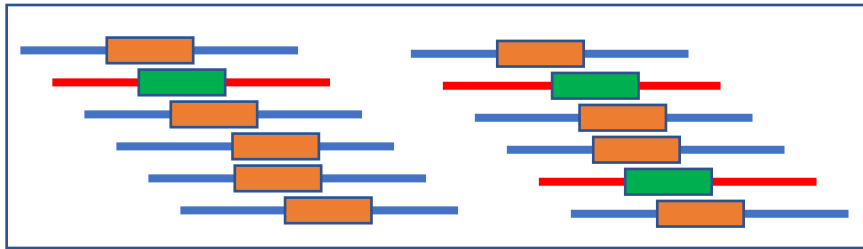


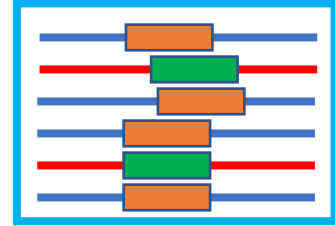
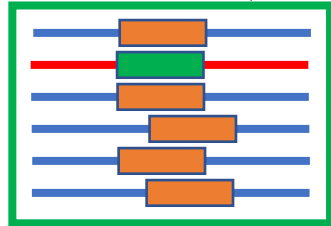
Proteins of CAZy family (e.g., GH1:  
<http://www.cazy.org/GH1.htm> )

**workflow of dbCAN-sub HMM construction**

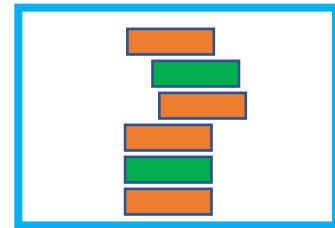
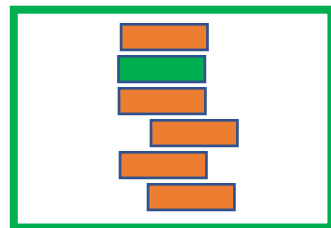


proteins of a CAZy family (e.g., GH1)

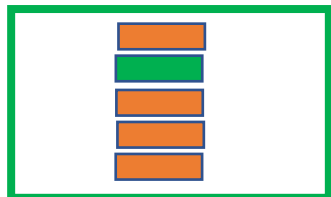
**GH1\_e0**  **eCAMI** *Bioinformatics 2020*  **GH1\_e1**



proteins of eCAMI subfamilies (e.g., GH1\_e0)



dbCAN GH1 domains of proteins of GH1\_e0 (hmmscan)

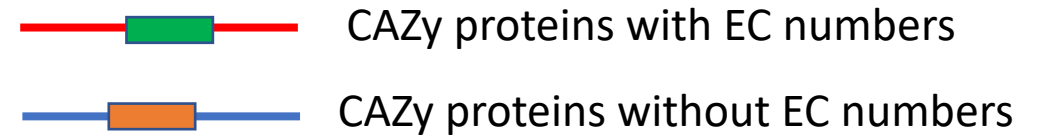


dereplicated dbCAN GH1 domains (cd-hit 95% sequence identity)

**HMM-GH1\_e0**

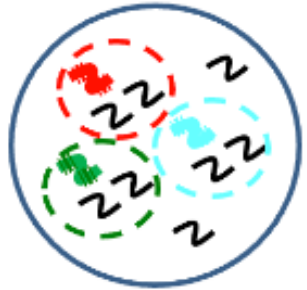
**HMM-GH1\_e1**

multiple sequence alignment and HMM (mafft & hmmbuild)



**Notes:**

1. some proteins do not contain the dbCAN domain (rare but possible)
2. some domains are removed after cd-hit
3. not all subfamilies have CAZy proteins with EC numbers
4. only subfamilies with sequence count  $\geq 4$  after cd-hit are used for HMM construction



# dbCAN-sub

Subfamily → EC → substrates

Database of CAZyme subfamilies for substrate annotation

AA    CBM    CE    GH    GT    PL

AA1   AA2   AA3   AA4   AA5   AA6   AA7   AA8   AA9   AA10   AA11   AA12   AA13   AA14   AA15   AA16   AA17

click each tab to access the six CAZyme classes  
all families of the selected class are shown  
click on each family will expand to show the subfamilies

The screenshot shows a web interface with a navigation menu at the top. The menu items are 'AA', 'CBM', 'CE', 'GH', 'GT', and 'PL'. The 'AA' item is highlighted with a blue border. Below the menu is a list of subfamilies: 'AA1', 'AA2', 'AA3', 'AA4', 'AA5', 'AA6', 'AA7', 'AA8', 'AA9', 'AA10', 'AA11', 'AA12', 'AA13', 'AA14', 'AA15', 'AA16', and 'AA17'. The 'AA1' item is highlighted with a blue background. Below this list is a grid of subfamily names, each with an 'e' suffix, such as 'AA1\_e0', 'AA1\_e1', etc. A blue arrow points from the 'AA1' item in the list to the 'AA1\_e0' item in the grid.

AA	CBM	CE	GH	GT	PL											
AA1	AA2	AA3	AA4	AA5	AA6	AA7	AA8	AA9	AA10	AA11	AA12	AA13	AA14	AA15	AA16	AA17
AA1_e0	AA1_e1	AA1_e2	AA1_e3	AA1_e4	AA1_e5	AA1_e6	AA1_e7	AA1_e8	AA1_e9	AA1_e10	AA1_e11	AA1_e12	AA1_e13	AA1_e14		
AA1_e15	AA1_e16	AA1_e17	AA1_e18	AA1_e19	AA1_e20	AA1_e21	AA1_e22	AA1_e23	AA1_e24	AA1_e25	AA1_e26	AA1_e27	AA1_e28			
AA1_e29	AA1_e30	AA1_e31	AA1_e32	AA1_e33	AA1_e34	AA1_e35	AA1_e36	AA1_e37	AA1_e38	AA1_e39	AA1_e40	AA1_e41	AA1_e42			
AA1_e43	AA1_e44	AA1_e45	AA1_e46													

click on AA1 to show all subfamilies of AA1

the subfamilies were classified by eCAMI: <https://github.com/yinlabniu/eCAMI>

these subfamilies are named with an “e” in them, e.g., AA1\_e0 (e means eCAMI)

click on each subfamily will open a new page

AA1\_e1

[https://bcb.unl.edu/dbCAN\\_sub/dbsub.php?contig=AA1\\_e1](https://bcb.unl.edu/dbCAN_sub/dbsub.php?contig=AA1_e1)

[<- Back to dbCAN-sub](#)

## Summary

summary of CAZy proteins in the subfamily  
(different steps lead to different counts)

Number of CAZy proteins	110
Number of CAZy proteins with ECs	1
Number of CAZy proteins with the corresponding HMM domains	107
Number of HMM domains	107
Number of HMM domains after cd-hit	33
Number of CAZy proteins with the corresponding HMM domains and with ECs	1

### Notes:

1. some proteins do not contain the dbCAN domain (rare but possible)
2. some domains are removed after cd-hit
3. not all subfamilies have CAZy proteins with EC numbers
4. only subfamilies with sequence count  $\geq 4$  after cd-hit are used for HMM construction

[Download HMM domains after cd-hit](#) [Download AA1\\_e1](#)

## Substrate Table

CAZy proteins with EC numbers in  
this subfamily

Show  entries Search:

EC	Count	CAZy protein ID	Substrates
<a href="#">1.10.3.2</a>	1	<a href="#">AFC76164.1</a>	<a href="#">lignin</a>

Showing 1 to 1 of 1 entries First Previous 1 Next La

[Download Substrate Table](#)

manually curated family-EC-substrate mapping table

# Search dbCAN-sub @ the dbCAN meta server

<https://bcb.unl.edu/dbCAN2/blast.php>

check here to search dbCAN-sub  
for substrate prediction

## Choose Sequence Type:

Protein sequence ([example](#)) ?  Nucleotide sequence ([example](#)) ?

## Select Which Tools To Run

HMMER: dbCAN (E-Value < 1e-15, coverage > 0.35)  DIAMOND: CAZy (E-Value < 1e-102)  HMMER: dbCAN-sub (E-Value < 1e-15, coverage > 0.35)   
CGCFinder (Distance <= 2, signature genes = CAZyme+TC)?

▼ Just paste some sequences here (**note: only FASTA format please!!!**)

Try [example](#) sequences

# Search dbCAN-sub @ the dbCAN meta server

<https://bcb.unl.edu/dbCAN2/blastation.php?jobid=20220701162102>

dbCAN-sub result tab  
for details

link to dbCAN-sub page

this col has dbCAN-sub result

Cite us: [NAR/gky418](#) and [gks479](#)

Result of job: 20220701162102

Overview

HMMER: dbCAN

DIAMOND: CAZy

HMMER: dbCAN\_sub

[Download SignalP output](#) [Download input file](#) [Download this table](#) (keep those with # of Tools >=2 will give you best result; and use dbCAN domain assignment is recommended)

Show 15 entries

Search:

Gene ID	EC#	HMMER	DIAMOND	dbCAN_sub	Signal Peptide	# of Tools
<a href="#">AT1G11720.1</a>	<a href="#">2.4.1.21</a>	<a href="#">CBM53</a> (154-237)+ <a href="#">CBM53</a> (329-423)+ <a href="#">CBM53</a> (496-584)+ <a href="#">GT5</a> (595-1038)	<a href="#">CBM53+GT5</a>	<a href="#">CBM53_e1+CBM53_e0+CBM53_e0+GT5_e38</a>	N	3
<a href="#">gj 222529846 ref YP_002573728.1</a>	<a href="#">3.2.1.4 3.2.1.176 3.2.1.73 3.2.1.- 3.2.1.78 3.2.1.151 3.2.1.8 3.2.1.55 3.2.1.74 3.2.1.91 3.2.1.14</a>	<a href="#">GH9</a> (36-466)+ <a href="#">CBM3</a> (491-576)+ <a href="#">CBM3</a> (724-804)+ <a href="#">CBM3</a> (923-1003)+ <a href="#">GH48</a> (1134-1755)	<a href="#">CBM3+GH48+GH9</a>	<a href="#">GH9_e22+CBM3_e0+CBM3_e16+CBM3_e16+GH48_e1</a>	N	3

dbCAN-sub result tab for details

link to dbCAN-sub page

these are the seq composition of CBM53\_e1 HMM

CBM53:91|GT5:87|GT31:2 means the CBM53\_e1 subfam contains 91 CAZy proteins from CBM53 fam, 87 from GT5, and 2 from GT31 (CAZymes are often multi-domain proteins)

Cite us: [NAR/gky418](#) and [gks479](#)

Result of job: 20220701162102

[Overview](#) [HMMER: dbCAN](#) [DIAMOND: CAZy](#) **[HMMER: dbCAN\\_sub](#)**

[Download dbCAN\\_sub output](#)

Show  entries

Search:

Query ID	Query Length	dbCAN Subfam	Subfam Composition	Subfam EC	substrate	E Value	Coverage
AT1G11720.1	89	<a href="#">CBM53_e1</a>	CBM53:91 GT5:87 GT31:2	<a href="#">"2.4.1.21:27"</a>	starch	6.3e-28	0.966292134831
AT1G11720.1	89	<a href="#">CBM53_e0</a>	CBM53:62 GT5:59	<a href="#">"2.4.1.21:13"</a>	starch	1.5e-28	0.977528089888
AT1G11720.1	89	<a href="#">CBM53_e0</a>	CBM53:62 GT5:59	<a href="#">"2.4.1.21:13"</a>	starch	5.8e-28	0.955056179775
AT1G11720.1	445	<a href="#">GT5_e38</a>	GT5:214 CBM53:145 GT31:2	<a href="#">"2.4.1.21:40"</a>	-	8.5e-237	0.997752808989

# CBM53\_e1 sequence composition is described in the dbCAN-sub webpage

[< Back to Job Overview](#)

Cite us: [NAR/gky418](#) and [gks479](#)

## CBM53\_e1

### Summary

Number of CAZy proteins	91
Number of CAZy proteins with ECs	27
Number of CAZy proteins with the corresponding HMM domains	91
Number of HMM domains	91
Number of HMM domains after cd-hit	85
Number of CAZy proteins with the corresponding HMM domains and with ECs	27

[Download HMM domains after cd-hit](#) [Download CBM53\\_e1](#)

### Substrate Table

Show  entries

Search:

EC	Count	CAZy protein ID	Substrates
-	-	-	<a href="#">starch</a>
<a href="#">2.4.1.21</a>	27	<a href="#">AEE28775.1</a> <a href="#">CAA64173.1</a> <a href="#">ABE21468.1</a> <a href="#">AFF28774.1</a>	-