser una herramienta para todos los componentes de la lignocelulosa, de momento solo los módulos de celulosa y hemicelulosa fueron realizados, en futuras versiones será incluida la lignina.

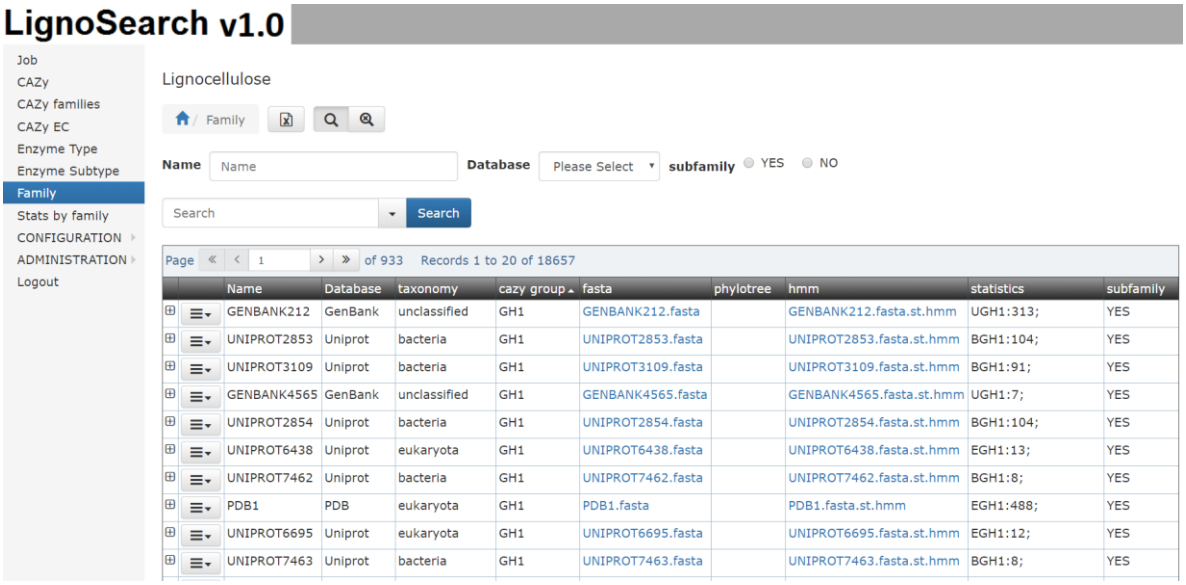


Figura 24. Herramienta de consulta visual LignoSearch.

Esta base de datos contiene no solo los resultados de los archivos finales de cómo quedaron las subfamilias, también está relacionada la familia de CAZy con la cual está relacionada cada subfamilia. Cada detalle de la subfamilia presenta una estadística de la cantidad de secuencias que quedaron por cada *cluster*, incluso se puede ver si este fue formado por la relación de una o más familias.

También permite ver cada una de las familias donde se presenta en detalle la información perteneciente a cada grupo, los archivos relacionados, e incluso se obtuvo la

filogenia de cada una de las secuencias, relacionando la secuencia con el organismo a la cual pertenece. A partir de la taxonomía de cada secuencia se armó el árbol filogenético basado en las especies que tenía cada familia (Ver Figura 25).

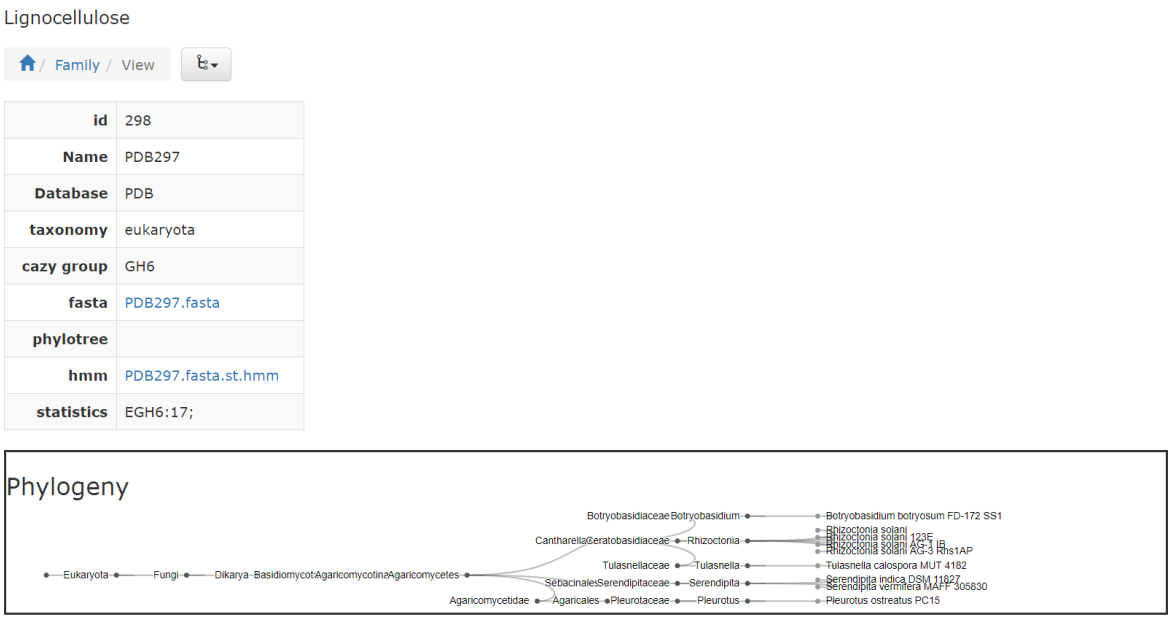


Figura 25. Vista de los taxones de una subfamilia específica en la herramienta LignoSearch.

La construcción del árbol filogenético permitió determinar que la gran mayoría de subgrupos quedaron constituidos en relación con subespecies de la familia principal, lo que permite hacer investigación más precisa en relación a cultivos de enzimas mucho más diferenciados.

Se desarrolló también un *script* de Python v3.5.2 llamado *searcher.py* (*Anexo XV*) que se encarga de ejecutar los trabajos de búsqueda. Acepta cinco parámetros que se muestran en la Tabla 31.

Tabla 31. Parámetros de ejecución del *script* searcher.py.

Nombre	Parámetros
Carpeta de salida de los resultados	-o
Archivo de entrada a analizar	-i
Software a ejecutar (HMMer3, Blast)	-s
Bases de datos de referencias de las subfamilias	-r
Base de datos biológica (***Genbank***, PDB, ***Uniprot***)	-d

Una ejecución normal por medio del *script* tiene por salida una carpeta que contiene los resultados de los alineamientos o búsquedas *HMM,* según el software que se haya elegido para este propósito, que se ejecutan contra los *datasets* definidos de subfamilias de enzimas celulolíticas y hemicelulolíticas, y también un reporte final que contiene los *hits* de los alineamientos con las secuencias y resultados de las búsquedas con los perfiles proteicos.

4.4. Resultados extras de aplicación general

Tras la ejecución de todos los pasos planteados para el proyecto y dados los hallazgos en los resultados de las secuencias finales, se detectaron algunos problemas relacionados con la metodología escogida para la ejecución del proyecto, a continuación, se enumeran todos en orden de aparición.

4.4.1.1. Familias del CAZy incompletas o sin ninguna secuencia relacionada

Buena parte de los problemas relacionados con los datos incompletos del proyecto en su etapa inicial se deben a que los identificadores de las secuencias que se tomaron de la página *web* del CAZy, fueron adquiridos manualmente por estudiantes, lo que insertó un error humano. Esto causó por ejemplo que una taxonomía para una base de datos quedara totalmente vacía, o que familias tuvieran secuencias incompletas.

Para el año 2016, ya cuando los datos habían sido recolectados, se publicó la herramienta llamada CAZy parser (Honorato, 2016) que permite hacer descarga automática de toda la base de datos de identificadores de la página *web* del CAZy. Tomar en cuenta este tipo de herramientas automáticas hubiera eliminado los problemas mencionados y, además, la recolección de datos se haría de una manera más rápida. Para el final de este proyecto y con el fin de complementar la aplicación *web* LignoSearch con los datos del CAZy, se utilizó CAZy parser para descargar todos los identificadores del CAZy.

4.4.1.2. Familias con una cantidad muy pequeña de secuencias

Existen familias del CAZy que solo poseen una sola secuencia, tal es el caso de la familia Glicosil Hidrolasa 27 de archaea que posee el identificador ADB63458.1 para *Genbank*. Haber aplicado la metodología planteada con los alineamientos pudo causar que

existieran *hits* de secuencias que realmente no correspondiera a la intención de mantener las mismas familias siempre, es más preciso reconstruir familias tras realizar alineamientos de múltiples secuencias de entrada, que hacerlo para una sola secuencia. Realizar modelos sobre familias de una sola secuencia puede causar un mal modelo que tal vez no describa bien las características fundamentales de la familia inicial.

4.4.1.3. Fragmentos de enzimas y péptidos que fueron obtenidos de CAZy, pero que no corresponden a enzimas

En el transcurso de este proyecto y tras realizar mediciones estadísticas básicas, se descubrió que ciertas secuencias descargadas del CAZy no son enzimas de carbohidratos, sino que son péptidos que no se pueden considerar como enzimas propiamente dichas. La Tabla 32 se muestra algunos casos particulares que presentan esta anormalidad, y la Figura 26 muestra la prueba de la existencia de estas secuencias en la página de CAZy.

Tabla 32. Muestra de secuencias que presentan anormalidades.

Id de la secuencia	Base de datos	Familia CAZy	Reino	Tamaño (Cantidad de aminoácidos)
P85974.1	Genbank	GH50	Bacteria	10
P29261	Uniprot	GH1	Eukariota	10
P80072	Uniprot	GH13	Bacteria	15

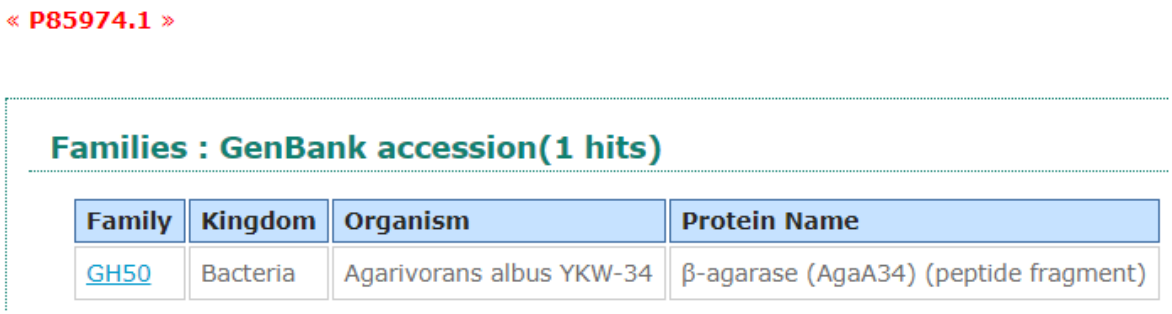


Figura 26. Prueba de la existencia de secuencias irregulares en CAZy.

CAZy no solo publica datos de enzimas sino también fragmentos de péptidos lo cual fue una limitante en la investigación.

4.4.1.4. Identificadores del CAZy contenían caracteres especiales

Algunos de los identificadores que fueron descargados para la base de datos PDB, principalmente, tenían caracteres especiales, como 1ZB5:B, siendo la proteína real solo 1ZB5 sin extensiones, esto ocasionó que para ese identificador se haya descargado unos caracteres incorrectos correspondientes a secuencias de no más de cuatro aminoácidos, lo que al momento de hacer los alineamientos causaron error al hacer match con las secuencias objetivo equivocadas.

4.4.1.5. Categoría Unclassified del CAZy inútil para la investigación

En este trabajo se incluyeron todas las categorías que estaban disponibles en el CAZy, bacterias, eucariotas, arqueas, virus e incluso la 'No clasificadas' o '*Unclassified*'. Esta última recibe este nombre porque se tratan de secuencias que, si bien se han detectado que cumplen con uno de los patrones de alguna de las familias del CAZy, aún no se saben de qué organismo pertenecen porque generalmente son secuencias de patentes a las cuales no se les publicó la especie de donde lo tomaron.

Para efectos del desarrollo de un proyecto como este en que uno de los enfoques principales era mantener las tres bases de datos separadas por el tipo biológico (PDB, *Genbank* y *Uniprot*) y por la taxonomía (bacterias, eucariotas, arqueas y virus), esta clasificación no fue útil, dado a que ella tiene secuencias de patentes que corresponden a diversas taxonomías, es decir, puede haber una mezcla de bacterias y eucariotas o arqueas y virus, o todas juntas, con lo que si un investigador quiere buscar por esta categoría no le va a dar una información correcta y útil para una especie objetivo que desee investigar.

4.4.1.6. Problemas con secuencias multifamilia del CAZy

Otra de las ideas principales del proyecto era mantener bases de datos que discriminaran de la manera más precisa si una secuencia pertenecía a una familia o no. La realidad es que ya por las propias definiciones del CAZy existen secuencias que pueden pertenecer a múltiples familias al mismo tiempo, por ejemplo, puede ser una secuencia que pertenezca a la familia Glicosil Hidrolasa 5 y también a la familia Glicosil Hidrolasa 6, o incluso a más familias. Este tipo de situaciones causan que, en el desarrollo del modelo, HMM (*Hidden Markov Model*), pensado para hacer búsqueda para la familia GH5, filtre secuencias que no sean propias de este grupo, y termine filtrando secuencias de la familia GH6. El problema es mayor en la medida que la familia que se evalúe posea una baja

cantidad de secuencias para crear el modelo, la probabilidad de error aumenta. De acuerdo con lo anterior era necesario verificar cuáles de estas proteínas son multiespecie, y de serlo preferiblemente quitarlas para que el modelo filtre bien solo familias específicas, o en el mejor de los casos, crear clasificaciones intermedias que consideren estos casos como si fuera una misma familia.

4.4.1.7. *Hits con secuencias hipotéticas y patentes*

También en el proyecto se planteó la condición de que los *dataset* de secuencias y modelos trabajaran con datos reales y que al momento del *script* arrojar un reporte de resultados, las enzimas relacionadas se pudieran adquirir ya sea comercialmente o en cultivos de laboratorios del grupo investigador.

La base de datos de secuencias no redundantes NR del NCBI la cual se utilizó en la primera parte el proyecto, contiene todo tipo de secuencias registradas, desde secuencias curadas hasta secuencias hipotéticas, de manera que, así como quedó el ejercicio realizado, al momento de hacer los alineamientos pareados de esta base de datos con las secuencias iniciales del proyecto, algunos hits de secuencias hipotéticas también fueron filtrados, con lo cual en adelante todo el desarrollo de los modelos son mucho más afines a encontrar este tipo de secuencias que si se hubieran quitado desde el principio, y que para el objetivo del proyecto son totalmente inútiles.

4.4.1.8. *Se aplicó el filtro de cobertura del query sobre el subject, pero no al contrario*

En el paso de aplicar el filtro de identidad y cobertura de los alineamientos, se utilizó para la cobertura la columna del formato tabla llamada *qcovs (query coverage subject),* este es un valor que da un porcentaje de cuánto cubre la secuencia *query* a la secuencia objetivo que para el filtro se planteó una cobertura de 80%.

Para los casos en que ambas longitudes de secuencias poseían un tamaño similar en cantidad de aminoácidos, haber aplicado este filtro no supone ninguna dificultad en el proyecto, el problema está cuando la secuencia de búsqueda o *query* es mucho mayor en tamaño que la secuencia objetivo o *subject,* porque una secuencia así puede contener la gran mayoría de aminoácidos de la secuencia objetivo dado a que su tamaño se lo permite y no necesariamente ser una secuencia similar en funcionalidad, evaluar la cobertura del *subject* sobre el *query* hubiera reducido el problema. La situación empeoró aún más cuando

se sabe que se incluyeron péptidos como si fueran enzimas dentro del procesamiento y cualquier péptido pudiera estar contenido en una enzima mucho más grande en tamaño lo que pudo haber incluido muchos falsos positivos innecesarios.

4.4.1.9. *Filtro de tamaño fijo eliminó familias CAZy enteras*

El filtro de tamaño de 250 aminoácidos que se aplicó para el proyecto, causó que muchas secuencias fueran descartadas dado a que no cumplían con un tamaño de secuencia mayor o igual a este valor, incluso hubo familias enteras que fueron descartadas completamente tras aplicar el filtro, como lo fue la familia GH25 de eucariota del PDB. El haber aplicado un filtro de secuencia dinámico que se tomó como base la longitud mínima de secuencia por familia habría solucionado el problema, de manera que, si el tamaño mínimo de secuencia para la familia es de 200, el filtro debe manejar este número y no un tamaño fijo como originalmente se hizo.

4.4.1.10. *El filtro de tamaño se aplicó sobre la longitud de la secuencia query y no sobre la secuencia subject*

El filtro de tamaño en el proyecto fue aplicado sobre el parámetro *qlen query length* y no sobre el parámetro *slen subject length* el cual más correcto. El resultado final que se logró de la manera como se aplicó se hubiera podido alcanzar solamente habiendo filtrado las secuencias con longitud mayor o igual a 250 de las secuencias del CAZy e incluso se hubiera logrado que el tiempo de procesamiento de los datos hubiera sido menor a como fue realmente.

4.4.1.11. *El filtro de identidad como se planteó no fue totalmente efectivo*

Se descubrió tras ejecutar todos los pasos que la forma como se aplicó el filtro de identidad filtraba la mayoría de las secuencias relacionadas con la misma familia dentro de los dominios de glicosil hidrolasas, pero no servía para aplicarlo como regla general para hacer filtro de todas las secuencias.

En un caso particular de la familia glicosil hidrolasa 4 la secuencia de identificador ACX49739.1 (Figura 27) para la proteína *siderophore 2,3-dihydroxybenzoate synthesis-like protein [uncultured marine bacterium 1n22]* hizo *match* con la secuencia de identificador WP_067593936.1 llamada *non-ribosomal peptide synthase/polyketide synthase [Nocardia*

terpenica], pasando el filtro de 80% de cobertura, 40% de identidad y longitud de la secuencia > 250 aminoácidos.

« ACX49739.1 »

Families : GenBank accession(1 hits)

Family	Kingdom	Organism	Protein Name
GH4	Bacteria	uncultured marine bacterium 1n22	ORF

Figura 27. Existencia de la enzima ACX49739.1 en CAZy.

La Figura 28 se muestra un resultado aproximado del alineamiento por consola. En este alineamiento en la herramienta *web* de Blastp se muestra la cobertura es de 83% e identidad de 36% lo cual indica que ambas proteínas hacen *match* para las reglas establecidas para el proyecto. La identidad está un poco reducida dado a que la versión *web* no utiliza un algoritmo tan preciso como si lo utiliza el alineamiento por consola de comandos.

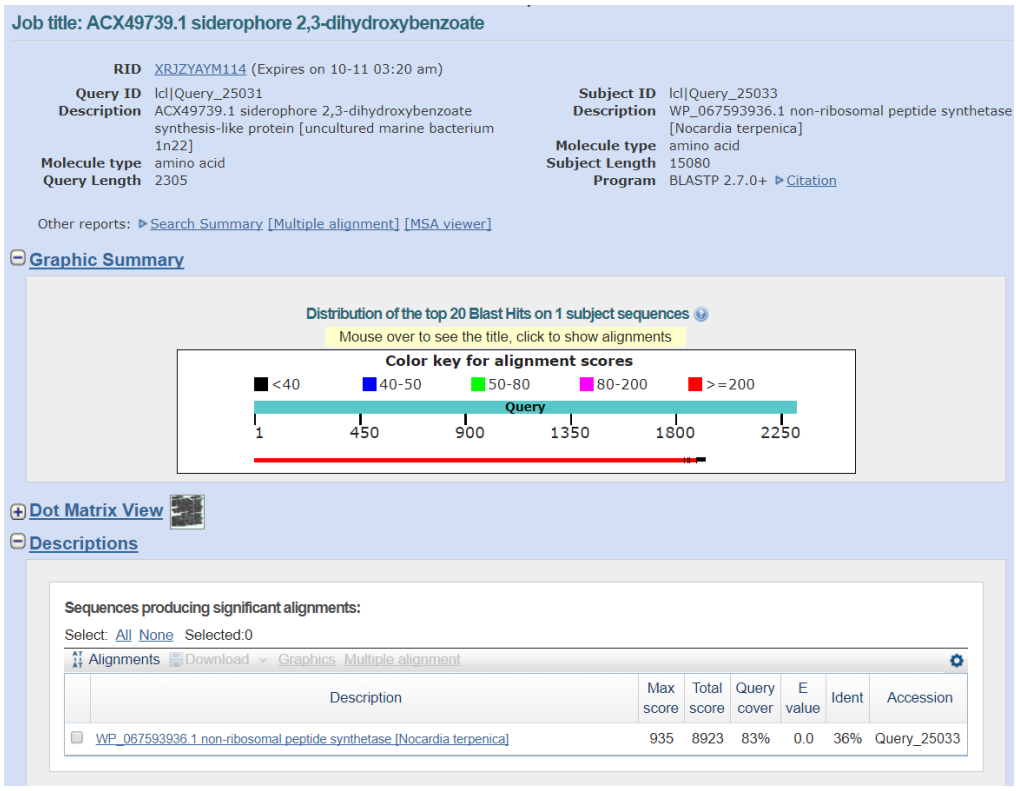


Figura 28. Alineamiento de la enzima ACX49739.1 contra WP_0675936.1.

Dada su inusual tamaño de 15.080 aminoácidos comparado al máximo tamaño para la familia glicosil hidrolasa 4 de 2.305, se decidió hacer un análisis más detallado de la composición de ambas proteínas y como eran relacionadas para lo cual se usó el software Interpro Scan (Jones et al., 2014).

Para el caso de la secuencia de CAZy ACX49739.1, la herramienta mostró todos los dominios para la proteína donde además del dominio de lactato deshidrogenasa GH4 también es necesario ver que existen los dominios sintetasa de AMP (Adenosin mono fosfato), dominio de adenilación de aminoácido y domino de condensación, estos tres últimos cubre más del 50% de los dominios funcionales de esta proteína (Figura 29).

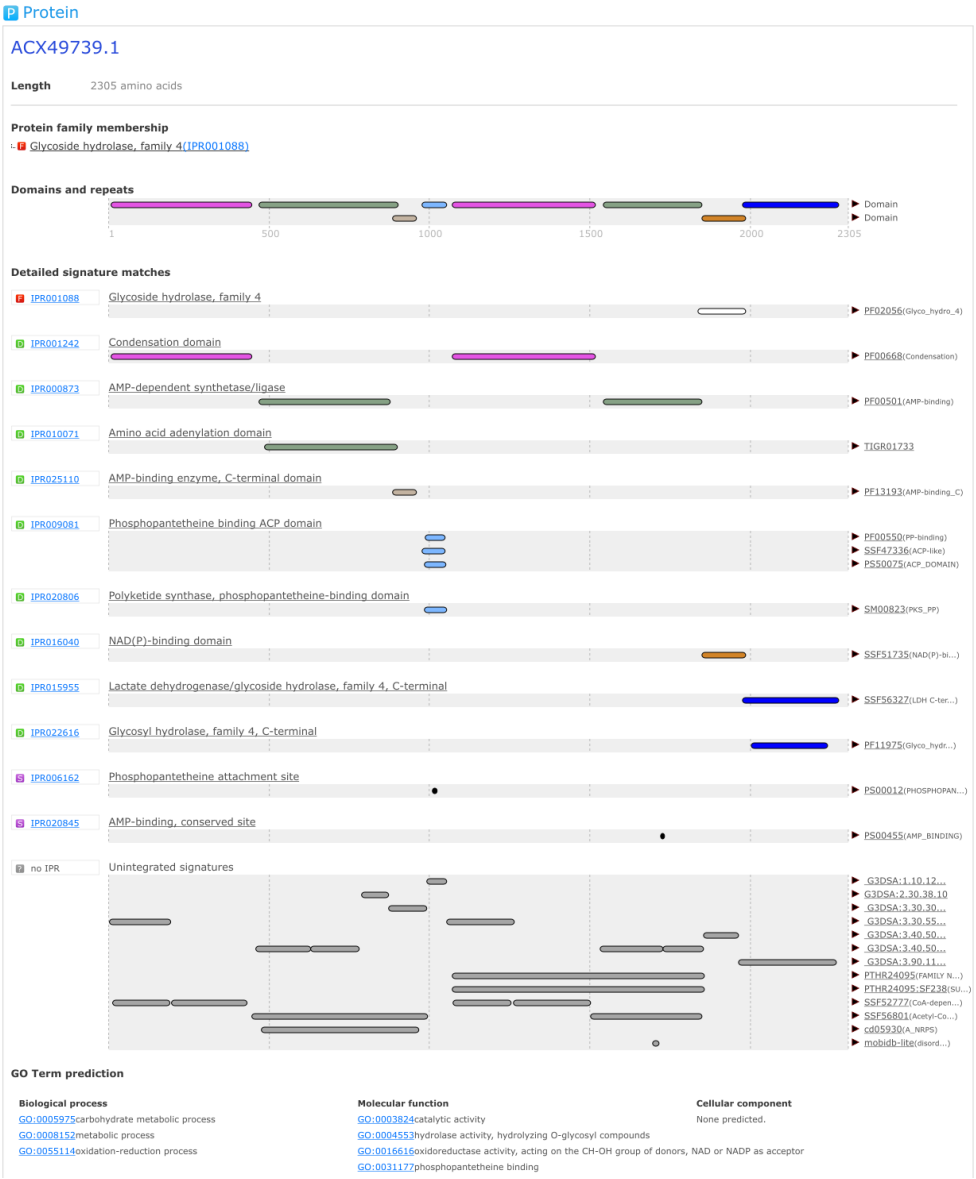


Figura 29. Ejecución de InterPro Scan para ACX49739.1.

Para el caso de la secuencia objetivo se realizó el mismo ejercicio, se detectó que ninguno de los dominios descritos por el InterProScan correspondían a la familia GH4 que, si estaba para la secuencia *query*, pero en cambios los dominios sintetasa de AMP (Adenosín mono fosfato), dominio de adenilación de aminoácido y domino de condensación si estaban presentes recorriendo casi que toda la proteína (Figura 30).

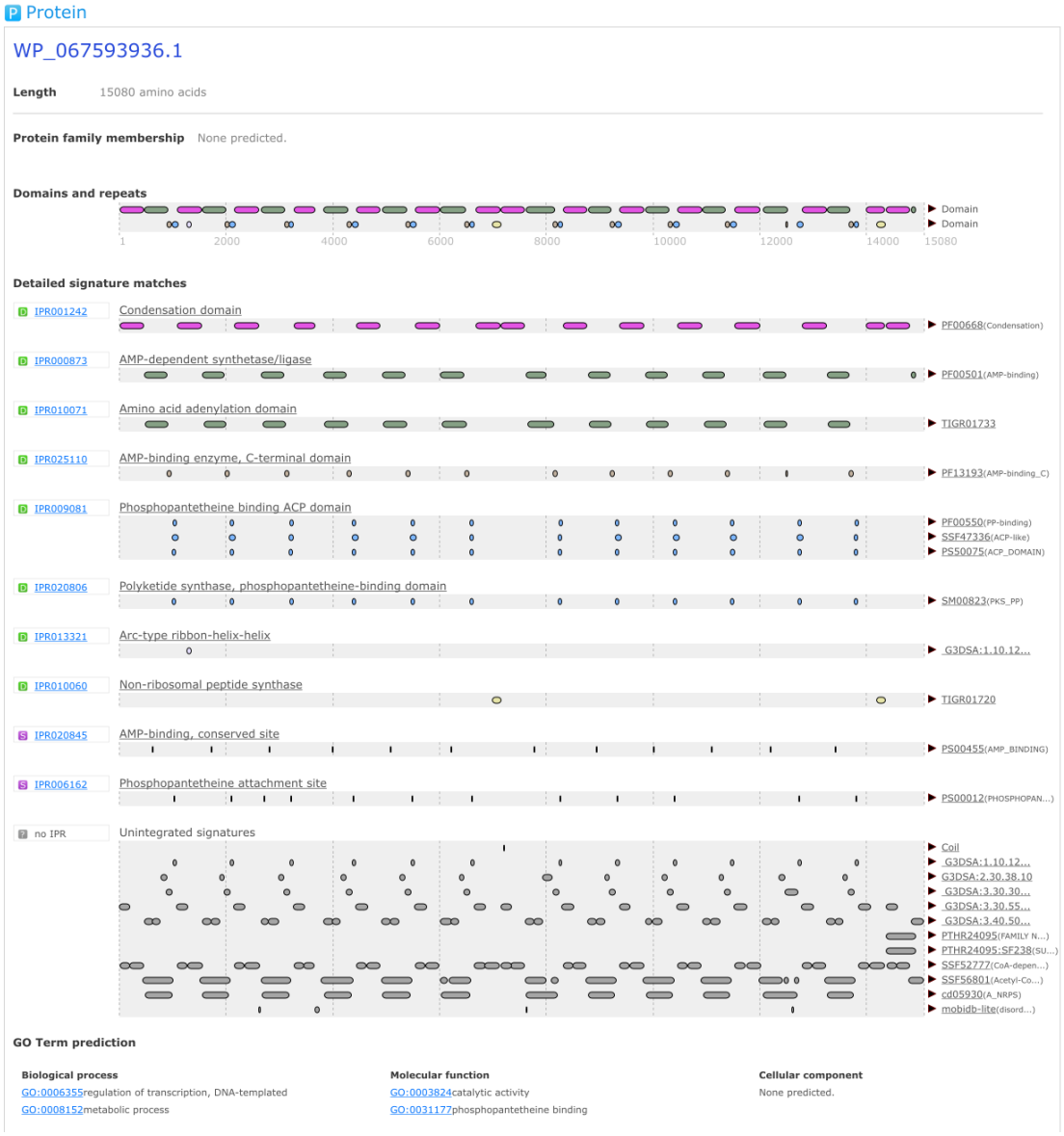


Figura 30. Ejecución de InterPro Scan para WP_067593936.1.

Se determinó que efectivamente existía la identidad entre ambas secuencias, pero esto no se debía a una relación existente de GH4, sino a la relación de las funciones de síntesis

de AMP de ambas proteínas, con lo cual se concluyó que, si bien el valor de identidad podría dar una idea de la probable relación de dos proteínas como pertenecientes a familias de glicosil hidrolasas, no era necesariamente una regla aplicable a todos los casos. Para trabajos futuros se recomienda hacer filtro por identidad de los dominios de glicosil hidrolasas, no por el valor de identidad de la secuencia alineada.

4.4.1.12. *La verificación y depuración de los subgrupos debe realizarla un profesional en biología*

La depuración de los grupos debe realizarla un experto que entienda de enzimas, la formación de los grupos como quedaron actualmente todavía contienen algunas secuencias que no corresponden al grupo, pero esa verificación si implica saber mucho más de la formación de dominios y propiamente de toda la secuencia. Lo mismo ocurre con la verificación del grupo haciendo la matriz de conjunción en la cual para determinar si efectivamente hay falsos positivos para la definición de un grupo específico, implica también un conocimiento la forma de cada familia, actualmente con la prueba manual solo se dio una revisión basado en las características más visibles por lo cual los falsos positivos que fueron puestos realmente requieren analizar mejor las secuencias para confirmar que se trata efectivamente de un falso positivo.

4.4.1.13. *Las enzimas que corresponden a la degradación de lignina no fueron totalmente abordadas en el proyecto*

Las enzimas mostradas en la Tabla 33 no fueron tenidas en cuenta para el desarrollo del proyecto y todas son enzimas que actúan en la degradación de lignina, esto ocurrió porque en la investigación solo fue tenido en cuenta la clasificación Glicosil Hidrolasas (GH) del CAZy, y todas esas enzimas pertenecen a la clasificación Actividades Auxiliares (AA) que no hacen degradación por hidrolización de enlaces glicosídicos.

Tabla 33. Enzimas degradadoras de lignina no tenidas en cuenta.

Enzima	Clasificación CAZY
Ligninperoxidasa (LiP)	AA (Actividades Auxiliares)
Manganesoperoxidasa (MnP)	AA (Actividades Auxiliares)
Peroxidasa versátil (PV)	AA (Actividades Auxiliares)
Peroxidasa manganeso independiente	AA (Actividades Auxiliares)

5. Conclusiones generales

5.1. Contribuciones de la Tesis

Con los resultados obtenidos con la tesis se pueden determinar las siguientes contribuciones del proyecto:

- Se desarrolló de una aplicación *web* de consulta visual sobre las diferentes familias y subfamilias que quedaron del CAZy.
- Una aplicación *web* que permite la ejecución de búsquedas de enzimas y que se puede usar como un servicio público para beneficio de investigaciones relacionadas con la degradación de celulosa y hemicelulosa.
- La aplicación de un nuevo enfoque que amplía la usabilidad en las bases de datos públicas como lo son PDB, GenBank y Uniprot en la investigación en celulosa.
- Una nueva forma de abordar la búsqueda de enzimas partiendo del tipo de función, familia CAZy o tipo de compuesto (Celulosa, hemicelulosa, lignina), que no se había contemplado en ninguno de los antecedentes.
- Mejoras en el desarrollo de la herramienta CAZy parser (Honorato, 2016) dados algunos problemas identificados en estos scripts.
- Los resultados de los hallazgos de enzimas aumentaron en un 370%, como lo fue para el ejemplo de la familia GH6.
- En los subgrupos que se diagnosticaron visualmente se notó que los *datasets* quedaron divididos por subespecies mejor definidas que las iniciales por el CAZy lo que permite que se hagan estudios más especializados para los investigadores de una especie especifica según sea la necesidad de cada investigación.
- Un nuevo impulso a esta rama para el grupo de investigación GIBI de la Universidad Católica de Manizales.

5.2. Impactos Potenciales de la Tesis

A nivel investigativo en el corto plazo si se sigue mejorando esta versión del proyecto y se logra hacer una depuración y una caracterización más precisa del contenido de cada

subfamilia, se pueden realizar búsquedas que den mayor certeza de enzimas degradadoras de celulosa y hemicelulosa.

Hacer investigación real sobre productos residuales tras la degradación de lignocelulosa puede ser algo que beneficie a muchos tipos de industria, como la farmacéutica, industrias de edulcorantes y las dedicadas a la investigación sobre biocombustibles entre otras.

El conocimiento hecho público del presente trabajo puede hacer de la investigación de enzimas degradadoras de celulosa y hemicelulosa algo más democrático por la inclusión de pequeños grupos de investigación con pocos recursos para hacer uso de una tecnología de fácil acceso para hacer búsqueda de enzimas requeridas, actualmente es algo que está casi que monopolizado por unas cuantas empresas que cobran grandes cantidades de dinero por la venta de enzimas.

Tal vez en un futuro el problema real de los desechos agroindustriales del material lignocelulósico pueda resolverse con la facilitación en la investigación que promueve esta herramienta.

Por lo pronto, si se mejora o se desarrolla una segunda versión del producto final de esta investigación teniendo en cuenta las recomendaciones descritas y haciendo una curación más precisa de cada subgrupo, se podrá hacer una oficialización de un producto que sirva para establecer mejor académicamente al grupo de investigación GIBI de la Universidad Católica de Manizales, dado a que es un desarrollo que puede alcanzar una categoría incluso mayor que las herramientas científicamente populares descritas en los antecedentes de este proyecto.

5.3. Recomendaciones y trabajos futuros

Basados en la discusión planteada en el capítulo anterior de resultados, se sugiere tomar en cuenta los siguientes cambios al momento de realizar una segunda versión del pipeline. Se tratan de algunos puntos que a manera de observación de quien ejecutó el proyecto se pueden tomar en cuenta, la decisión final debe estar consensuada con el grupo de investigación que lo desarrollará. A continuación, se describen los más importantes:

- Utilizar la herramienta CAZy parser (Honorato, 2016) para automatizar la descarga de los identificadores de cada familia del CAZy para evitar errores de falta de datos.
- Descartar familias del CAZy que contengan una cantidad muy pequeña de secuencias.

- Eliminar caracteres especiales de los identificadores de las secuencias para evitar traer secuencias que no correspondan.
- Verificar que las secuencias obtenidas a partir de los identificadores del CAZy, se traten de enzimas propiamente, no fragmentos de enzimas, ni péptidos, si se encuentran deben ser removidas antes de comenzar a hacer los alineamientos.
- Omitir la categoría *Unclassified* del CAZy dado a que contiene secuencias de múltiples especies, no útil para la investigación.
- Verificar cuáles de las proteínas del CAZy son multiespecie, y de serlo preferiblemente quitarlas para que el modelo filtre bien solo familias específicas, o en el mejor de los casos, crear clasificaciones intermedias que consideren estos casos como si fuera una misma familia.
- Descartar secuencias hipotéticas y patentes de la base de datos NR para los alineamientos, o directamente de los hits encontrados.
- Aplicar un filtro de cobertura que haga la operación inversa de calcular el porcentaje de la secuencia objetivo sobre la secuencia query.
- Aplicar un filtro de tamaño de secuencia que no sea fijo como el de 250 que se usó, sino un filtro más dinámico que se valga de los tamaños mínimos de secuencia por cada familia.
- Aplicar el filtro de tamaño no sobre la secuencia query, sino sobre la secuencia objetivo.
- Tras hacer el alineamiento y haber filtrado las secuencias, se debe realizar un segundo alineamiento, pero sobre los dominios glicosil hidrolasas correspondientes a cada familia, esto para validar que la identidad haya sido por tener esta capacidad degradadora y no por otro motivo.

6. Referencias bibliográficas

Agustini, L., Efiyanti, L., Faulina, S. A., & Santoso, E. (2012). Isolation and Characterization of Cellulase-and Xylanase- Producing Microbes Isolated from Tropical Forests in Java and Sumatra. *International Journal of Environment and Bioenergy International Journal of Environment and Bioenergy Journal Int. J. Environ. Bioener*, *3*(33), 154–167. Retrieved from www.ModernScientificPress.com/journals/ijee.aspx

Alvarez, T. M., Goldbeck, R., Santos, C. R. dos, Paixão, D. A. A., Gonçalves, T. A., Franco Cairo, J. P. L., … Squina, F. M. (2013). Development and Biotechnological Application of a Novel Endoxylanase Family GH10 Identified from Sugarcane Soil Metagenome. *PLoS ONE*, *8*(7). http://doi.org/10.1371/journal.pone.0070014

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., … Yeh, L.-S. L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, *32*(Database issue), D115-9. http://doi.org/10.1093/nar/gkh131

Bairoch, A. (1994). The ENZYME data bank. *Nucleic Acids Research*, *22*(17), 3626–3627. http://doi.org/10.1093/nar/22.17.3626

Bastien, G., Arnal, G., Bozonnet, S., Laguerre, S., Ferreira, F., Fauré, R., … O'Donohue, M. (2013). Mining for hemicellulases in the fungus-growing termite Pseudacanthotermes militaris using functional metagenomics. *Biotechnology for Biofuels*, *6*(1), 78. http://doi.org/10.1186/1754-6834-6-78

Baxevanis, A. D., & Bateman, A. (2015). The importance of biological databases in biological discovery. *Current Protocols in Bioinformatics*, *2015*(June), 1.1.1-1.1.8. http://doi.org/10.1002/0471250953.bi0101s50

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. http://doi.org/10.1093/nar/28.1.235

Brink, J. Van Den, & Vries, R. P. De. (2011). Fungal enzyme sets for plant polysaccharide degradation, 1477–1492. http://doi.org/10.1007/s00253-011-3473-2

Busk, P. K., Lange, M., Pilgaard, B., & Lange, L. (2014). Several genes encoding enzymes with the same activity are necessary for aerobic fungal degradation of cellulose in nature. *PLoS ONE*, *9*(12), 1–22. http://doi.org/10.1371/journal.pone.0114138

Busk, P. K., Pilgaard, B., Lezyk, M. J., Meyer, A. S., & Lange, L. (2017). Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC*

Bioinformatics, *18*(1), 214. http://doi.org/10.1186/s12859-017-1625-9

Cantarel, B. I., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Research*, *37*(SUPPL. 1), 233–238. http://doi.org/10.1093/nar/gkn663

Cao, G., & Sheng, Y. (2016). Biobutanol Production from Lignocellulosic Biomass: Prospective and Challenges. *Journal of Bioremediation & Biodegradation*, *7*(4). http://doi.org/10.4172/2155-6199.1000363

Carlos, J., Salgado, S., Meleiro, L. P., Carli, S., & John, R. (2018). Glucose tolerant and glucose stimulated β-glucosidases – A review. *Bioresource Technology*, (July). http://doi.org/10.1016/j.biortech.2018.07.137

Cha, J., Yoon, J., & Cha, C. (2018). Functional characterization of a thermostable endoglucanase belonging to glycoside hydrolase family 45 from Fomitopsis palustris.

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, *2*(4). http://doi.org/10.1371/journal.pone.0000383

Coutinho, T. J. D., Franco, G. R., & Lobo, F. P. (2015). Homology-Independent Metrics for Comparative Genomics. *Computational and Structural Biotechnology Journal*, *13*, 352–357. http://doi.org/10.1016/j.csbj.2015.04.005

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, *14*(9), 755–763. http://doi.org/btb114 [pii]

Eddy, S. R. (2009). a New Generation of Homology Search Tools Based on Probabilistic Inference. *Genome Informatics*, *23*(1), 205–211. http://doi.org/10.1142/9781848165632_0019

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584. http://doi.org/doi: 10.1093/nar/30.7.1575

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279–D285. http://doi.org/10.1093/nar/gkv1344

Geetha, K., & Gunasekaran, P. (2017). Purification of Endoxylanase from Bacillus pumilus B20 for Production of Prebiotic Xylooligosaccharide Syrup ; An In vitro Study, *15*(4). http://doi.org/10.15171/ijb.1494

Gírio, F. M., Fonseca, C., Carvalheiro, F., Duarte, L. C., & Marques, S. (2010). Hemicelluloses for fuel ethanol : A review. *Bioresource Technology*, *101*(13), 4775–4800. http://doi.org/10.1016/j.biortech.2010.01.088

Glaser, A. N., & Nikaido, H. (2007). *Microbial biotechnology. Fundamentals of Applied Microbiology, Second Edition*. *Journal Of Bacteriology*.

Glazer N, Alexander & Nikaido, H. (2007). *Microbial Biotechnology* (Second Edi). New York, USA: CAMBRIDGE UNIVERSITY PRESS.

Gupta, V. K., Carrott, P. J. M., Singh, R., Chaudhary, M., & Kushwaha, S. (2016). Cellulose: A review as natural, modified and activated carbon adsorbent. *Bioresource Technology*. http://doi.org/10.1016/j.biortech.2016.05.106

Henrissat, B., Vegetales, M., & Grenoble, F. (1991). A classification of glycosyl hydrolases based sequence similarities amino acid. *Biochemical Journal*, *280*((Pt 2)), 309–316. http://doi.org/10.1007/s007920050009

Honorato, R. V. (2016). CAZy-parser a way to extract information from the Carbohydrate-Active enZYmes Database. *The Open Journal of Open Source Software*, *42*(2016), 2013. http://doi.org/10.1093/nar/gkt1178.1

Jin, Y., Jayakody, L. N., Liu, J., Yun, E. J., Turner, T. L., Oh, J., … Engineering, B. (2018). Direct conversion of cellulose into ethanol and ethyl- β -, 0–2. http://doi.org/10.1002/bit.26799

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. http://doi.org/10.1093/bioinformatics/btu031

Kubicek, C. P. (2013). Systems biological approaches towards understanding cellulase production by Trichoderma reesei. *Journal of Biotechnology*, *163*(2), 133–142. http://doi.org/10.1016/j.jbiotec.2012.05.020

Kubicek, C. P., Druzhinina, I. S., & Atanasova, L. (2012). *Fungi and Lignocellulosic Biomass Library of Congress Cataloging-in-Publication Data*.

Li, L., Stoeckert, C. J. J., & Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research. *Genome Research*, *13*(9), 2178–2189. http://doi.org/10.1101/gr.1224503.candidates

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. http://doi.org/10.1093/bioinformatics/btl158

Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013, *42*(November 2013), 490–495. http://doi.org/10.1093/nar/gkt1178

Manavalan, T., Manavalan, A., Thangavelu, K. P., & Heese, K. (2015). Characterization of a novel endoglucanase from Ganoderma lucidum, 1–11. http://doi.org/10.1002/jobm.201400808

Martínez, Á. T., Speranza, M., Ruiz-Dueñas, F. J., Ferreira, P., Camarero, S., Guillén, F., … Del Río, J. C. (2005). Biodegradation of lignocellulosics: Microbial, chemical, and enzymatic aspects of the fungal attack of lignin. *International Microbiology*, *8*(3), 195–204. http://doi.org/im2305029 [pii]

Meinke, A. (1995). Enhancement of the endo-beta-1,4-glucanase activity of an exocellobiohydrolase by deletion of a surface loop. *The Journal of Biological Chemistry*.

Murphy, C., Powlowski, J., Wu, M., Butler, G., & Tsang, A. (2011). Curation of characterized glycoside hydrolases of Fungal origin, *2011*, 1–14. http://doi.org/10.1093/database/bar020

Park, B. H., Karpinets, T. V., Syed, M. H., Leuze, M. R., & Uberbacher, E. C. (2010). CAZymes Analysis Toolkit (cat): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology*, *20*(12), 1574–1584. http://doi.org/10.1093/glycob/cwq106

Pauly, M., Gille, S., & Liu, L. (2013). Hemicellulose biosynthesis. http://doi.org/10.1007/s00425-013-1921-1

Pearson, W. R. (2014). An Introduction to Sequence Similarity (“ Homology ”) Searching, 1–9. http://doi.org/10.1002/0471250953.bi0301s42.An

Peng, F., Ren, J. L., Xu, F., & Sun, R. (2011). Chemicals from Hemicelluloses : A Review.

Rawat, R., Kumar, S., & Singh, B. (2015). An acidothermophilic functionally active novel GH12 family endoglucanase from Aspergillus niger HO : purification , characterization and molecular interaction studies, 103–117. http://doi.org/10.1007/s10482-014-0308-z

Rossi, M. F., Mello, B., & Schrago, C. G. (2017). Performance of Hidden Markov Models in Recovering the Standard Classification of Glycoside Hydrolases. *Evolutionary Bioinformatics Online*, *13*, 1176934317703401. http://doi.org/10.1177/1176934317703401

Shallom, D., & Shoham, Y. (2003). Microbial hemicellulases. *Current Opinion in Microbiology*, *6*(3), 219–228. http://doi.org/10.1016/S1369-5274(03)00056-0

Sievers, Fabian;Wilm, Andreas;Dineen, David;Gibson, Toby J;Karplus, Kevin;Li, Weizhong;Lopez, Rodrigo;McWilliam, Hamish;Remmert, Michael;Soding, Johannes;Thompson, Julie D;Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biolog*.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., … Vaughan, R. (2002). The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, *30*(1), 21–6. http://doi.org/10.1093/nar/gki098

Strasser, K., Mcdonnell, E., Nyaga, C., Wu, M., Wu, S., Almeida, H., … Tsang, A. (2015). mycoCLAP , the database for characterized lignocellulose-active proteins of fungal origin : resource and text mining curation support, 1–10. http://doi.org/10.1093/database/bav008

Ummartyotin, S., & Manuspiya, H. (2015). A critical review on cellulose : From fundamental to an approach on sensor technology. *Renewable and Sustainable Energy Reviews*, *41*, 402–412. http://doi.org/10.1016/j.rser.2014.08.050

Yan, R., Vuong, T. V, Wang, W., & Master, E. R. (2017). Action of a GH115 α -glucuronidase from Amphibacillus xylanus at alkaline condition promotes release of 4- O -methylglucopyranosyluronic acid from glucuronoxylan and arabinoglucuronoxylan. *Enzyme and Microbial Technology*, *104*(May), 22–28. http://doi.org/10.1016/j.enzmictec.2017.05.004

Yao, L., Yoo, C. G., Meng, X., Li, M., Pu, Y., & Ragauskas, A. J. (2018). Biotechnology for Biofuels A structured understanding of cellobiohydrolase I binding to poplar lignin fractions after dilute acid pretreatment, 1–11. http://doi.org/10.1186/s13068-018-1087-y

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., & Xu, Y. (2012). DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, *40*(W1), 445–451. http://doi.org/10.1093/nar/gks479

Yoon, B. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis, 402–415.

7. Anexos

Anexo I. Familia Glicosil Hidrolasas del CAZy.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo II. Archivos iniciales CAZy

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo III. Script de Python para filtrado de secuencias de alineamientos pareados en formato tabla.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co

Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo IV. Pasos siguientes de OrthoMCL.

orthomclBlastParser salida.tab complaint >> similarSequences.txt

Se crea la base de datos orthocml vacia y se le otorgan los privilegios sobre la base de datos.

CREATE DATABASE orthomcl;
GRANT ALL PRIVILEGES ON orthomcl. TO orthouser@localhost;*
orthomclInstallSchema orthomcl.conf

Antes de hacer la carga de los datos se debe crear un indice unico para la tabla SimilarSequences, no es necesario siempre en las ejecuciones de orthomcl, solo en esta.
Es necesario dado a que al hacer el paso anterior de todos contra todos se efectuo el blast de la base de datos simplificada contra la base de datos en archivos individuales lo que causo que existieran hits repetidos al no simplificar.
CREATE UNIQUE INDEX unique_similarSequences ON SimilarSequences (`QUERY_ID`, `SUBJECT_ID`, `QUERY_TAXON_ID`, `SUBJECT_TAXON_ID`);

Se carga cargan los datos de las secuencias
orthomclLoadBlast orthomcl.conf similarSequences.txt

Se efectua el agrupamiento de las secuencias con orthomcl pairs.
orthomclPairs orthomcl.conf pairs.log cleanup=no

Se descargan los pares
orthomclDumpPairsFiles orthomcl.conf

Se corrio mcl para los valores de 1 y 1.5
mcl mclInput --abc -I 1.5 -o mclOutput
Se pasa el formato MCL a groups de orthoMCL

orthomclMclToGroups PDB 0 < mclOutput_1.5 > groups_1.5.txt

Anexo V. Estadística de las familias iniciales.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo VI. Listado de casos anormales de las familias iniciales del CAZy.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo VII. Histogramas de longitudes de las secuencias iniciales.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo VIII. Estadística de los alineamientos.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo IX. Estadística de las secuencias filtradas.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo X. Secuencias pertenecientes a múltiples familias.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo XI. Resultados y estadística de la ejecución del MCL.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo XII. Asociación de las secuencia a los identificadores por cluster.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo XIII. Estadística de las secuencias finales.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo XIV. Estadística de los clusters del OrthoMCL

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co
Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.

Anexo XV. Script para busqueda de enzimas degradadoras de celulosa y hemicelulosa.

Solicitar enlace de acceso a los documentos a la coordinación de la Maestría en Bioinformática y Biología Computacional escribiendo a los correos:

Gloria María Restrepo - grestrepo@ucm.edu.co

Narmer Fernando Galeano - ngaleano@ucm.edu.co

Debe hacer referencia al anexo que requiere.