CBMDB: A Database for Accessing, Analyzing, and Mining CBM Information

Xu Lin, Xiaqi Xie, Xueyan Wang, Zhimin Yu, Xiaoyi Chen * and Fan Yang *

School of Biological Engineering, Dalian Polytechnic University, Ganjingziqu, Dalian 116034, China
* Correspondence: chen-xy@dlpu.edu.cn (X.C.); yang_fan@dlpu.edu.cn (F.Y.)

Abstract: Carbohydrate-binding modules (CBMs) are important substrate-binding domains that are mainly contained within carbohydrate-active enzymes. To elucidate the mechanism of enzyme-carbohydrate recognition and to promote the process of enzymatic engineering, it is important to explore more potential CBMs. However, the information and analytic tools of CBMs provided by current databases are limited. Here, a simple, user-friendly, and comprehensive CBM database (CBMDB) that integrates multidimensional information and analysis tools was constructed. Based on a data query function and analysis tools provided by the CBMDB, including sequence similarity searches, pairwise alignment, multiple sequence alignment, structure similarity searches, and phylogenetic visualization, information retrieval and analysis of known CBMs could be easily performed. Notably, unknown proteins with potential CBM functions could also be examined based on existing CBM data.

Keywords: online database; web interface; carbohydrate-binding module; multidimensional information

1. Introduction

Carbohydrate-binding modules (CBMs) are a sort of protein that can bind to carbohydrates and usually constitute large multi-module enzymes as a relatively independent module [1]. CBMs have been found to promote the association between an enzyme and a substrate, increasing the effective enzyme concentration on the polysaccharide surface, thus enhancing its enzymatic activity [2–5]. This proximity effect has been demonstrated by several studies, where it is noticed that enzymes fail to hydrolyze insoluble substrates when CBMs are genetically removed [5]. In addition, CBMs have been proposed to be highly specific, to be able to recognize different polysaccharides, and to target enzyme binding to specific sites of the substrate [6,7]. This substrate-targeting effect has been used as a tool for the elucidation of protein–carbohydrate interaction mechanisms [8]. The particular aromatic residues on the surface of CBMs, especially tryptophan and highly conserved tyrosine residues, play an important role in the binding of CBMs to ligands. The appropriate topology reflecting the structure of the substrate and the orientation of conserved residues located in the binding site of CBMs guide the effective binding and recognition of CBMs to the substrates [9]. Moreover, some studies have shown that the binding of CBMs to a crystalline substrate leads to the disruption of polysaccharide structures and, therefore, enhances substrate availability for enzymatic hydrolysis [10,11]. With the development of enzyme engineering, it is important to explore more potential CBMs to elucidate the mechanism of protein–carbohydrate recognition and to promote its application in enzyme molecular modification and drug development.

To date, an increasing number of CBMs have been identified and characterized. The total number of CBM entries is 277,412, which come from 88 families, and at least one 3D structure has been reported for 408 modules [9]. All of this information can be found in some comprehensive databases, such as NCBI (https://www.ncbi.nlm.nih.gov/), PFAM (http://pfam.xfam.org/), PDB (https://www.rcsb.org/), CAZy (http://www.cazy.org/), and PubMed (https://pubmed.ncbi.nlm.nih.gov/). NCBI’s Protein Data Bank mainly
records the amino acid sequences of CBMs contained in relative enzymes. PFAM provides information on protein families. The 3D structure of CBMs, together with relative enzymes, can be found in PDB records. CAZy provides the GenBank accession number of CBMs. Related literature about CBMs can be searched for on PubMed. With the rapid development of high-throughput sequencing technology, a large amount of potential CBMs need to be identified, which presents a challenge for the collection, integration, and analysis of huge amounts of information. However, none of the current databases provide independent information about CBMs, and the information on CBMs is not consolidated. Furthermore, some analytical tools for CBM research, such as BLAST, Clustal Omega, structure searches, and viewing tools, are limited in these databases. All of these problems make CBM work time-consuming.

To facilitate the search for and analysis of CBM information, a simple, user-friendly, and comprehensive CBM database (CBMDB) that integrates multidimensional information and analytical tools was constructed. All of the CBM data were collected and downloaded from various public sources, such as NCBI, PFAM, PDB, CAZy, and PubMed, based on indexes and API fetches. Some classic sequence and structure analysis tools, such as BLAST, Clustal Omega, and Jmol, were integrated into CBMDB. Our CBMDB platform should help researchers to access CBM information more efficiently and provide insights into exploring potential CBMs and speculating on their functions.

2. Materials and Methods

2.1. Database Architecture and Web Interface

To collect, organize, and save information about carbohydrate-binding modules (CBMs) and to develop a useful web interface, a CBM database (CBMDB) was developed using conventional web development techniques. The gathered data were integrated into a MySQL relational database (version 5.6.50). PhpMyAdmin (version 4.4) was used as the database administration tool. The interactive interface of the website was built using HTML5, CSS, and JavaScript. Bootstrap was used for layout design. PHP (version 7.1) and Bash (version 4.2.46(1)) scripts were used for server-side development. Apache was used as the web server. CBMDB was deployed in an Alibaba Cloud host running CentOS7 (version 7.2.1511). The database website is freely available at http://cbmdb.org.cn/.

2.2. Data Acquisition and Compilation

Comprehensive and reliable data for the CBMDB were collected from various resources including NCBI (https://www.ncbi.nlm.nih.gov/), PDB (https://www.rcsb.org/), CAZy (http://www.cazy.org/), and PubMed (https://pubmed.ncbi.nlm.nih.gov/). By searching CAZy with the CBM family name, GenBank accession numbers of carbohydrate-active enzymes were returned. Edirect [12], a searching and batch download tool provided by NCBI, was used to download the amino acid sequences of enzymes containing CBMs using the GenBank accession numbers. Article information on CBM and its related carbohydrate-active enzymes, including title and index number, was downloaded from the PubMed Library (https://pubmed.ncbi.nlm.nih.gov/). By searching CAZy with the CBM family name, GenBank accession numbers of carbohydrate-active enzymes were returned. Edirect [12], a searching and batch download tool provided by NCBI, was used to download the amino acid sequences of enzymes containing CBMs using the GenBank accession numbers. Article information on CBM and its related carbohydrate-active enzymes, including title and index number, was downloaded from the PubMed Library (https://pubmed.ncbi.nlm.nih.gov/). Shell scripts and Python scripts were used to extract the enzyme type and organism information from the downloaded data. Annotation information for CAZy enzymes and CBMs, and their amino acid sequences obtained from other databases were downloaded in FASTA format. In a file with FASTA formatting, the sequence annotation and sequence were split into two lines, where the sequence annotation line had a unique symbol as a handle. Based on this feature, Python scripts were used to split these data into annotations and sequences using handles, where the annotations were then separated by regular expressions and separators. These split data were then stored in a database in the form of a table.

HMMER [13] (http://hmmer.org/, version 3.3.2) was used for analyzing the primary structure of carbohydrate-active enzymes and for obtaining the position information of CBMs. HMMER’s hmsmscan program was applied for annotating the amino acid sequence of enzymes against the PFAM database. The domain hits table was selected as the output.
format of domain annotations. The domain table had 22 whitespace-delimited fields followed by a free text target sequence description. A Python script was used to derive the target name, the position information of primary structures of CBMs, and the E-value and to filter the information. There may be more than one domain per sequence in the domain hits table, so we filtered the annotation results according to the E-value and the target name. The annotation results of the query sequences that were annotated as ‘CBM’ in the ‘target names’ column and with an E-value of less than 0.1 were selected as the basis for extracting CBM sequences. For the extraction of CBM sequences, Python scripts were used based on the location information of the CBM provided by the analysis results of the HMMER package. The specific extraction process was to extract the characters in the string according to the position index, where the amino acid sequence was equivalent to the string and the position of the CBM provided the position index. The extracted CBM sequences were then combined with information on their source enzymes and stored in the database in tabular form. Throughout the data-processing process, Shell scripts were used to automate the running of programs and to batch process large amounts of data.

2.3. Data Organization

To classify, organize, and save the data, the CBMDB was organized into two tables. The first table contained the following fields:
- Enzyme ID: A unique identifier to tag this table entry. Each identifier began with the character ‘>’ followed by letters and digits. That identifier was consistent with the GenBank accession number;
- Enzyme Type: The types of carbohydrate-active enzymes;
- Species Origin: The scientific name of the organism with the carbohydrate-active enzymes;
- Enzyme Seq: The amino acid sequence of the carbohydrate-active enzymes;
- CBM Family: The classification of the CBM.

The second table contained the following fields:
- CBM Enzyme ID: A unique identifier to tag this table entry. It consisted of the Enzyme ID in the first table and a numeric suffix connected by ‘#’;
- CBM Family: The classification of the CBM. It was extracted from the annotation results of hmmscan;
- CBM Position: The position of the CBM in the primary structure of carbohydrate-active enzymes;
- CBM Seq: The amino acid sequence of the CBM. It was obtained from the total amino acid sequence of enzymes by position information in the annotation results.

2.4. Integration of Web Tools

For the analysis of CBM sequences, various tools were set up in the CBMDB. Users can click “upload” on the home page to use these tools. Brief descriptions of these tools are listed below.

The information search tool on the home page was implemented by connecting to the database through PHP and using MySQL query statements. Wildcards and regular expressions added to the MySQL statement make this search tool capable of fuzzy queries.

BLAST [14]: Users can align their sequences with CBM sequences in our database and perform pairwise alignment. Users are permitted to input query sequences or to submit files in FASTA format. The output is shown in the normal BLAST output. Users can use the above two functions in the “BLAST” and “Global Align” sections.

Clustal Omega [15]: Clustal Omega is a popular kit for performing multiple sequence alignment. It is swift enough to make very large alignments, and the precision of the protein alignments is high when compared to alternative packages. This tool can be found in the “Multiple sequence alignment” section.

HMMER [13]: HMMER is a package based on profile hidden Markov models, which are computational algorithms for predicting probabilistic protein models and functions. Re-
lying on the strength of probability models, HMMER searches for homologous sequences in sequence databases and performs either single sequences or multiple sequence alignments.

Phylogeny tree [16]: This tool provides a view of the phylogenetic tree and appearance modifications including layout, hint labels, node labels, node shapes, node bars, branch labels, and scale bars.

Application program interface (API) [17]: RCSB PDB Data API supports the HTTP GET method in accessing PDB data through a set of endpoints (or URLs). The API can quickly search the PDB number of proteins on the basis of the query sequence to view and analyze the three-dimensional structure of proteins. This function and sequence similarity alignment was integrated into the web page using BLAST for alignment.

Jmol [18]: Jmol was integrated into the website’s tools so that users can directly view interactive images of the protein structures on the web page. Jmol (http://www.jmol.org/) is an open-source software for visualizing chemical structures in 3D. The Jmol applet can be inserted into our web page to generate three-dimensional protein structure visualizations without local 3D model files by using the following website: https://chemapps.stolaf.edu/jmol/jmol.php (accessed on 1 August 2022). The 3D models can be obtained either from the RCSB using a PDB ID or from the NIH CACTVS server using a chemical identifier.

3. Results and Discussion

3.1. Data Statistics

Comprehensive information was generated, accumulated, and stored in the CBMDB by entry. This database combined the amino acid sequences of CBMs and CBM-containing enzymes, CBMs’ families, relevant literature, original species, and enzyme types. The amino acid sequences of 533,496 carbohydrate-active enzymes were downloaded from NCBI. In total, 143,015 sequences were obtained after deleting redundancies. There were 143,316 CBM sequences that came from different enzymes or different positions of the same enzyme; these were extracted by HMMER annotation. Currently, the CBMDB has collected and produced more than 280,000 entries covering 88 known CBM families. Data stored in the CBMDB were categorized into two parts: CBM-containing enzymes and CBMs. These data are available to users through the CBMDB’s web interface.

3.2. Web Interface: Data Query and Analysis Tool

The CBMDB has a user-friendly web interface, which supports users and allows them to smoothly query various CBM information, to download data, and to analyze and visualize information using the online tools. The web interface of the CBMDB is composed primarily of two parts: data query tools and analysis tools (Figure 1). In order to teach users to learn how to use our website quickly, the interface of each section was set up in a simple and straightforward way. The scripting language PHP was useful for implementing and managing connections between the front-end and back-end of the website, for performing data queries, and for using analytics tools swiftly and facilely. This service is meant to dramatically decrease the amount of time researcher spend searching for CBM information and configuring analysis tools on various websites.
3.3. Data Query

All the data stored in the database can be queried directly via a keyword search box on the home page of the CBMDB web browser (Figure 2a). Users can query the information about CBMs automatically by using the keyword search function, using information such as a short amino acid sequence of CBMs, the GenBank accession number of carbohydrate-active enzymes, the type or scientific name of enzymes, or the CBM family. Every keyword produces the results page containing integrated outputs, including the amino acid sequence of CBMs, the GenBank accession number, the PDB ID, the enzyme category, the organism, the CBM family category, and the literature. In addition, users can view the 3D model of a protein with the given PDB ID on the web page, and the PDB file corresponding to the PDB ID can be freely downloaded. Users can also query the database by using fuzzy or incomplete keywords or sequences. For example, if users enter a part of an amino acid sequence, information on all enzymes and CBMs containing this sequence is returned.

![Figure 1. Architecture of the CBMDB website.](image)

**Figure 1.** Architecture of the CBMDB website.

**Figure 2.** Cont.
3.4. Analysis Tools

The CBMDB integrates five web-based tools for performing further analyses, including sequence similarity searches, pairwise alignment, multiple sequence alignment, structure similarity searches, domain annotation, and phylogenetic visualization (Figure 2a).

Sequence similarity searches and pairwise alignment are performed by the “blastp” or “blastx” search against sequences deposited in the CBMDB and searched for sequences in the database that have a high similarity with the query sequence; these are then extracted (Figure 2b). Sequence similarity searches and pairwise alignments are implemented using “blastp” and “blastx”, two search applications in the sequence alignment software BLAST+ [14]. “Blastp” is a general amino acid sequence identification and similarity search application. It is suitable for aligning protein sequences in protein databases. “Blastx” is an application for identifying potential protein products encoded by a nucleotide query. It is suitable for aligning nucleic acid sequences in protein databases; therefore, when the query sequence entered by the user is a nucleic acid sequence, the user should select the blastx option for alignment. Both of these applications are based on the Needleman–Wunsch algorithm [19], which is an algorithm used in bioinformatics to align protein or nucleotide sequences. The algorithm assigns a score to every possible alignment and finds all possible alignments that have the highest score.

In the Structure Similarity Search module, these sequences are applied as a parameter of the PDB application programming interface (API) for searching the structural information...
of the protein with the highest sequence similarity to the query sequence (Figure 2b). PDB API is a web application programming interface [17]. A client device can send a request as a Hypertext Transfer Protocol (HTTP) request via web application programming interfaces and receives a response message, usually in JavaScript object notation (JSON) or extensible markup language (XML) format. The PDB API can provide access to information about PDB structures.

The multiple sequence alignment tool is used to align three or more sequences together in a computationally efficient and accurate manner. Sequences in FASTA format can be entered directly into the box on the web page. Additionally, files containing valid sequences in FASTA format can be uploaded and used as the input for multiple sequence alignments (Figure 2b). Multiple sequence alignments are achieved with the Clustal Omega package, which can align virtually any number of protein sequences quickly and accurately. Clustal Omega uses a modified version of mBed [20] to produce guide trees. The vectors generated by the mBed method are then clustered extremely fast (compared to standard methods such as K-Means or UPGMA), and then, Clustal Omega computes the comparison using the very accurate HHalign package [21].

The Domain Annotation module, which is based on the algorithmic principles of hidden Markov models, is applied to annotate and analyze domains in proteins. To annotate CBM and other domains in the query sequence, the module uses the hmmscan program in the HMMER package to compare query sequences with the CBMDB and PFAM databases. The input data required are amino acid sequences in FASTA format. The type and scoring of the comparison results for these domains are then obtained (Figure 2b). Hidden Markov models (HMMs) are Markov chains with implicitly unknown parameters [22]. HMMs consist of a finite number of states, each of which appears with a certain probability, and the transitions between states are determined by a table of transition probabilities, each of which can produce an observed state. In hidden Markov models, only the observed states are visible in the hidden Markov model, while the real Markov state chain is not visible, hence the name “hidden Markov model”.

The phylogenetic visualization tool can display interactive phylogenetic trees and produce beautifully rendered, sometimes publication-ready figures. The phylogenetic tree files in Newick format can be uploaded and rendered. It also allows for automated annotation of extended features, such as confidence intervals to nodes as well as the size, color, and appearance using extended data in the tree file. To avoid slowing down the browser when visualizing, the size of the phylogenetic tree should be limited to less than 4000 nodes.

Using the data query and analysis tools described above, the CBMDB provides information that can be used by researchers in their research practices.

### 3.5. Application

To illustrate the application of CBMDB, a known CBM (GenBank accession number: ABN53395.1) was selected for analysis. Firstly, the accession number was used as a search term to collect basic information about the CBM. The search results page showed that the target CBM originates from the strain *Hungateiclostridium thermocellum* and belongs to the CBM6 family. Amino acid sequences and the position information of the target CBM and the enzyme containing this CBM were also available (Figure 3a). Further, using the Sequence Similarity Search module, a sequence BLAST between the query CBM and subject CBMs was performed. Based on the sequence BLAST results, multiple alignment was then conducted. As shown in Figure 3b,c, ABN53395.1 showed the highest identity scores with three characterized CBMs, including CBM6 from positions 399 to 515 (accession number EFB38194.1, 100.0% identity), CBM6 from positions 393 to 510 (accession number EFL62381.1, 77.8% identity), and CBM6 from positions 2108 to 2221 (accession number QDV75083.1, 60.2% identity) (Figure 3b). The sequence alignment of ABN53395.1 with EFB38194.1, EFL62381.1, and QDV75083.1 revealed two conserved substrate binding residues (Tyr22, Trp84) [23]. The CBM was then analyzed using the Domain Annotation
module. As shown in Figure 3d, the highest score, c-Evalue, and i-Evalue are $5.3 \times 10^{-24}$, $1.1 \times 10^{-20}$, and 74.2 respectively, indicating that the query sequence belongs to CBM6 and that the positions of CBM6 in the query sequence are 402–514. The phylogenetic relationship between the CBM ABN53395.1 and the known CBMs in the database was then evaluated using the Phylogenetic Tree module. The results suggested that ABN53395.1 was clearly clustered into the clade of CBM family 6 in the phylogenetic tree based on the neighbor-joining method of the amino sequences of CBMs, in which the level of sequence similarity between ABN53395.1 and EFL62831.1 was 77.8% (Figure 3e). Finally, the structure information of ABN53395.1 was obtained by the Structure Similarity Search module through the PDB API. The result revealed a typical beta sandwich structure of the CBM6 family [2]. Additionally, the 3D models of the CBMs, such as 2Y8K, 5G56 and 1UY0, which showed high sequence similarity to the query sequence, were found on this results page (Figure 3f).

(a)

(b)

Figure 3. Cont.
Figure 3. Cont.
Another important application of the CBMDB is the examination of unknown proteins with potential CBM functions. Here, the C-Terminus module of an endotype xanthanase (ALX66163.1) from the genus Microbacterium was analyzed as an example. Sequence alignment was run with the Sequence Similarity Search module [24]. The subject “AMM07675.1#8969, CBM6, Position: 917-1041” showed the highest amino sequence similarity to the query sequence (30.1% identity) (Figure 4a), suggesting that the unknown C-terminus of the xanthanase might function as a CBM and belong to the CBM6 family. The multiple sequence alignment results showed a low sequence identity of the unknown C-terminus to CBM6 (AMM07675.1, 30.1% identity), CBM4_9 (ADQ44988.1, 26.5% identity), and CBM6 (AYY13113.1, 25.3% identity). However, several conserved substrate binding residues (Tyr and Trp) were still observed (Figure 4b). After structure similarity searches, the 3D model of CBM6 (1W9W) with a 34.9% identity to query sequence was obtained and displayed a typical beta sandwich structure of the CBM6 family [2] (Figure 4c). The phylogenetic results demonstrated that the query protein was clustered into the clade of CBM family 6 (Figure 4d). All of these analyses implied that the unknown C-terminus of the xanthanase probably represented a novel branch of the CBM6 family.
Figure 4. Cont.
3.6. Comparison to Other Databases

To further demonstrate the contribution of our database in efficiently accessing CBM information, in exploring potential CBMs, and in speculating on their functions, the CBMDB was compared with two other commonly used CBM-related databases: CAZy (http://www.cazy.org/) and dbCAN (https://bcb.unl.edu/dbCAN2/).

CAZy has been used for nearly three decades and has become the main access point for researchers to collect and classify information on CAZymes, some of which contain CBM domains [25]. The dbCAN, another popular database, integrates with the HMMER, DIAMOND, and Hotpep search tools and provides automated CAZyme annotations for newly sequenced genomes [26]. Both databases mainly focus on providing information on CAZymes and are not specifically directed at CBMs. Furthermore, only limited information on CBMs, including classification, strain origin, and GenBank accession number are available. Compared to CAZy and dbCAN, the CBMDB integrates more direct data related to CBMs, including the amino acid sequence, the structure, source enzymes, the position of CBMs in the enzyme, etc. In addition, unlike the CAZy and dbCAN, the CBMDB combines more powerful analysis tools that enable multidimensional information access, such as the sequence alignment tools for classification and conserved site confirmation, the PDB-based search tool for structural information, the visual analysis tool for phylogenetic trees, and the HMMs-based tools for CBMs annotation. Consequently, the CBMDB will be a unique and powerful resource for CBM-related research.
4. Conclusions

In this work, we presented the CBM database and its web interface, which can be accessed online (http://cbmdb.org.cn/); collects multidimensional information from various public sources; and integrates five powerful web-based analysis tools for performing further analyses, including sequence similarity searches, pairwise alignment, multiple sequence alignment, structure similarity searches, and phylogenetic visualization. The user-friendly web interface supports users and allows them to smoothly query various CBM information, to download data, and to analyze and visualize information using online tools. Notably, unknown proteins with potential CBM functions could also be predicted based on existing CBM data. Our CBM database is valuable for researchers who wish to access CBM information more efficiently and will provide insights into exploring potential CBMs and speculating on their functions. In the forthcoming future, the database will be expanded, the data structure will be reorganized based on new CBM-related data, and analytical tools will be further optimized.

Author Contributions: X.L. and X.X. contributed equally to this paper. Conceptualization, X.L., XX., X.W. and Z.Y.; methodology, X.L. and X.X.; software, X.L. and X.X.; validation, X.L. and XX.; investigation, X.L. and X.X.; resources, X.L. and XX.; data curation, X.L., XX., X.W. and Z.Y.; writing—original draft preparation, X.L.; writing—review and editing, X.L. and F.Y.; visualization, F.Y., XL., XX., X.W. and Z.Y.; supervision, X.C. and F.Y.; project administration, X.C. and F.Y.; funding acquisition, X.C. and F.Y. All authors have read and agreed to the published version of the manuscript.

Funding: Financial support provided by the National Natural Sciences Foundation of China (32072160, 31671796, 31801469), the Natural Science Foundation of Liaoning Province (J2020041, 2020-MS-276), and the Liaoning BaiQianWan Talents Program is also greatly acknowledged.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sidar, A.; Albuquerque, E.D.; Voshol, G.P.; Ram, A.E.J.; Vijgenboom, E.; Punt, P.J. Carbohydrate Binding Modules: Diversity of Domain Architecture in Amylases and Cellulases From Filamentous Microorganisms. Front. Bioeng. Biotechnol. 2020, 8, 871. [CrossRef] [PubMed]
2. Boraston, A.B.; Bolam, D.N.; Gilbert, H.J.; Davies, G.J. Carbohydrate-Binding Modules: Fine-Tuning Polysaccharide Recognition. Biochem. J. 2004, 382, 769–781. [CrossRef] [PubMed]
3. Guillén, D.; Sánchez, S.; Rodriguez-Sanoja, R. Carbohydrate-Binding Domains: Multiplicity of Biological Roles. Appl. Microbiol. Biotechnol. 2010, 85, 1241–1249. [CrossRef] [PubMed]
4. Levy, I.; Shoseyov, O. Cellulose-Binding Domains: Biotechnological Applications. Biotechnol. Adv. 2002, 20, 191–213. [CrossRef]
5. Shoseyov, O.; Shani, Z.; Levy, I. Carbohydrate Binding Modules: Biochemical Properties and Novel Applications. Microbiol. Mol. Biol. Rev. 2006, 70, 283–295. [CrossRef]
6. Linder, M.; Teeri, T.T. The Roles and Function of Cellulose-Binding Domains. J. Biotechnol. 1997, 57, 15–28. [CrossRef]
7. Wang, J.; Stuckey, J.A.; Wishart, M.J.; Dixon, J.E. A Unique Carbohydrate Binding Domain Targets the Lafora Disease Phosphatase to Glycogen. J. Biol. Chem. 2002, 277, 2377–2380. [CrossRef]
8. McCartney, L.; Gilbert, H.; Bolam, D.; Boraston, A.; Knox, P. Glycoside Hydrolase Carbohydrate-Binding Modules as Molecular Probes for the Analysis of Plant Cell Wall Polymers. Anal. Biochem. 2004, 326, 49–54. [CrossRef]
9. Barral, P.; Suárez, C.; Batanero, E.; Alfonso, C.; Alché, J.D.D.; Rodriguez-García, M.I.; Villalba, M.; Rivas, G.; Rodriguez, R. An Olive Pollen Protein with Allergenic Activity, Ole e 10, Defines a Novel Family of Carbohydrate-Binding Modules and Is Potentially Implicated in Pollen Germination. Biochem. J. 2005, 390, 77–84. [CrossRef]
10. Gao, P.-J.; Chen, G.-J.; Wang, T.-H.; Zhang, Y.-S.; Liu, J. Non-Hydrolytic Disruption of Crystalline Structure of Cellulose by Cellulose Binding Domain and Linker Sequence of Celllobiohydrolase I from Penicillium Janthinellum. Sheng Wu Wu Li Xue Bao Acta Biochim. Et Biophys. Sin. 2001, 33, 13–18.
11. Wang, L.; Zhang, Y.; Gao, P. A Novel Function for the Cellulose Binding Module of Celllobiohydrolase I. Sci. China C Life Sci. 2008, 51, 620–629. [CrossRef]
12. Kan, J. Entrez Direct: E-Utilities on the Unix Command Line. In Entrez Program. Util. Help. [Internet]; National Center for Biotechnology Information: Bethesda, MD, USA, 2013.

13. Hmmer.org. Available online: http://hmmer.org/ (accessed on 1 June 2022).

14. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and Applications. BMC Bioinform. 2009, 10, 421. [CrossRef]

15. Sievers, F.; Higgins, D.G. Clustal Omega for Making Accurate Alignments of Many Protein Sequences: Clustal Omega for Many Protein Sequences. Protein Sci. 2018, 27, 135–145. [CrossRef]

16. Jovanovic, N.; Mikheyev, A.S. Interactive Web-Based Visualization and Sharing of Phylogenetic Trees Using Phylogeny.IO. Nucleic Acids Res. 2019, 47, W266–W269. [CrossRef]

17. Dotsika, F. Semantic APIs: Scaling up towards the Semantic Web. Int. J. Inf. Manag. 2010, 30, 335–342. [CrossRef]

18. Hanson, R.M. Jmol–A Paradigm Shift in Crystallographic Visualization. J. Appl. Crystallogr. 2010, 43, 1250–1260. [CrossRef]

19. Needleman, S.B.; Wunsch, C.D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. J. Mol. Biol. 1970, 48, 443–453. [CrossRef]

20. Blackshields, G.; Sievers, F.; Shi, W.; Wilm, A.; Higgins, D.G. Sequence Embedding for Fast Construction of Guide Trees for Multiple Sequence Alignment. Algorithms Mol. Biol. 2010, 5, 21. [CrossRef]

21. Sievers, F. Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega. Mol. Syst. Biol. 2011, 7, 539. [CrossRef]

22. Eddy, S.R. What Is a Hidden Markov Model? Nat. Biotechnol. 2004, 22, 1315–1316. [CrossRef]

23. Abbott, D.W.; van Buuren, A.L. Using Structure to Inform Carbohydrate Binding Module Function. Curr. Opin. Struct. Biol. 2014, 28, 32–40. [CrossRef]

24. Yang, F.; Li, H.; Sun, J.; Guo, X.; Zhang, X.; Tao, M.; Chen, X.; Li, X. Novel Endotype Xanthanase from Xanthan-Degrading Microbacterium Sp. Strain XT11. Appl. Environ. Microbiol. 2019, 85, e01800–e01818. [CrossRef]

25. Cantarel, B.L.; Coutinho, P.M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B. The Carbohydrate-Active EnZymes Database (CAZy): An Expert Resource for Glycogenomics. Nucleic Acids Res. 2009, 37, D233–D238. [CrossRef]

26. Zhang, H.; Yohe, T.; Huang, L.; Entwistle, S.; Wu, P.; Yang, Z.; Busk, P.K.; Xu, Y.; Yin, Y. DbCAN2: A Meta Server for Automated Carbohydrate-Active Enzyme Annotation. Nucleic Acids Res. 2018, 46, W95–W101. [CrossRef]