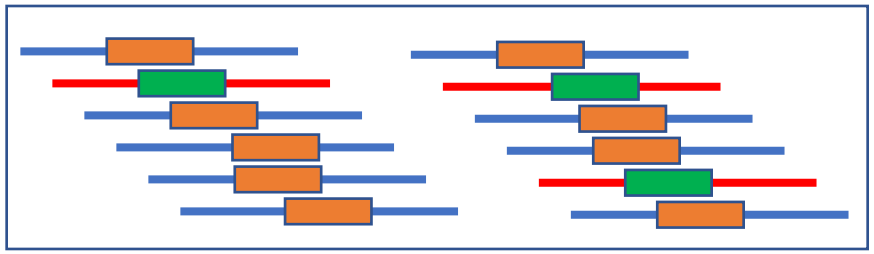
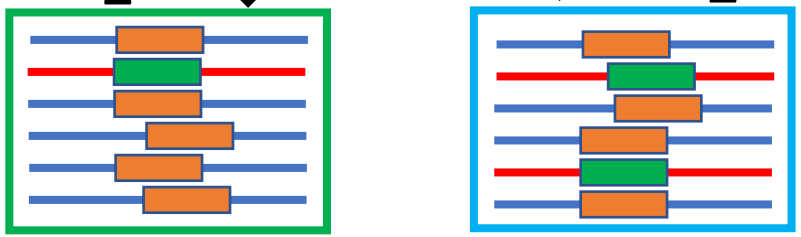
workflow of dbCAN-sub HMM construction

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

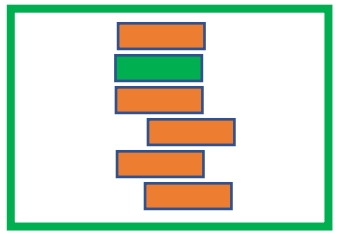


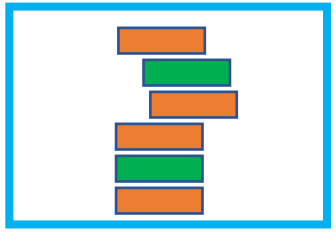
proteins of a CAZy family (e.g., GH1)

GH1_e0 ↓ *eCAMI* *Bioinformatics 2020* ↓ **GH1_e1**

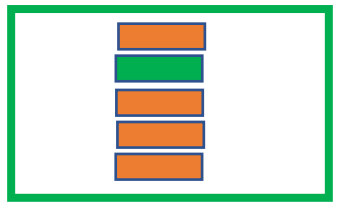


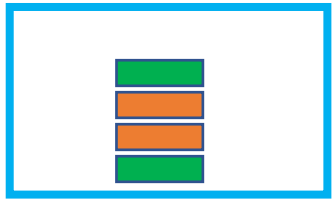
proteins of eCAMI subfamilies (e.g., GH1_e0)





dbCAN GH1 domains of proteins of GH1_e0 (hmmscan)





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

HMM-GH1_e0

HMM-GH1_e1

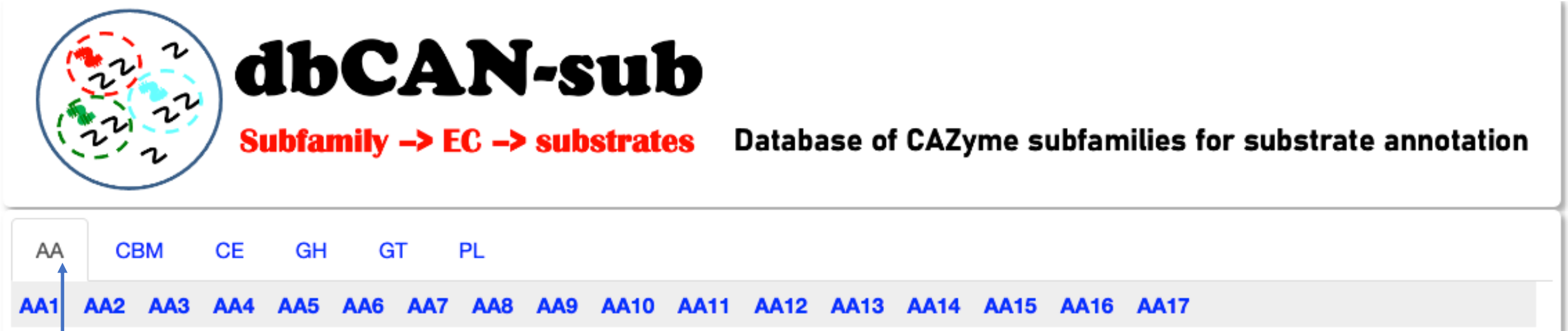
multiple sequence alignment and HMM (mafft & hmmbuild)

CAZy proteins with EC numbers

CAZy proteins without EC numbers

Notes:

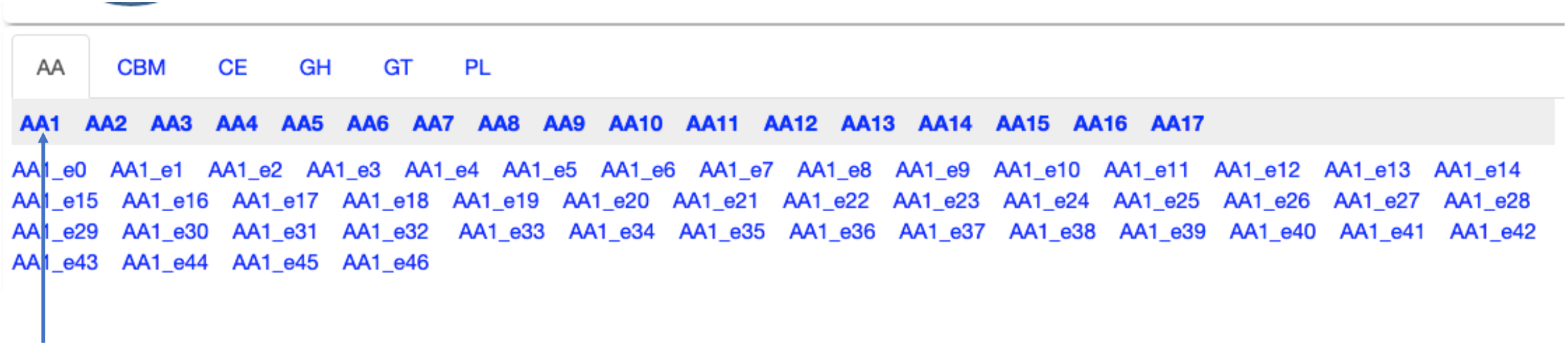
1. some proteins do not contain the dbCAN domain (rare but possible)
2. some domains are removed after cd-hit
3. not all subfamilies have CAZy proteins with EC numbers
4. only subfamilies with sequence count >=4 after cd-hit are used for HMM construction



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



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

<- 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 >=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 15 entries Search:

EC	Count	CAZy protein ID	Substrates
1.10.3.2	1	AFC76164.1	lignin

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

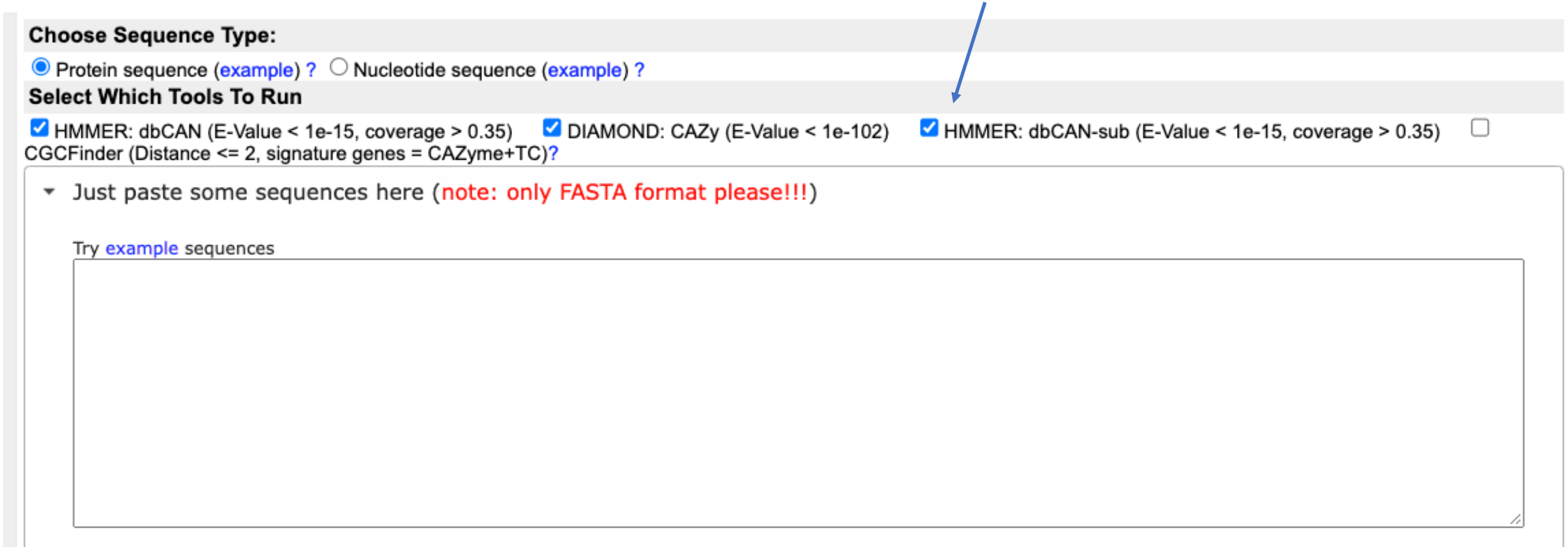
Download Substrate Table

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



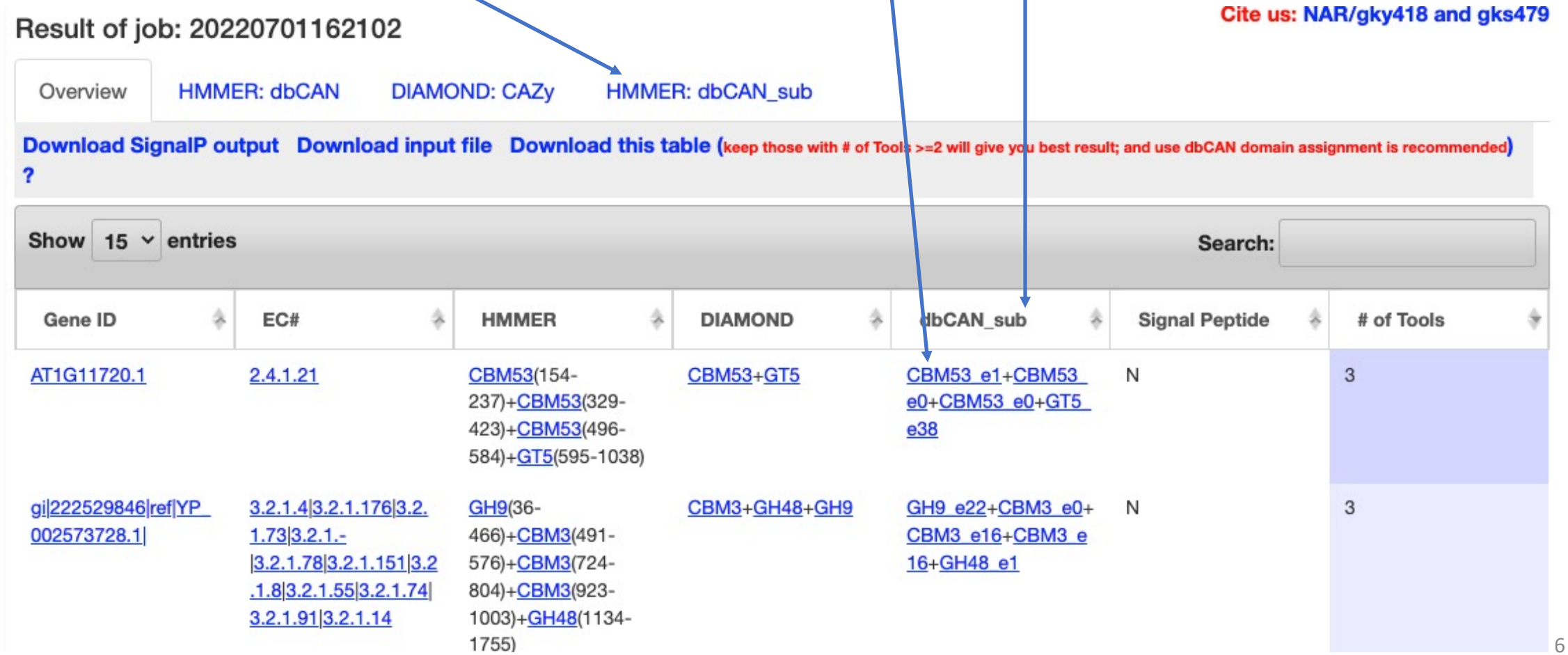
Search dbCAN-sub @ the dbCAN meta server

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



Result of job: 20220701162102

Cite us: NAR/gky418 and gks479

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
AT1G11720.1	2.4.1.21	CBM53(154-237)+CBM53(329-423)+CBM53(496-584)+GT5(595-1038)	CBM53+GT5	CBM53_e1+CBM53_e0+CBM53_e0+GT5_e38	N	3
gi|222529846|ref|YP_002573728.1|	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	GH9(36-466)+CBM3(491-576)+CBM3(724-804)+CBM3(923-1003)+GH48(1134-1755)	CBM3+GH48+GH9	GH9_e22+CBM3_e0+CBM3_e16+CBM3_e16+GH48_e1	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 15 entries Search:

Query ID	Query Length	dbCAN Subfam	Subfam Composition	Subfam EC	substrate	E Value	Coverage
AT1G11720.1	89	CBM53_e1	CBM53:91|GT5:87|GT31:2	"2.4.1.21:27"	starch	6.3e-28	0.966292134831
AT1G11720.1	89	CBM53_e0	CBM53:62|GT5:59	"2.4.1.21:13"	starch	1.5e-28	0.977528089888
AT1G11720.1	89	CBM53_e0	CBM53:62|GT5:59	"2.4.1.21:13"	starch	5.8e-28	0.955056179775
AT1G11720.1	445	GT5_e38	GT5:214|CBM53:145|GT31:2	"2.4.1.21:40"	-	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 15 entries Search:

EC	Count	CAZy protein ID	Substrates
-	-	-	starch
2.4.1.21	27	AEE28775.1 CAA64173.1 ABE21468.1 AEE28774.1	-