# Project title: Subfamily classification and analysis of CAZy family

**Goal: Classify the CAZy family into subfamilies, analyze and visualize them using Bioinformatics tools and methods.**

In this tutorial, CAZy family GH31 was used to explain bioinformatics methods to analyze and visualize them.

**Paper Link:** A subfamily classification to choreograph the diverse activities within glycoside hydrolase family 31. **DOI:** https://doi.org/10.1016/j.jbc.2023.103038

**Required materials are available at https://github.com/sivanr92/Class_project**

**Updated: Xinpeng, 9/19/2023**

**Updated: Siva, 02 Oct 2023, 3.20 AM.**

**Updated: Xinpeng 02 Oct.**

**Methods to identify the subfamily from a family:**

1. Dataset and preprocess
2. Domain annotation using HMM and dbCAN
3. Extraction of modules (based on the annotation)
4. Construction of sequence similarity networks (SSN) using SSNpipe and analyzing SSNs based on characterized IDs from CAZy and EC numbers.
5. Visualization of SSN networks using Cytoscape
6. Phylogenentic Analysis
7. Interpretation and Discussion

**Detailed tutorials and methods are available in the following pages.**

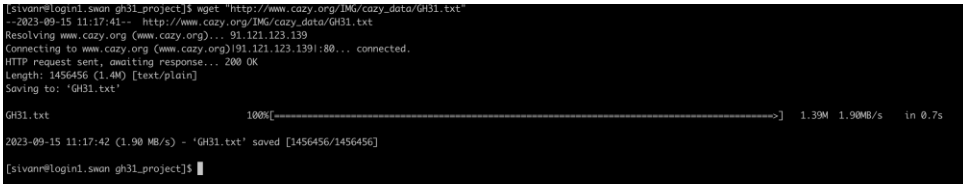
Note: Denotes the program name available at the GitHub.

**Detailed methods section:**

## 1. Dataset and preprocess:

- Dataset can be downloaded from CAZy database
  http://www.cazy.org/
- Select GH classes on CAZy database (Your family under the classification, we will use GH31 as example, please choose your own family)
  http://www.cazy.org/Glycoside-Hydrolases.html
- Go to GH class 31 and download the dataset.
  http://www.cazy.org/GH31.html
- Find the page you can find the dataset
  http://www.cazy.org/IMG/cazy_data/GH31.txt
  Download that using the wget in the terminal as use click and download.
  Command line:

```
$ wget "http://www.cazy.org/IMG/cazy_data/GH31.txt"
```



```
[sivanr@login1.swan gh31_project]$ wget "http://www.cazy.org/IMG/cazy_data/GH31.txt"
--2023-09-15 11:17:41--  http://www.cazy.org/IMG/cazy_data/GH31.txt
Resolving www.cazy.org (www.cazy.org)... 91.121.123.139
Connecting to www.cazy.org (www.cazy.org)|91.121.123.139|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1456456 (1.4M) [text/plain]
Saving to: ‘GH31.txt’

GH31.txt                          100%[=================================================================================>]   1.39M  1.90MB/s    in 0.7s

2023-09-15 11:17:42 (1.90 MB/s) - ‘GH31.txt’ saved [1456456/1456456]

[sivanr@login1.swan gh31_project]$ █
```

Use GH31.txt to get a list of genbank ids of GH31 family.

Use the following command in the terminal make to make the unique list of IDs

```
$ cut -f 4 GH31.txt > list_ids.txt
```

Remove redundant ids from the list

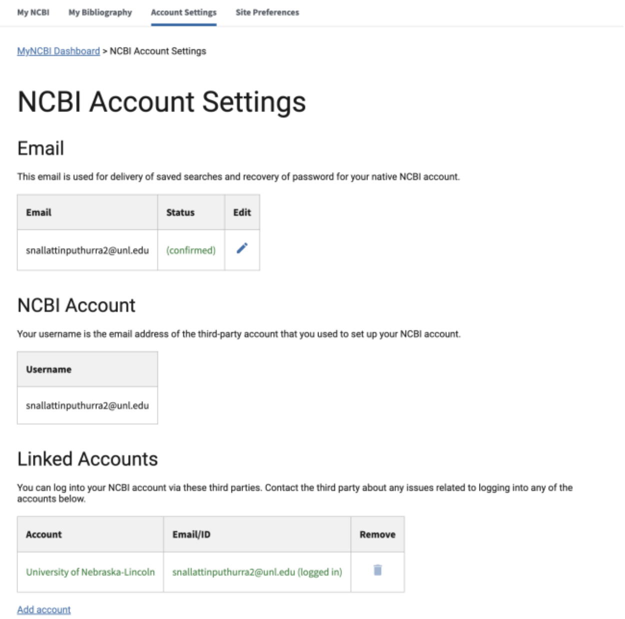
```
$ cat list_ids.txt | sort | uniq > uniq_list_ids.txt
```

24307 unique ids were unique.

To download the sequences from the NCBI. Use the batch service available from NCBI Entrez Direct: E-utilities

Full tutorial Link: https://www.ncbi.nlm.nih.gov/books/NBK179288/

- Register in NCBI using the for using the API-KEY. Helps for fast and terminal download.
- NCBI register using UNL mail id. (Use Institution search)
- Navigate to account settings.
- Find the Generate API-KEY button and generate.



MyNCBI Dashboard > NCBI Account Settings

NCBI Account Settings

Email

This email is used for delivery of saved searches and recovery of password for your native NCBI account.

| Email | Status | Edit |
|---|---|---|
| snallattinputhurra2@unl.edu | (confirmed) | ✎ |

NCBI Account

Your username is the email address of the third-party account that you used to set up your NCBI account.

| Username |
|---|
| snallattinputhurra2@unl.edu |

Linked Accounts

You can log into your NCBI account via these third parties. Contact the third party about any issues related to logging into any of the accounts below.

| Account | Email/ID | Remove |
|---|---|---|
| University of Nebraska-Lincoln | snallattinputhurra2@unl.edu (logged in) | 🗑 |

Add account

Install EDirect using one of the following.

```
$ sh -c "$(curl -fsSL https://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect.sh)"
```

**OR**

```
$ sh -c "$(wget -q https://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect.sh -O -)"
```

```
$ export PATH=${HOME}/edirect:${PATH}. (After installation, an automatic path will be shown. Export that)
```

After that export your unique API-KEY, using the following command,

```
$ export NCBI_API_KEY=unique_api_key
```

**Program to download the sequences can be found at** “batch_download_protein_seqeunces.sh”

Note: remember to edit your mail id and batch size (500 is maximum as of now)

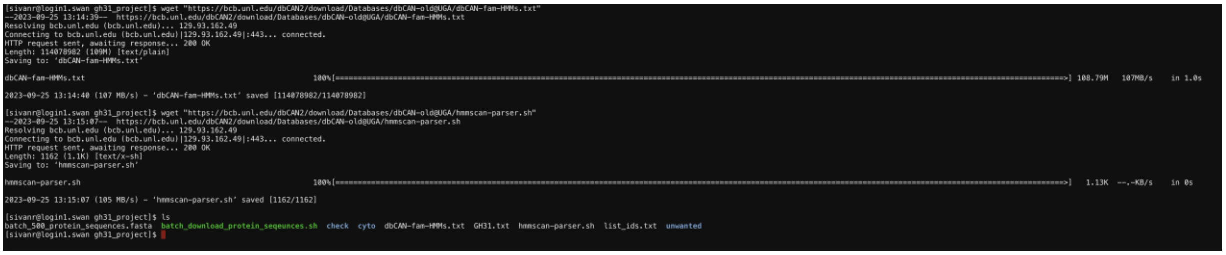
24306 ids were found and downloaded.

## 2. Domain annotation using HMM and dbCAN.

### (i) download dbCAN-fam-HMMs.txt, hmmscan-parser.sh

Download following files from https://bcb.unl.edu/dbCAN2/download/Databases/dbCAN-old@UGA/

- hmmscan-parser.sh
- dbCAN-fam-HMMs.txt



```
[sivanr@login1.swan gh31_project]$ wget "https://bcb.unl.edu/dbCAN2/download/Databases/dbCAN-old@UGA/dbCAN-fam-HMMs.txt"
--2023-09-25 13:14:39--  https://bcb.unl.edu/dbCAN2/download/Databases/dbCAN-old@UGA/dbCAN-fam-HMMs.txt
Resolving bcb.unl.edu (bcb.unl.edu)... 129.93.162.49
Connecting to bcb.unl.edu (bcb.unl.edu)|129.93.162.49|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 114078982 (109M) [text/plain]
Saving to: 'dbCAN-fam-HMMs.txt'

dbCAN-fam-HMMs.txt                              100%[======================================================================================================>] 108.79M   107MB/s    in 1.0s

2023-09-25 13:14:40 (107 MB/s) - 'dbCAN-fam-HMMs.txt' saved [114078982/114078982]

[sivanr@login1.swan gh31_project]$ wget "https://bcb.unl.edu/dbCAN2/download/Databases/dbCAN-old@UGA/hmmscan-parser.sh"
--2023-09-25 13:15:07--  https://bcb.unl.edu/dbCAN2/download/Databases/dbCAN-old@UGA/hmmscan-parser.sh
Resolving bcb.unl.edu (bcb.unl.edu)... 129.93.162.49
Connecting to bcb.unl.edu (bcb.unl.edu)|129.93.162.49|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1162 (1.1K) [text/x-sh]
Saving to: 'hmmscan-parser.sh'

hmmscan-parser.sh                               100%[======================================================================================================>]   1.13K  --.-KB/s    in 0s

2023-09-25 13:15:07 (105 MB/s) - 'hmmscan-parser.sh' saved [1162/1162]

[sivanr@login1.swan gh31_project]$ ls
batch_500_protein_sequences.fasta  batch_download_protein_seqeunces.sh  check  cyto  dbCAN-fam-HMMs.txt  GH31.txt  hmmscan-parser.sh  list_ids.txt  unwanted
[sivanr@login1.swan gh31_project]$
```

### (ii) download HMMER 3.0 package [hmmer.org] and install it properly

http://hmmer.org/download.html, File: http://eddylab.org/software/hmmer/hmmer-3.4.tar.gz

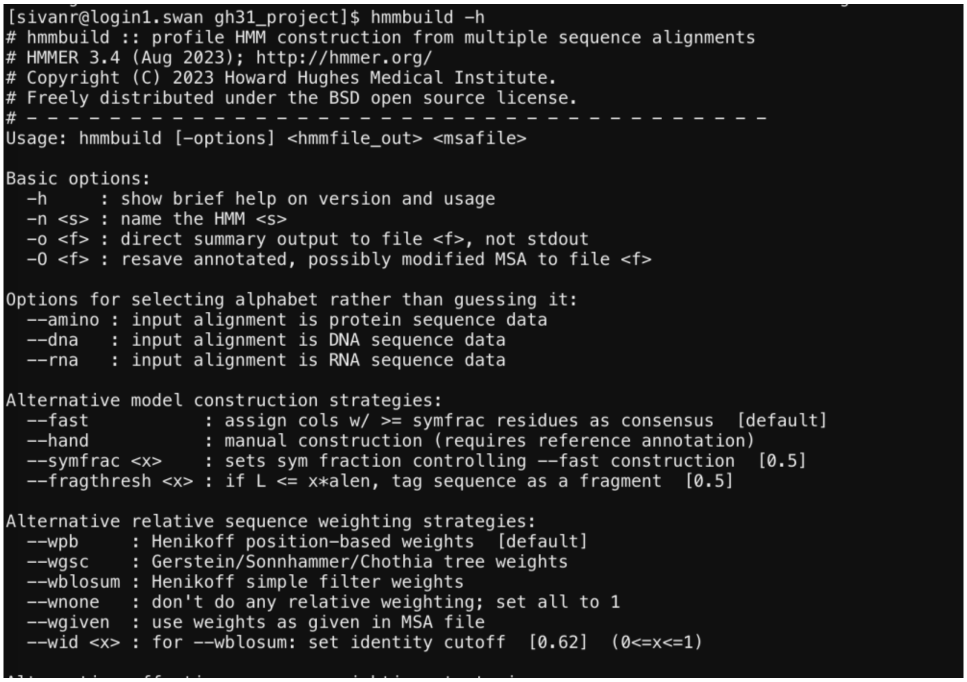
HMMER is available in HCC.

- Can be loaded with the following commands.

```
$ ml hmmer/3.3
```

- Check the installation by running:

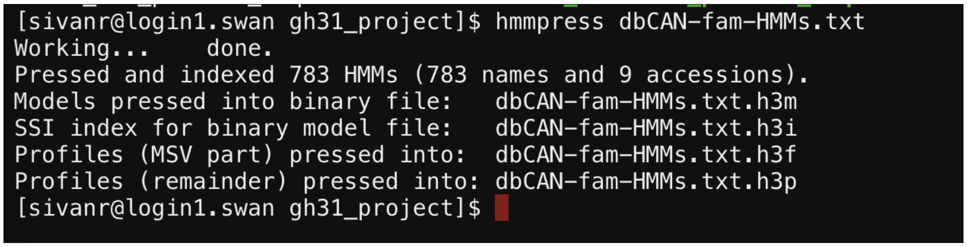
```
$ hmmbuild -h
```



```
[sivanr@login1.swan gh31_project]$ hmmbuild -h
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.4 (Aug 2023); http://hmmer.org/
# Copyright (C) 2023 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Usage: hmmbuild [-options] <hmmfile_out> <msafile>

Basic options:
  -h     : show brief help on version and usage
  -n <s> : name the HMM <s>
  -o <f> : direct summary output to file <f>, not stdout
  -O <f> : resave annotated, possibly modified MSA to file <f>

Options for selecting alphabet rather than guessing it:
  --amino : input alignment is protein sequence data
  --dna   : input alignment is DNA sequence data
  --rna   : input alignment is RNA sequence data

Alternative model construction strategies:
  --fast           : assign cols w/ >= symfrac residues as consensus  [default]
  --hand           : manual construction (requires reference annotation)
  --symfrac <x>    : sets sym fraction controlling --fast construction  [0.5]
  --fragthresh <x> : if L <= x*alen, tag sequence as a fragment  [0.5]

Alternative relative sequence weighting strategies:
  --wpb     : Henikoff position-based weights  [default]
  --wgsc    : Gerstein/Sonnhammer/Chothia tree weights
  --wblosum : Henikoff simple filter weights
  --wnone   : don't do any relative weighting; set all to 1
  --wgiven  : use weights as given in MSA file
  --wid <x> : for --wblosum: set identity cutoff  [0.62]  (0<=x<=1)
```

**(iii) format HMM db: hmmpress dbCAN-fam-HMMs.txt**

```
$ hmmpress dbCAN-fam-HMMs.txt
```



```
[sivanr@login1.swan gh31_project]$ hmmpress dbCAN-fam-HMMs.txt
Working...    done.
Pressed and indexed 783 HMMs (783 names and 9 accessions).
Models pressed into binary file:   dbCAN-fam-HMMs.txt.h3m
SSI index for binary model file:   dbCAN-fam-HMMs.txt.h3i
Profiles (MSV part) pressed into:  dbCAN-fam-HMMs.txt.h3f
Profiles (remainder) pressed into: dbCAN-fam-HMMs.txt.h3p
[sivanr@login1.swan gh31_project]$
```

**(iv) run: hmmscan --domtblout yourfile.out.dm dbCAN-fam-HMMs.txt yourfile > yourfile.out**

```
$ hmmscan --domtblout gh31_hmmscan.out.dm dbCAN-fam-HMMs.txt
batch_500_protein_sequences.fasta > gh31_hmmscan.out
```

Use the code below to run hmmscan on HCC

Program available at “hmmscan_run.slurm”

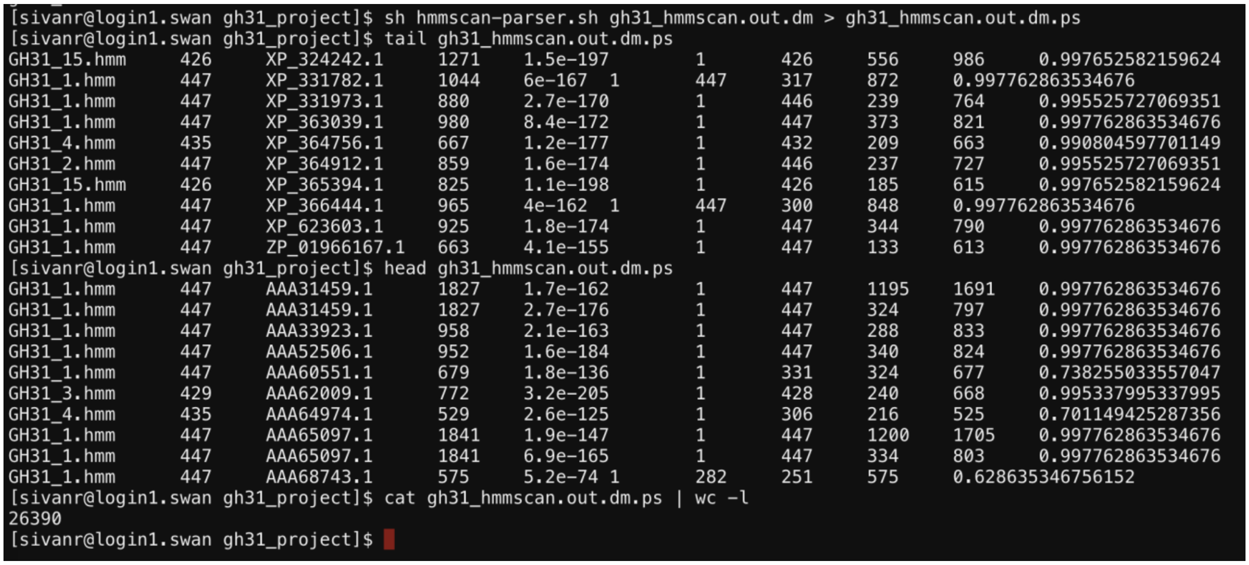
Students can use batch and guest partition.

Sample code in below.

```
#!/bin/bash
#SBATCH --time=24:00:00
#SBATCH --mem=100gb
#SBATCH --job-name=hmmscan
#SBATCH --error=/Your/path/gh31_project/job.%J.err
#SBATCH --output=/ Your/path/gh31_project/job.%J.out
#SBATCH --partition=batch,guest
ml hmmer/3.3
hmmscan --domtblout gh31_hmmscan.out.dm dbCAN-fam-HMMs.txt
batch_500_protein_sequences.fasta > gh31_hmmscan.out
```

**(v) run: sh hmmscan-parser.sh yourfile.out.dm > yourfile.out.dm.ps (if alignment > 80aa, use E-value < 1e-5, otherwise use E-value < 1e-3; covered fraction of HMM > 0.3)**

```
$ sh hmmscan-parser.sh gh31_hmmscan.out.dm >
gh31_hmmscan.out.dm.ps
```

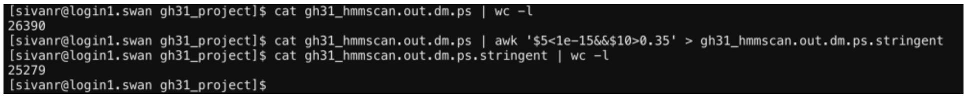


```
[sivanr@login1.swan gh31_project]$ sh hmmscan-parser.sh gh31_hmmscan.out.dm > gh31_hmmscan.out.dm.ps
[sivanr@login1.swan gh31_project]$ tail gh31_hmmscan.out.dm.ps
GH31_15.hmm     426     XP_324242.1     1271    1.5e-197        1       426     556     986     0.997652582159624
GH31_1.hmm      447     XP_331782.1     1044    6e-167  1       447     317     872     0.997762863534676
GH31_1.hmm      447     XP_331973.1     880     2.7e-170        1       446     239     764     0.995525727069351
GH31_1.hmm      447     XP_363039.1     980     8.4e-172        1       447     373     821     0.997762863534676
GH31_4.hmm      435     XP_364756.1     667     1.2e-177        1       432     209     663     0.990804597701149
GH31_2.hmm      447     XP_364912.1     859     1.6e-174        1       446     237     727     0.995525727069351
GH31_15.hmm     426     XP_365394.1     825     1.1e-198        1       426     185     615     0.997652582159624
GH31_1.hmm      447     XP_366444.1     965     4e-162  1       447     300     848     0.997762863534676
GH31_1.hmm      447     XP_623603.1     925     1.8e-174        1       447     344     790     0.997762863534676
GH31_1.hmm      447     ZP_01966167.1   663     4.1e-155        1       447     133     613     0.997762863534676
[sivanr@login1.swan gh31_project]$ head gh31_hmmscan.out.dm.ps
GH31_1.hmm      447     AAA31459.1      1827    1.7e-162        1       447     1195    1691    0.997762863534676
GH31_1.hmm      447     AAA31459.1      1827    2.7e-176        1       447     324     797     0.997762863534676
GH31_1.hmm      447     AAA33923.1      958     2.1e-163        1       447     288     833     0.997762863534676
GH31_1.hmm      447     AAA52506.1      952     1.6e-184        1       447     340     824     0.997762863534676
GH31_1.hmm      447     AAA60551.1      679     1.8e-136        1       331     324     677     0.738255033557047
GH31_3.hmm      429     AAA62009.1      772     3.2e-205        1       428     240     668     0.995337995337995
GH31_4.hmm      435     AAA64974.1      529     2.6e-125        1       306     216     525     0.701149425287356
GH31_1.hmm      447     AAA65097.1      1841    1.9e-147        1       447     1200    1705    0.997762863534676
GH31_1.hmm      447     AAA65097.1      1841    6.9e-165        1       447     334     803     0.997762863534676
GH31_1.hmm      447     AAA68743.1      575     5.2e-74 1       282     251     575     0.628635346756152
[sivanr@login1.swan gh31_project]$ cat gh31_hmmscan.out.dm.ps | wc -l
26390
[sivanr@login1.swan gh31_project]$
```

**26390 modules were found.**

**(vi) run: cat yourfile.out.dm.ps | awk '$5<1e-15&&$10>0.35' > yourfile.out.dm.ps.stringent (this allows you to get the same result as what is produced in our dbCAN2 webpage)**

```
$
```



```
[sivanr@login1.swan gh31_project]$ cat gh31_hmmscan.out.dm.ps | wc -l
26390
[sivanr@login1.swan gh31_project]$ cat gh31_hmmscan.out.dm.ps | awk '$5<1e-15&&$10>0.35' > gh31_hmmscan.out.dm.ps.stringent
[sivanr@login1.swan gh31_project]$ cat gh31_hmmscan.out.dm.ps.stringent | wc -l
25279
[sivanr@login1.swan gh31_project]$
```

After filtering using the E- Value and coverage cutoff of <1e-15 and >0.35, **25279 modules were found**.

Cols in yourfile.out.dm.ps:
1. Family HMM
2. HMM length
3. Query ID
4. Query length
5. E-value (how similar to the family HMM)
6. HMM start
7. HMM end
8. Query start
9. Query end
10. Coverage

** On our dbCAN2 website, we use E-value < 1e-15 and coverage > 0.35, which is more stringent than the default ones in hmmscan-parser.sh

### 3. Extract only GH31 modules from the overview.txt file by matching with the sequences.

Use a given python program to extract the modules from the dataset.

Refer python Program filter_modules_26sep2023.py

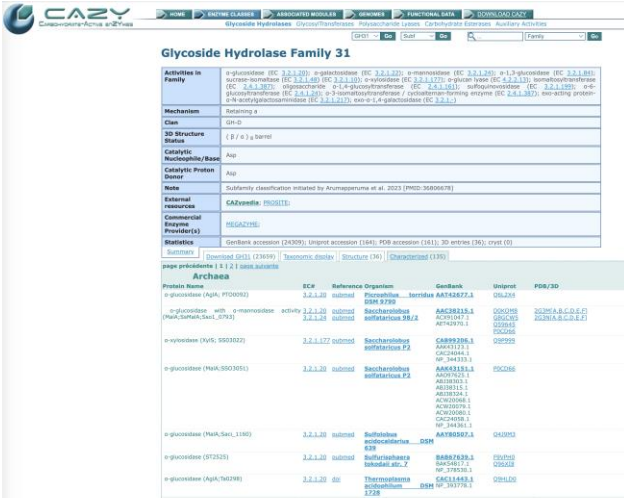
25279 modules were found.

Note: module load in HCC.

```
$ ml biopython
```

**Extract manually characterized IDs from the GH31 page of the CAZy.**

- Cilck on the “Characterized (135)”
- Tabular column page will be shown.
- Copy the Characterized IDs, EC numbers from the page and paste them in Excel and select only the IDs which only have hyperlinks (Characterized by CAZy) and make a list.



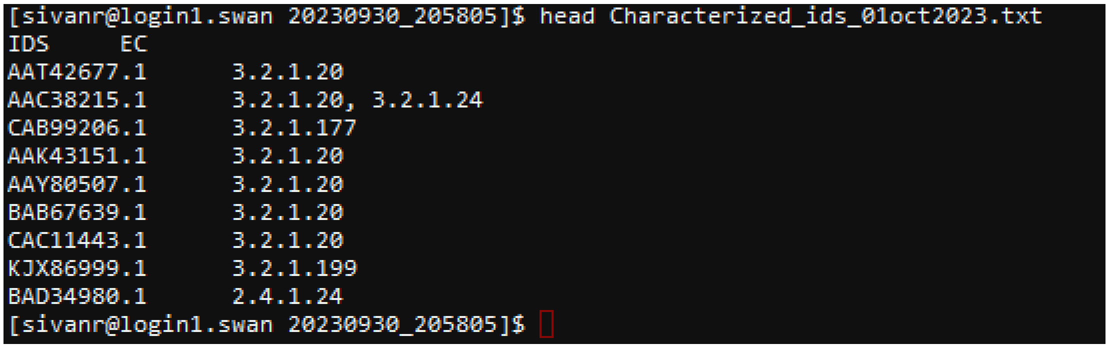
Glycoside Hydrolase Family 31

| | |
|---|---|
| Activities in Family | α-glucosidase (EC 3.2.1.20); α-galactosidase (EC 3.2.1.22); α-mannosidase (EC 3.2.1.24); α-1,3-glucosidase (EC 3.2.1.84); sucrase-isomaltase (EC 3.2.1.48) (EC 3.2.1.10); α-xylosidase (EC 3.2.1.177); α-glucan lyase (EC 4.2.2.13); isomaltosyltransferase (EC 2.4.1.387); oligosaccharide α-1,4-glucosyltransferase (EC 2.4.1.161); sulfoquinovosidase (EC 3.2.1.199); α-6-glucosyltransferase (EC 2.4.1.24); α-3-isomaltosyltransferase / cycloalternan-forming enzyme (EC 2.4.1.387); exo-acting protein-α-N-acetylgalactosaminidase (EC 3.2.1.217); exo-α-1,4-galactosidase (EC 3.2.1.-) |
| Mechanism | Retaining α |
| Clan | GH-D |
| 3D Structure Status | ( β / α ) $_8$ barrel |
| Catalytic Nucleophile/Base | Asp |
| Catalytic Proton Donor | Asp |
| Note | Subfamily classification initiated by Arumapperuma et al. 2023 [PMID:36806678] |
| External resources | CAZypedia; PROSITE; |
| Commercial Enzyme Provider(s) | MEGAZYME; |
| Statistics | GenBank accession (24309); Uniprot accession (164); PDB accession (161); 3D entries (36); cryst (0) |

Archaea

| Protein Name | EC# | Reference | Organism | GenBank | Uniprot | PDB/3D |
|---|---|---|---|---|---|---|
| α-glucosidase (AglA; PTO0092) | 3.2.1.20 | pubmed | Picrophilus torridus DSM 9790 | AAT42677.1 | Q6L2X4 | |
| α-glucosidase with α-mannosidase activity (MalA;SsMalA;Saci_0793) | 3.2.1.20<br>3.2.1.24 | pubmed<br>pubmed | Saccharolobus solfataricus 98/2 | AAC38215.1<br>ACX91047.1<br>AET42970.1 | D0KQM8<br>G8GCW5<br>Q59645<br>P0CD66 | 2G3M[A,B,C,D,E,F]<br>2G3N[A,B,C,D,E,F] |
| α-xylosidase (XylS; SSO3022) | 3.2.1.177 | pubmed | Saccharolobus solfataricus P2 | CAB99206.1<br>AAK43123.1<br>CAC24044.1<br>NP_344333.1 | Q9P999 | |
| α-glucosidase (MalA;SSO3051) | 3.2.1.20 | pubmed | Saccharolobus solfataricus P2 | AAK43151.1<br>AAO97625.1<br>ABJ38303.1<br>ABJ38315.1<br>ABJ38324.1<br>ACW20068.1<br>ACW20079.1<br>ACW20080.1<br>CAC24058.1<br>NP_344361.1 | P0CD66 | |
| α-glucosidase (MalA;Saci_1160) | 3.2.1.20 | pubmed | Sulfolobus acidocaldarius DSM 639 | AAY80507.1 | Q4J9M3 | |
| α-glucosidase (ST2525) | 3.2.1.20 | pubmed | Sulfurisphaera tokodaii str. 7 | BAB67639.1<br>BAK54817.1<br>NP_378530.1 | F9VPH0<br>Q96XI9 | |
| α-glucosidase (AglA;Ta0298) | 3.2.1.20 | doi | Thermoplasma acidophilum DSM 1728 | CAC11443.1<br>NP_393778.1 | Q9HLD0 | |

Place it in a file with the following format.
IDs EC_Number (example below)

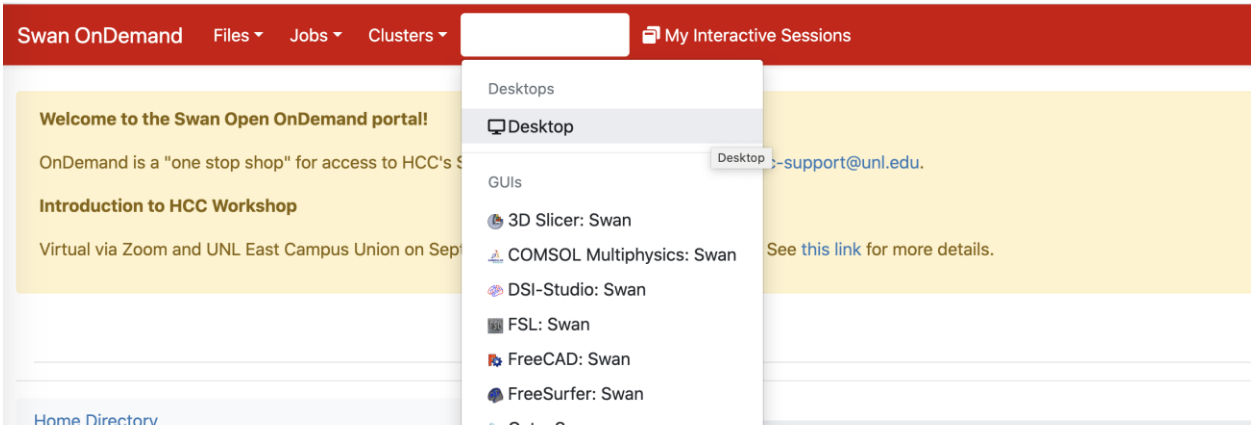


```
[sivanr@login1.swan 20230930_205805]$ head Characterized_ids_01oct2023.txt
IDS     EC
AAT42677.1      3.2.1.20
AAC38215.1      3.2.1.20, 3.2.1.24
CAB99206.1      3.2.1.177
AAK43151.1      3.2.1.20
AAY80507.1      3.2.1.20
BAB67639.1      3.2.1.20
CAC11443.1      3.2.1.20
KJX86999.1      3.2.1.199
BAD34980.1      2.4.1.24
[sivanr@login1.swan 20230930_205805]$
```

## 4. Construction of sequence similarity networks

### SSNpipe

### Donwload SSNpipe in the HCC virtual computer by invoking the desktop mode in the interactive application option



Welcome to the Swan Open OnDemand portal!

OnDemand is a "one stop shop" for access to HCC's S

-support@unl.edu.

Introduction to HCC Workshop

Virtual via Zoom and UNL East Campus Union on Sept

See this link for more details.

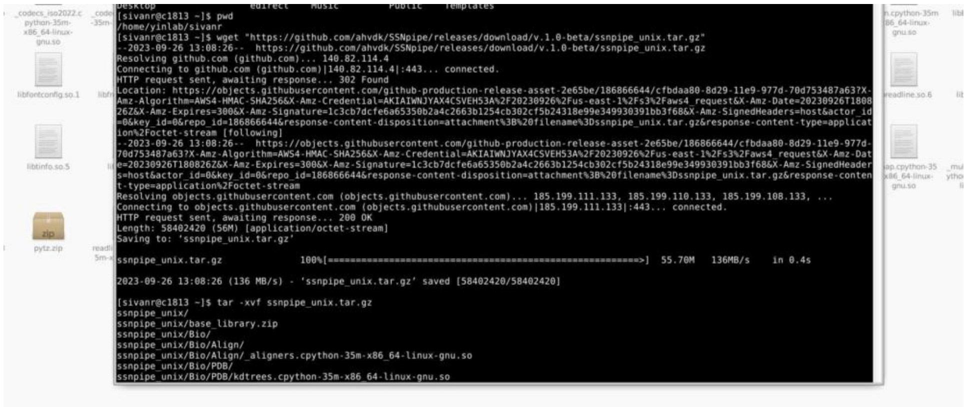
**Open terminal and execute the following command**

[https://github.com/ahvdk/SSNpipe/releases/download/v.1.0-beta/ssnpipe_unix.tar.gz](https://github.com/ahvdk/SSNpipe/releases/download/v.1.0-beta/ssnpipe_unix.tar.gz)

$ wget “https://github.com/ahvdk/SSNpipe/releases/download/v.1.0-beta/ssnpipe_unix.tar.gz”

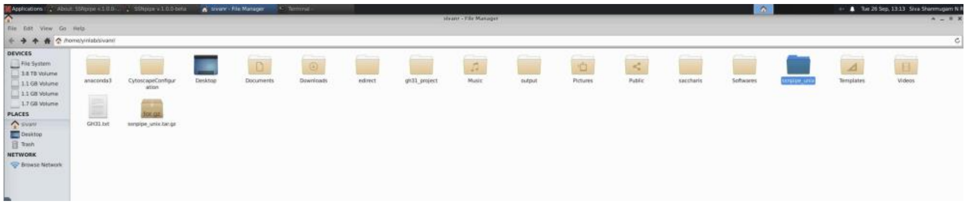
**Extract using following command**

$ tar –xvf ssnpipe_unix.tar.gz

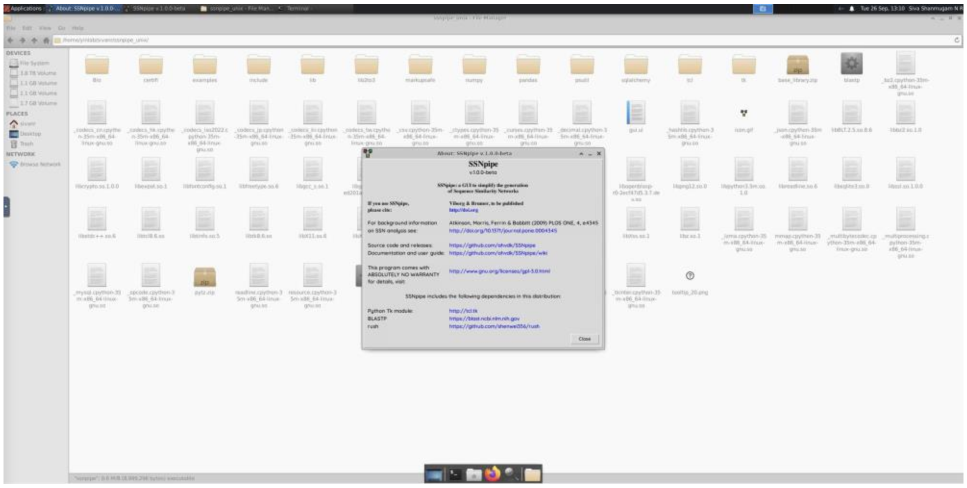


```
Desktop             edirect    Music        Public    Templates
[sivanr@c1813 ~]$ pwd
/home/yinlab/sivanr
[sivanr@c1813 ~]$ wget "https://github.com/ahvdk/SSNpipe/releases/download/v.1.0-beta/ssnpipe_unix.tar.gz"
--2023-09-26 13:08:26--  https://github.com/ahvdk/SSNpipe/releases/download/v.1.0-beta/ssnpipe_unix.tar.gz
Resolving github.com (github.com)... 140.82.114.4
Connecting to github.com (github.com)|140.82.114.4|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://objects.githubusercontent.com/github-production-release-asset-2e65be/186866644/cfbdaa80-8d29-11e9-977d-70d753487a63?X-
Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWNJYAX4CSVEH53A%2F20230926%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20230926T1808
26Z&X-Amz-Expires=300&X-Amz-Signature=1c3cb7dcfe6a65350b2a4c2663b1254cb302cf5b24318e99e349930391bb3f68&X-Amz-SignedHeaders=host&actor_id
=0&key_id=0&repo_id=186866644&response-content-disposition=attachment%3B%20filename%3Dssnpipe_unix.tar.gz&response-content-type=applicat
ion%2Foctet-stream [following]
--2023-09-26 13:08:26--  https://objects.githubusercontent.com/github-production-release-asset-2e65be/186866644/cfbdaa80-8d29-11e9-977d-
70d753487a63?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWNJYAX4CSVEH53A%2F20230926%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Dat
e=20230926T180826Z&X-Amz-Expires=300&X-Amz-Signature=1c3cb7dcfe6a65350b2a4c2663b1254cb302cf5b24318e99e349930391bb3f68&X-Amz-SignedHeader
s=host&actor_id=0&key_id=0&repo_id=186866644&response-content-disposition=attachment%3B%20filename%3Dssnpipe_unix.tar.gz&response-conten
t-type=application%2Foctet-stream
Resolving objects.githubusercontent.com (objects.githubusercontent.com)... 185.199.111.133, 185.199.110.133, 185.199.108.133, ...
Connecting to objects.githubusercontent.com (objects.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 58402420 (56M) [application/octet-stream]
Saving to: ‘ssnpipe_unix.tar.gz’

ssnpipe_unix.tar.gz            100%[==========================================================>]  55.70M   136MB/s    in 0.4s

2023-09-26 13:08:26 (136 MB/s) - ‘ssnpipe_unix.tar.gz’ saved [58402420/58402420]

[sivanr@c1813 ~]$ tar -xvf ssnpipe_unix.tar.gz
ssnpipe_unix/
ssnpipe_unix/base_library.zip
ssnpipe_unix/Bio/
ssnpipe_unix/Bio/Align/
ssnpipe_unix/Bio/Align/_aligners.cpython-35m-x86_64-linux-gnu.so
ssnpipe_unix/Bio/PDB/
ssnpipe_unix/Bio/PDB/kdtrees.cpython-35m-x86_64-linux-gnu.so
```

**Move to ssnpipe_unix folder in the explorer**



**Then open the application in GUI.**



About: SSNpipe v.1.0.0-beta

SSNpipe
v.1.0.0-beta

SSNpipe: a GUI to simplify the generation of Sequence Similarity Networks

If you use SSNpipe, please cite: Viberg & Brunner, to be published http://doi.org

For background information on SSN analysis see: Atkinson, Morris, Ferrin & Babbitt (2009) PLOS ONE, 4, e4345 http://doi.org/10.1371/journal.pone.0004345

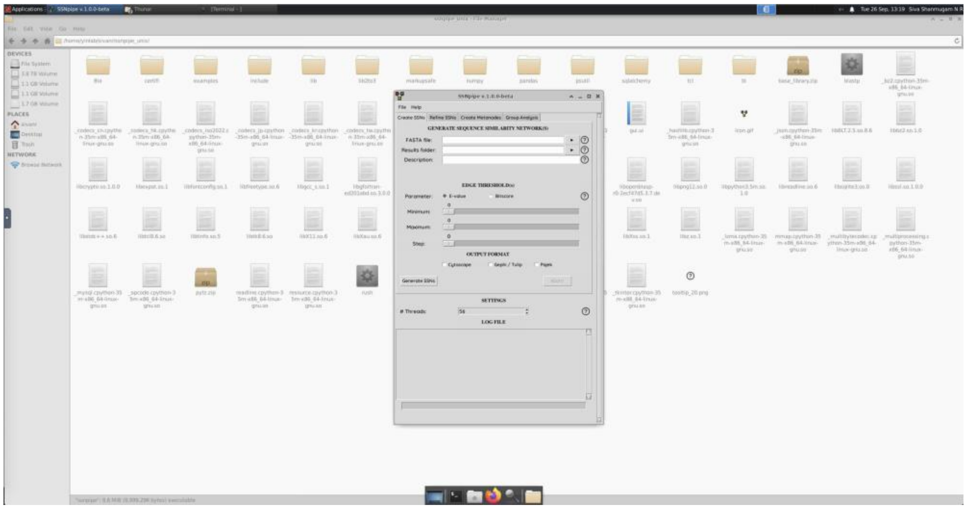
Source code and releases: https://github.com/ahvdk/SSNpipe
Documentation and user guide: https://github.com/ahvdk/SSNpipe/wiki

This program comes with ABSOLUTELY NO WARRANTY for details, visit: http://www.gnu.org/licenses/gpl-3.0.html

SSNpipe includes the following dependencies in this distribution:

Python Tk module: http://tcl.tk
BLASTP: https://blast.ncbi.nlm.nih.gov
rush: https://github.com/shenwei356/rush

**You can find the application opened in the top left corner**



GENERATE SEQUENCE SIMILARITY NETWORK(S)

FASTA file:
Results folder:
Description:

EDGE THRESHOLD(s)

Parameter: ● E-value ○ Bitscore
Minimum: 0
Maximum: 0
Step: 0

OUTPUT FORMAT

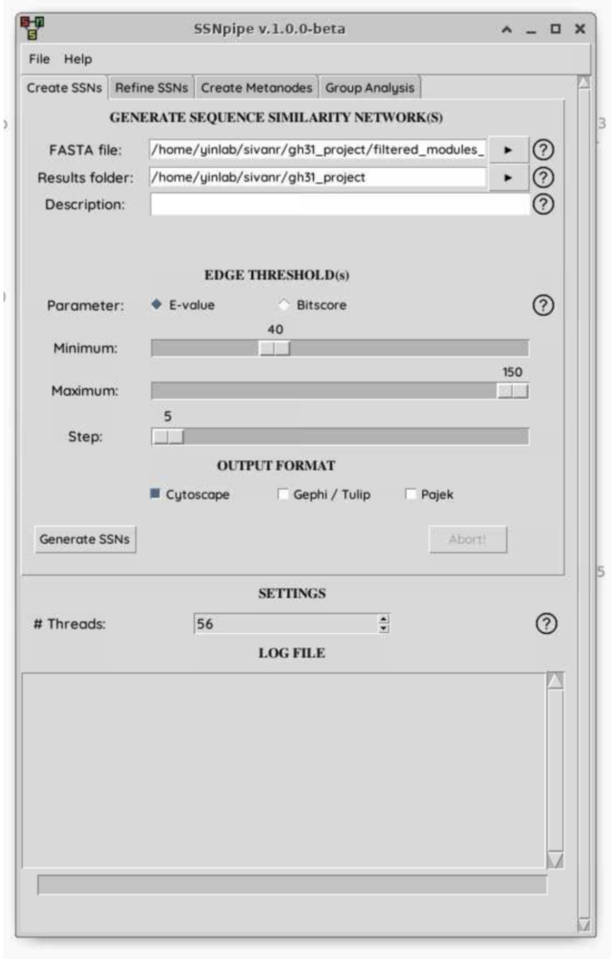
□ Cytoscape □ Gephi / Tulip □ Pajek

SETTINGS

# Threads: 56

LOG FILE

Set the input file, output file path, parameters below with the e-value step of 5 from minimum to maximum.



FASTA file: /home/yinlab/sivanr/gh31_project/filtered_modules_

Results folder: /home/yinlab/sivanr/gh31_project

Description:

Parameter: ● E-value ○ Bitscore

Minimum: 40

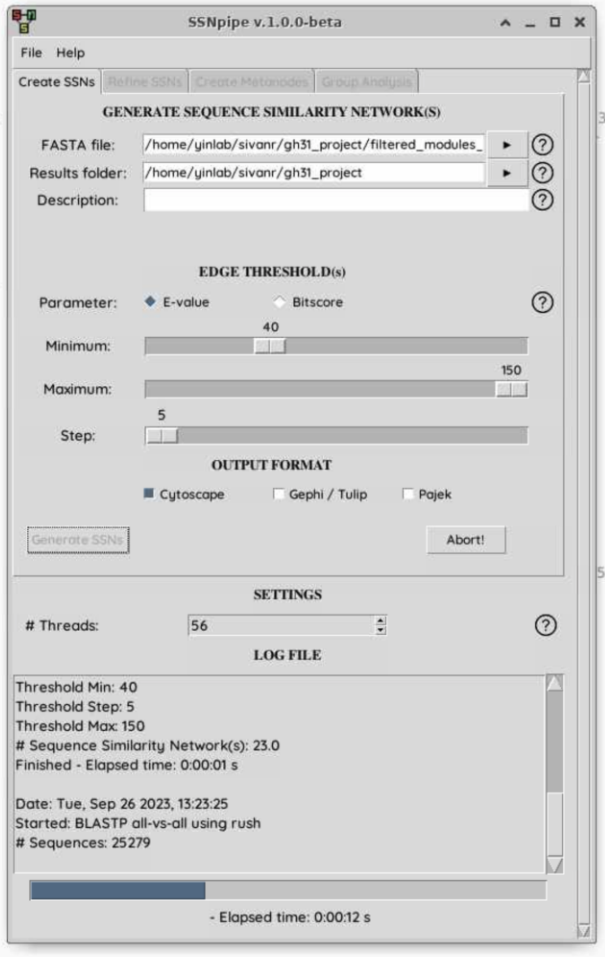
Maximum: 150

Step: 5

☒ Cytoscape □ Gephi / Tulip □ Pajek

# Threads: 56

Click on Generate SSNs, the program will run for few hours depending on the database size and computational power.



GENERATE SEQUENCE SIMILARITY NETWORK(S)

FASTA file: /home/yinlab/sivanr/gh31_project/filtered_modules_

Results folder: /home/yinlab/sivanr/gh31_project

Description:

EDGE THRESHOLD(s)

Parameter: ● E-value ○ Bitscore

Minimum: 40

Maximum: 150

Step: 5

OUTPUT FORMAT

■ Cytoscape □ Gephi / Tulip □ Pajek

SETTINGS

# Threads: 56

LOG FILE

```
Threshold Min: 40
Threshold Step: 5
Threshold Max: 150
# Sequence Similarity Network(s): 23.0
Finished - Elapsed time: 0:00:01 s

Date: Tue, Sep 26 2023, 13:23:25
Started: BLASTP all-vs-all using rush
# Sequences: 25279
```

- Elapsed time: 0:00:12 s

Note:

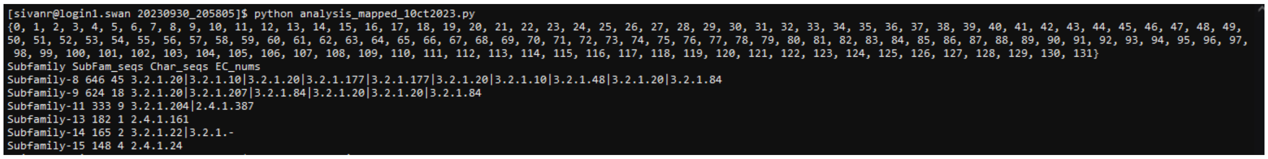
Depending on the network files size you may need to create a metanodes (3rd tab in the SSNpipe interface)

Generate group Analysis files from the 4th tab.

Results can be processed with the following python program and EC-Numbers can be mapped.

**Program name: analysis_mapped_1oct2023.py**

Example for the result of “E-value -140”.



```
[sivanr@login1.swan 20230930_205805]$ python analysis_mapped_10ct2023.py
{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49,
50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131}
Subfamily SubFam_seqs Char_seqs EC_nums
Subfamily-8 646 45 3.2.1.20|3.2.1.10|3.2.1.20|3.2.1.177|3.2.1.177|3.2.1.20|3.2.1.10|3.2.1.48|3.2.1.20|3.2.1.84
Subfamily-9 624 18 3.2.1.20|3.2.1.207|3.2.1.84|3.2.1.20|3.2.1.20|3.2.1.84
Subfamily-11 333 9 3.2.1.204|2.4.1.387
Subfamily-13 182 1 2.4.1.161
Subfamily-14 165 2 3.2.1.22|3.2.1.-
Subfamily-15 148 4 2.4.1.24
```

Use only one step for the project. Select based on your family. GH31 uses optimized subfamily clusters at E-value of 1E-115.

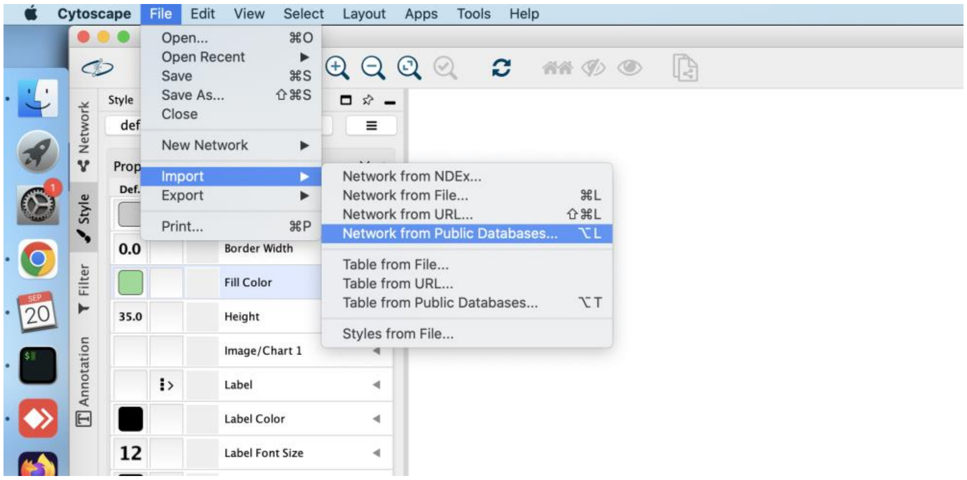
**5. Visualize the network using the Cytoscape and interpret the network. Create images.**

More details on visualization and analysis are available at: https://cytoscape.org/

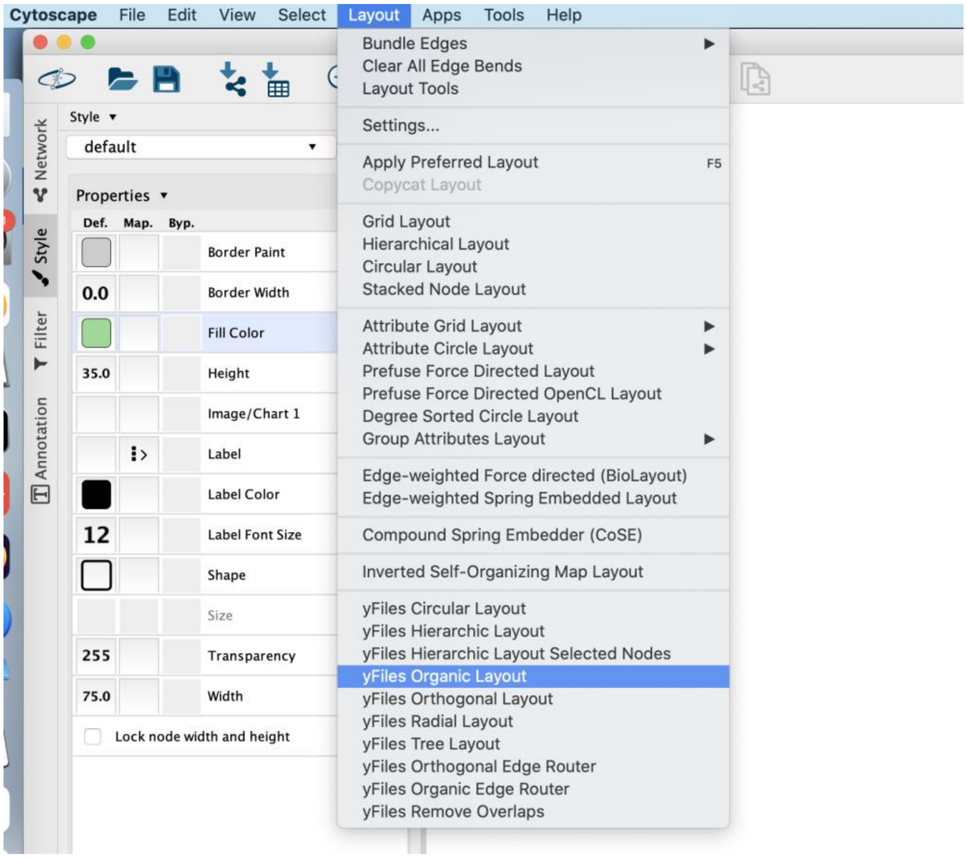
Cytoscape can be accessed by HCC Virtual Desktop in the terminal with the following commands.

- Open terminal in the virtual desktop
- Type “ml cytoscape/3.10”
- After loaded type “cytoscape.sh”
- Import file using import option (Ctrl + L)
- Use organic layout for visualization
- Analyze the results of the E-value.

Import the saved network file into the Cytoscape using the



- Use network from File and load them
- Network will be loaded.
- Use a *yFiles Organic layout* from the layout panel. (Install if not available)

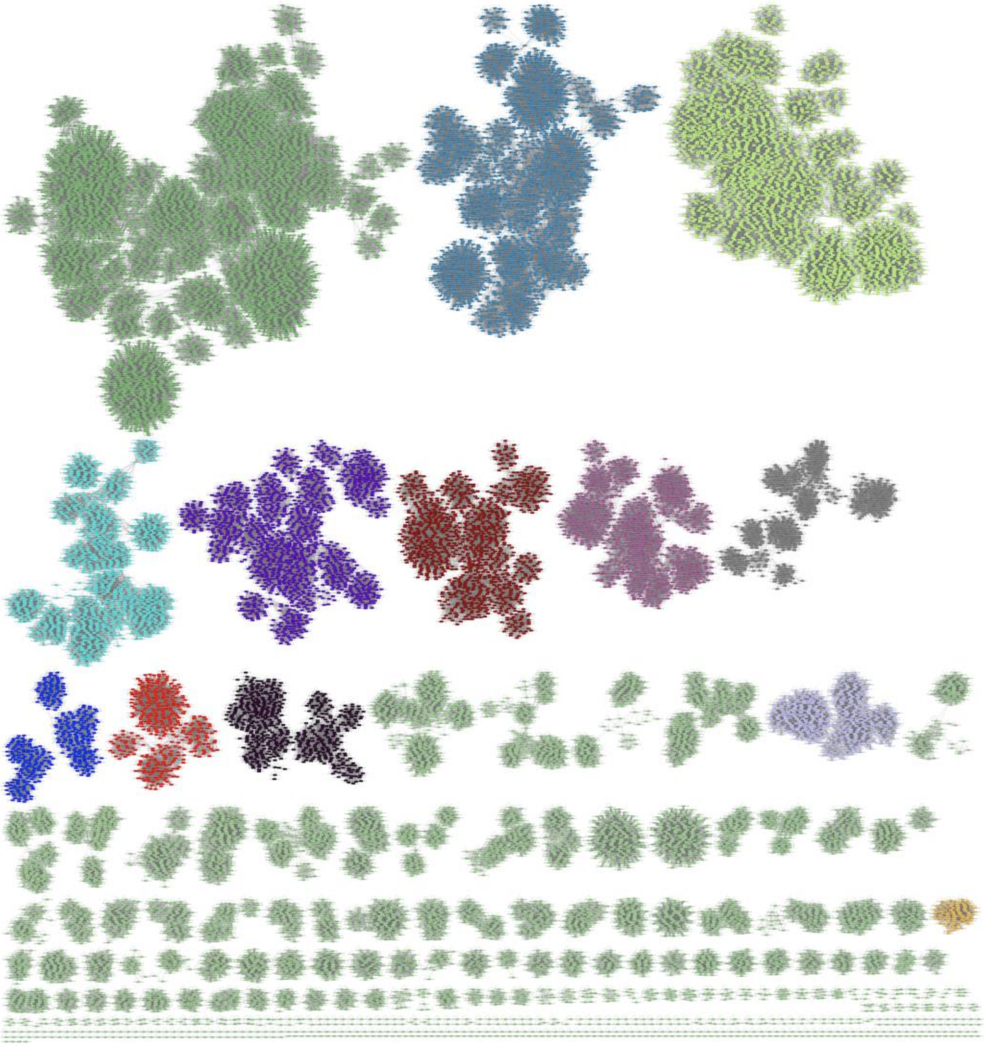


| Def. | Map. | Byp. | |
|---|---|---|---|
| ⬜ | | | Border Paint |
| 0.0 | | | Border Width |
| 🟩 | | | Fill Color |
| 35.0 | | | Height |
| | | | Image/Chart 1 |
| | ⁝> | | Label |
| ⬛ | | | Label Color |
| 12 | | | Label Font Size |
| □ | | | Shape |
| | | | Size |
| 255 | | | Transparency |
| 75.0 | | | Width |

□ Lock node width and height

- You may notice a change in a layout after the Algorithm is selected.
- Check different nodes and subfamilies, how they are aligned, etc.
- Group them based on the connected nodes and color them for easy visualization.

For example, I have used the network results obtained from the blast results with the cutoff of E-value 1e-140 and their sequence similarity networks (SSN) “NETWORK_cs_ev_140_.txt”



Modified colors based on the connected nodes and subfamilies. (This can be done by selecting the connected nodes in each cluster.)

### 6. Phylogenentic Analysis.

Analyse the results using data obtained from the subfamily random 30 seqs in subfamily where seqs are more than 30 and rest are taken as such (>20 anyways).

**Refer to python program (**phylo_process_random_1oct2023.py**)**

Use python program to select random 30 sequences each and use them for phylogeny buliding using RAxML.

## HCC module load MAFFT

```
$ ml mafft/7.520
```

```
$ mafft --localpair --maxiterate 1000 --genafpair --thread 10
phylo_input_fasttree_140.fasta > phylo_mafft_align_1e-140.fasta
```

## HCC module load FASTTREE

```
$ ml fasttree/2.1
```

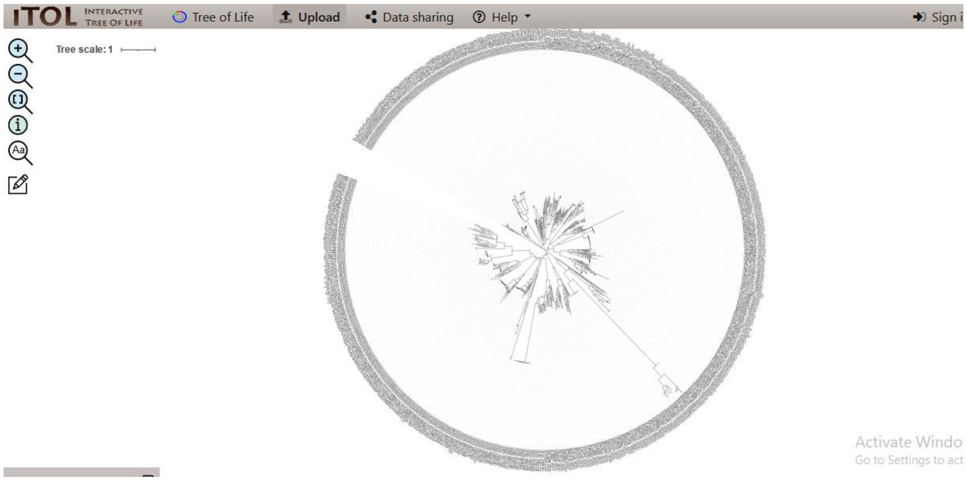
```
$ fasttree -wag -boot 100 -out tree_e-140.nwk
phylo_mafft_align_1e-140.fasta
```

“tree_e-140.nwk” can be visualized using the iTOL available online.

[https://itol.embl.de/](https://itol.embl.de/)

Explore the options over the webserver and save the image for the best SSN Evalue

Example for the SSN 1E-140 build tree



Also available at: [https://itol.embl.de/tree/13423816579310521696229988](https://itol.embl.de/tree/13423816579310521696229988)

## 7. Interpretation and analysis of the subfamilies based on the Phylogeny and subfamilies from the selected E-value.

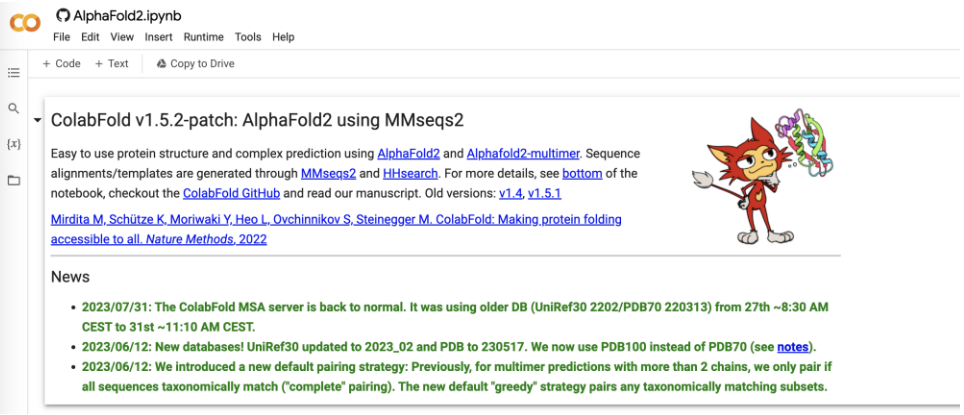
**Interpretation steps and protocol:**

(i) Explain each subfamily in terms of EC code, Taxonomical diversity, enzyme activity and structures.

(ii) Select any one characterized ID and their EC numbers from each subfamily from the results (Output of "analysis_mapped_1oct2023.py")

(iii) If PDB ID is available take the structure, or else model them using AlphaFold2.

a. Use Google Colab code for protein structure modelling, https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb



ColabFold v1.5.2-patch: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.4, v1.5.1

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

News

- 2023/07/31: The ColabFold MSA server is back to normal. It was using older DB (UniRef30 2202/PDB70 220313) from 27th ~8:30 AM CEST to 31st ~11:10 AM CEST.
- 2023/06/12: New databases! UniRef30 updated to 2023_02 and PDB to 230517. We now use PDB100 instead of PDB70 (see notes).
- 2023/06/12: We introduced a new default pairing strategy: Previously, for multimer predictions with more than 2 chains, we only pair if all sequences taxonomically match ("complete" pairing). The new default "greedy" strategy pairs any taxonomically matching subsets.

b. https://blast.ncbi.nlm.nih.gov/Blast.cgi (helpful to identify the similar sequences/ templates for model building)

(iv) Identify the common catalytic domain between each subfamily.

(v) Use PyMol to load and visualize the structures. (Available in HCC virtual Desktop)