Identify co-occurring CAZy families

Author: Emma EM Hobbs

This notebook identifies CAZy families that appear in the same species together, the frequency the CAZy families appear together. Presuming that there is an evolutionary advantage to both families being presing the the CAZome (hence genertng a high frequency of co-occurring), therefore, this approach could be used to identify CAZy families that may be frequently present in the same CAZyme together.

Imports

```
import pandas as pd

from collections import Counter
from copy import copy

from tqdm.notebook import tqdm
```

Load data

Load in the data contained the CSV file `genbank_acc_multi_fams.csv`. The CSV was generated by querying the local CAZyme database to retrieve the NCBI (GenBank) protein version accessions from the local CAZyme database, and their associated CAZy family annotations, for proteins associated with more than one family in the local CAZyme database (the CSV file was generated using the bash script `get_gbk_multi_fams.csv`.

Owing to the size of the file (which contains 398,182 lines), `get_gbk_multi_fams.csv` is read in chunks and parsed into a dictionary. Each row in the CSV file contains a unique family-protein accession pair, therefore, a protein accession will be present in multiple lines. Each key in the dictionary was a NCBI protein version accession, which was paired with the set of CAZy family annotations for the protein and that were retrieved from local CAZyme database.

```
# load in df with each pair of genbank accessions and cazy family
annotation on a unique row
gbk_fam_dict = {}  # {gbk_protein_acc: {families}}

all_families = set()

print('Parsing df')
for chunk in tqdm(

pd.read_csv("../Data/cooccuring_families/genbank_acc_multi_fams.csv",
chunksize=1),
    desc="Parsing acc and fams",
    total=398183,
):
```

```
    # each chunk is one row

    acc = list(chunk['genbank_accession'])[0]  # protein version
accession
    fam = list(chunk['family'])[0]

    # check if protein is already listed in the dictionary
    try:
        gbk_fam_dict[acc].add(fam)  # if in dict, add the CAZy family
to the protein

    except KeyError:  # protein not listed in dict, so add protein
        gbk_fam_dict[acc] = {fam}

    all_families.add(fam)

print(
    f'Parsed df, which included {len(list(gbk_fam_dict.keys()))} '
    f'protein version accessions and {len(all_families)} CAZy
families')
```

```
Parsing df
```


```
{"model_id":"b110e12056b9444fbb8ef423d93bd17e","version_major":2,"vers
ion_minor":0}
```


```
Parsed df, which included 191838 protein version accessions and 360
CAZy families
```


```
{"model_id":"946dbbe50eb548919f67158fc0ac4e31","version_major":2,"vers
ion_minor":0}
```


```
# convert sets to lists and order to faciliate comparison across the
proteins
for acc in tqdm(gbk_fam_dict, desc='Formating dictionary'):
    families = list(gbk_fam_dict[acc])
    families.sort()

    gbk_fam_dict[acc] = str(families)
```


```
{"model_id":"f9ff55e19ebc40f3a7ab8ecc94c71070","version_major":2,"vers
ion_minor":0}
```


Calculate incidence

Use the `Counter` object from the Python package `collections` to count the frequency that each group of CAZy families is listed in `gbk_fam_dict`. This will calculate how many proteins contain each group of CAZy families, and thus can be used to identify groups of CAZy families that frequently appear in the same CAZyme.

```
counter = Counter(gbk_fam_dict.values())
counter
```

```
Counter({"['CBM48', 'GH13']": 45737,
         "['CBM1', 'GH7']": 252,
         "['CBM20', 'GH13']": 1129,
         "['PL5', 'PL7']": 3,
         "['CBM59', 'GH26']": 21,
         "['CBM3', 'GH10', 'GH5']": 6,
         "['AA5', 'CBM32']": 146,
         "['GH13', 'GH133']": 409,
         "['CBM18', 'GH19']": 1170,
         "['CBM13', 'GH10']": 298,
         "['CBM34', 'GH13']": 8244,
         "['CBM50', 'GH25']": 1307,
         "['CBM3', 'GH9']": 422,
         "['CBM2', 'GH11']": 224,
         "['CBM22', 'CBM9', 'GH10']": 183,
         "['CBM5', 'GH5']": 113,
         "['CBM3', 'GH44']": 5,
         "['CBM17', 'CBM28', 'GH5']": 54,
         "['CBM3', 'GH5']": 855,
         "['CBM46', 'GH5']": 210,
         "['GH26', 'GH5']": 15,
         "['CBM26', 'GH13']": 630,
         "['CBM2', 'GH6']": 702,
         "['CBM2', 'CBM3', 'GH9']": 125,
         "['CBM4', 'GH9']": 354,
         "['CBM2', 'GH5']": 1472,
         "['CBM20', 'CBM34', 'GH13']": 72,
         "['CBM20', 'GH14']": 153,
         "['CE2', 'GH5']": 9,
         "['CBM11', 'GH26', 'GH5']": 7,
         "['CBM32', 'GH84']": 541,
         "['CBM6', 'CE0', 'GH10']": 7,
         "['CBM50', 'GH73']": 2589,
         "['CBM41', 'CBM48', 'GH13']": 4194,
         "['CBM13', 'GH64']": 74,
         "['GH94', 'GT84']": 2266,
         "['CBM10', 'GH5']": 106,
         "['CBM2', 'GH18']": 1651,
         "['CBM66', 'GH32']": 501,
         "['CBM0', 'CBM40', 'GH33']": 101,
         "['CBM14', 'GH18']": 670,
         "['CBM13', 'GT27']": 1148,
         "['CBM50', 'GH24']": 66,
         "['CBM20', 'GH15']": 359,
         "['CBM1', 'GH5']": 208,
         "['CBM1', 'GH6']": 143,
         "['CBM19', 'GH18']": 174,
```

```
"['GT31', 'GT7']": 270,
"['CBM8', 'GH9']": 7,
"['AA9', 'CBM1']": 244,
"['CBM70', 'PL8']": 482,
"['CBM2', 'GH10']": 457,
"['CBM5', 'GH18']": 5750,
"['CBM91', 'GH43']": 9730,
"['CBM25', 'GH13']": 234,
"['CBM1', 'GH10']": 161,
"['CBM1', 'GH45']": 121,
"['CBM38', 'GH32']": 389,
"['CBM43', 'GH72']": 644,
"['CBM3', 'GH44', 'GH5']": 6,
"['CBM42', 'GH54']": 221,
"['CBM12', 'GH18']": 1299,
"['CBM12', 'CBM5', 'GH18']": 141,
"['CBM22', 'CE4', 'GH11']": 5,
"['CBM25', 'GH13', 'GH14']": 48,
"['CBM43', 'GH17']": 1267,
"['CBM3', 'GH48', 'GH9']": 18,
"['CBM76', 'GH44']": 5,
"['CBM2', 'GH48']": 358,
"['GT2', 'GT4']": 4244,
"['GT47', 'GT64']": 325,
"['CBM50', 'GH23']": 14178,
"['CBM18', 'CBM50', 'GH18']": 145,
"['CBM22', 'GH11', 'GH16']": 2,
"['CBM6', 'GH3']": 413,
"['CBM17', 'GH5']": 6,
"['CBM39', 'GH16']": 223,
"['CBM22', 'GH10']": 446,
"['CBM58', 'GH13']": 85,
"['CBM30', 'GH9']": 13,
"['CBM55', 'GH18']": 6,
"['CBM6', 'GH5']": 351,
"['CBM35', 'GH30']": 23,
"['CBM18', 'GH16']": 444,
"['CBM2', 'GH12']": 220,
"['CBM35', 'GH26']": 195,
"['GT1', 'GT2']": 49,
"['CBM22', 'CBM6', 'CBM91', 'GH43']": 11,
"['CBM22', 'CBM3', 'CBM6', 'GH10', 'GH43']": 2,
"['CBM36', 'GH11']": 55,
"['GT2', 'GT4', 'GT99']": 56,
"['CBM48', 'CBM68', 'GH13']": 999,
"['CBM6', 'GH11']": 20,
"['CBM6', 'CE4', 'GH11']": 13,
"['CBM2', 'CE4']": 136,
"['CBM2', 'CBM4', 'GH9']": 127,
"['CBM53', 'GT5']": 155,
```

```
"['AA3', 'AA8', 'CBM1']": 47,
"['CBM1', 'GH3']": 15,
"['CBM13', 'GH62']": 325,
"['CBM16', 'GH16']": 76,
"['GH13', 'GT5']": 150,
"['AA3', 'AA8']": 93,
"['CBM49', 'GH9']": 184,
"['CBM72', 'GH5']": 26,
"['CBM13', 'GH16']": 504,
"['CBM27', 'CBM35', 'GH26']": 8,
"['CBM40', 'GH33']": 567,
"['CBM11', 'CBM30', 'GH51']": 6,
"['AA10', 'CBM2']": 471,
"['CBM5', 'CBM73', 'GH18']": 775,
"['CBM32', 'GH87']": 675,
"['CBM21', 'GH13']": 7,
"['GH17', 'GT2']": 1412,
"['CBM27', 'GH5']": 46,
"['CE4', 'GH153']": 4586,
"['AA10', 'CBM5']": 1655,
"['AA15', 'CBM14']": 45,
"['GT111', 'GT8']": 824,
"['CBM16', 'GH5']": 8,
"['CBM16', 'PL18']": 19,
"['CBM13', 'GH10', 'GH62']": 90,
"['CBM32', 'GH65']": 176,
"['CBM4', 'GH16']": 151,
"['CBM61', 'GH53']": 574,
"['CBM65', 'GH5']": 21,
"['CBM23', 'GH26']": 54,
"['CBM22', 'CBM5', 'CBM9', 'GH10']": 3,
"['CBM24', 'GH71']": 90,
"['GT2', 'GT8']": 210,
"['CBM13', 'PL1']": 113,
"['CBM1', 'GH11']": 63,
"['CBM13', 'GH27']": 333,
"['CBM45', 'GH13']": 97,
"['AA10', 'CBM73']": 1794,
"['CBM1', 'PL1']": 26,
"['AA10', 'CBM3', 'GH5']": 1,
"['CBM6', 'GH96']": 10,
"['CBM6', 'GH16']": 468,
"['CBM94', 'GT54']": 271,
"['CBM22', 'CBM3', 'GH10']": 20,
"['GT2', 'GT74']": 8,
"['CBM63', 'GH5']": 267,
"['CE8', 'PL1']": 298,
"['CE4', 'GT4']": 425,
"['CBM5', 'GH19']": 950,
"['CBM12', 'CE4']": 277,
```

```
"['CBM71', 'GH2']": 228,
"['CBM6', 'GH10']": 31,
"['CBM0', 'CBM22']": 287,
"['CBM2', 'CBM35', 'PL10']": 13,
"['CBM18', 'GH18']": 249,
"['CBM6', 'CE6']": 10,
"['CBM20', 'CBM34', 'CBM48', 'GH13']": 7,
"['CBM0', 'GT54']": 18,
"['GT49', 'GT8']": 165,
"['GT2', 'GT45']": 109,
"['CBM3', 'GH74']": 89,
"['CBM3', 'GH48']": 97,
"['CBM3', 'GH44', 'GH9']": 2,
"['AA15', 'CBM1']": 33,
"['CBM2', 'CE4', 'GH11']": 13,
"['CBM2', 'CBM63', 'GH5']": 16,
"['GT101', 'GT8']": 127,
"['CBM13', 'GH26']": 34,
"['CBM2', 'PL11']": 89,
"['CBM6', 'GH86']": 53,
"['CBM25', 'GH14']": 4,
"['CBM32', 'GH85']": 713,
"['CBM47', 'GH98']": 76,
"['CBM13', 'CE0']": 176,
"['CBM6', 'GH43']": 1255,
"['CBM13', 'GH30']": 557,
"['CBM13', 'CBM6', 'GH43']": 34,
"['CBM1', 'GH74']": 35,
"['CBM13', 'GH43']": 773,
"['CBM13', 'GH5']": 133,
"['CBM38', 'CBM66', 'GH32']": 70,
"['CBM26', 'GH31']": 14,
"['GT0', 'GT2']": 1305,
"['CBM50', 'GH18']": 2559,
"['CBM18', 'CE4']": 145,
"['CBM32', 'GH123', 'GH33']": 11,
"['CBM32', 'GH101']": 384,
"['CBM73', 'GH19']": 1032,
"['CBM60', 'GH11']": 60,
"['CBM1', 'GH18']": 78,
"['CBM34', 'CBM48', 'GH13']": 40,
"['CBM20', 'GH77']": 637,
"['CBM92', 'GH5']": 47,
"['CBM43', 'GH5']": 18,
"['CBM92', 'GH30']": 27,
"['CBM35', 'GH98']": 310,
"['GH5', 'PL31']": 12,
"['CBM1', 'CE1']": 49,
"['GT102', 'GT103', 'GT99']": 62,
"['CBM22', 'CBM6', 'GH43']": 102,
```

```
"['CBM22', 'CBM91', 'GH43']": 108,
"['CBM59', 'GH5']": 51,
"['CE4', 'GH11']": 3,
"['CBM21', 'GH15']": 55,
"['CBM69', 'GH13']": 56,
"['CBM73', 'GH18']": 310,
"['CBM5', 'CBM73']": 191,
"['CBM13', 'CBM66', 'GH32']": 32,
"['CBM10', 'CBM2', 'GH5']": 46,
"['CBM10', 'CBM35', 'CBM5', 'GH5']": 7,
"['CBM10', 'GH26']": 20,
"['CBM6', 'GH28']": 17,
"['CBM32', 'GH35']": 435,
"['CBM93', 'GH33']": 265,
"['CBM0', 'CBM57', 'GH137', 'GH2']": 104,
"['GH33', 'GH78']": 120,
"['GH142', 'GH143']": 115,
"['CBM32', 'GH2']": 622,
"['CBM0', 'GH20']": 20,
"['CBM35', 'GH27']": 478,
"['GH106', 'GH43']": 29,
"['GH105', 'PL33']": 41,
"['CBM32', 'GH31']": 637,
"['GH43']": 478,
"['CBM51', 'GH27']": 484,
"['CBM32', 'GH43']": 368,
"['GH16', 'GH43']": 86,
"['CBM32', 'GH38']": 36,
"['CE12', 'CE8']": 187,
"['GH42', 'GH43']": 35,
"['GH18', 'GH20']": 82,
"['CBM32', 'PL8']": 110,
"['GT4', 'GT97']": 102,
"['CBM16', 'GH111']": 1,
"['CE4', 'GH18', 'GT2']": 716,
"['CBM10', 'CBM2', 'GH45']": 10,
"['CBM1', 'CE15']": 16,
"['PL1']": 5,
"['CBM12', 'CBM6', 'GH18']": 3,
"['CBM0', 'GH26']": 2,
"['AA10', 'CBM12']": 341,
"['CBM5', 'GH46']": 29,
"['CBM32', 'GH33']": 18,
"['CBM61', 'GH66']": 11,
"['CBM37', 'CBM4', 'GH9']": 7,
"['CBM37', 'GH48']": 22,
"['GT0', 'GT2', 'GT4']": 19,
"['CBM22', 'CE0', 'GH10']": 8,
"['CBM22', 'CE0', 'GH11']": 3,
"['CBM51', 'GH98']": 97,
```

```
"['GH6', 'GT2']": 11,
"['CBM67', 'GH78']": 815,
"['GH10', 'GH11']": 3,
"['CBM41', 'GH13']": 118,
"['CBM25', 'CBM26', 'GH13']": 39,
"['CBM35', 'CBM61', 'GH31']": 136,
"['CBM35', 'GH31']": 230,
"['CBM37', 'GH5']": 4,
"['CBM22', 'CBM37', 'GH11']": 3,
"['GH20', 'GH35']": 8,
"['AA8', 'CBM1']": 3,
"['CBM6', 'CBM92', 'GH3']": 69,
"['AA5', 'CBM13']": 238,
"['CBM13', 'CBM35', 'GH97']": 3,
"['GT10', 'GT25']": 36,
"['CBM1', 'PL3']": 13,
"['CBM42', 'GH43']": 462,
"['CBM3', 'CBM35', 'GH26', 'GH44']": 46,
"['CBM0', 'GH33']": 5,
"['CE4', 'GT2']": 654,
"['CBM47', 'GT0']": 9,
"['CBM5', 'CBM73', 'GH19']": 157,
"['CBM13', 'GH19']": 76,
"['CBM4', 'GH6']": 47,
"['CBM61', 'GH16']": 45,
"['CBM35', 'GH2']": 45,
"['CBM25', 'CBM41', 'CBM48', 'GH13']": 62,
"['CBM2', 'GH74']": 242,
"['CBM2', 'CE3']": 137,
"['GT101', 'GT2']": 42,
"['GH13', 'GH77']": 205,
"['CBM56', 'CBM6', 'GH81']": 14,
"['GT2', 'GT41']": 48,
"['AA5', 'CE3']": 22,
"['GH0', 'GH13']": 776,
"['GH26', 'GT2']": 137,
"['GH16', 'GH20']": 43,
"['GT0', 'GT4']": 223,
"['CBM72', 'GH26', 'GH5']": 2,
"['CBM32', 'CBM4', 'GH16']": 12,
"['GH1', 'GT4']": 154,
"['GT2', 'GT4', 'GT41']": 36,
"['CBM2', 'GH9']": 112,
"['CBM32', 'CBM35']": 8,
"['CBM12', 'CBM32']": 3,
"['CBM6', 'GH18']": 29,
"['CBM12', 'CBM6', 'GH5']": 1,
"['GT26', 'GT4']": 32,
"['CBM13', 'CBM6', 'GH16']": 10,
"['CBM2', 'GH8']": 26,
```

```
"['GT2', 'GT25']": 437,
"['CBM48', 'GH13', 'GH77']": 84,
"['CBM50', 'GH0']": 1914,
"['GT3', 'GT35']": 90,
"['GH57', 'GT5']": 1,
"['GH8', 'GT2']": 6,
"['CBM32', 'PL6']": 43,
"['CBM16', 'CBM6']": 2,
"['CBM5', 'GH10']": 1,
"['CBM2', 'CBM57']": 3,
"['CBM56', 'CBM6']": 2,
"['CBM6', 'PL0']": 3,
"['CBM85', 'GH10']": 55,
"['CBM32', 'GH3']": 64,
"['CBM10', 'CBM2', 'GH9']": 14,
"['CBM10', 'CBM2']": 3,
"['CBM35', 'PL1']": 51,
"['CBM56', 'CBM6', 'GH16']": 9,
"['CBM6', 'GH128']": 22,
"['CBM32', 'PL7']": 145,
"['CBM2', 'CBM35', 'PL11']": 4,
"['CBM13', 'PL3']": 57,
"['CBM2', 'CBM35', 'PL3']": 1,
"['CBM2', 'CBM35', 'PL1']": 1,
"['CBM32', 'CBM6']": 50,
"['CBM13', 'CBM35', 'GH43']": 15,
"['CBM6', 'GH0']": 21,
"['CBM56', 'GH81']": 59,
"['PL6', 'PL7']": 3,
"['CBM32', 'CBM6', 'GH128', 'GH16']": 11,
"['CBM10', 'CBM2', 'GH10']": 7,
"['CBM35', 'PL10']": 6,
"['CBM13', 'CBM6', 'GH30']": 2,
"['CBM13', 'CBM6', 'GH5']": 9,
"['CBM32', 'GH16']": 260,
"['CBM10', 'CBM60', 'CE4', 'GH11']": 16,
"['CBM6', 'CBM81', 'GH5']": 4,
"['CBM6', 'PL31']": 1,
"['CBM16', 'CBM32', 'PL18']": 12,
"['CBM2', 'CBM22', 'CBM6', 'GH10', 'GH43']": 4,
"['CBM0', 'CBM10', 'GH26']": 1,
"['CBM32', 'CBM47']": 2,
"['CBM13', 'GH53']": 104,
"['CBM85', 'GH5']": 6,
"['CBM48', 'CE0']": 99,
"['CBM51', 'CBM57']": 7,
"['CBM8', 'CE0']": 29,
"['CBM50', 'GH19']": 145,
"['GT0', 'GT2', 'GT4', 'GT41']": 1,
"['CBM51', 'CBM6']": 3,
```

```
"['CBM88', 'PL11']": 3,
"['CBM9', 'CE15']": 5,
"['CBM4', 'CE6', 'GH10']": 1,
"['CBM9', 'CE4', 'GH8']": 1,
"['CBM88', 'CBM9', 'CE1']": 1,
"['CBM88', 'CBM9', 'CBM91', 'CE6', 'GH43']": 1,
"['CBM88', 'CBM9', 'GH30']": 1,
"['CBM88', 'CBM9', 'GH10']": 1,
"['CBM6', 'CBM88', 'CBM9', 'CBM91', 'GH43']": 1,
"['CBM4', 'GH10']": 106,
"['CBM9', 'GH11']": 1,
"['CBM9', 'GH8']": 1,
"['CBM22', 'GH146']": 3,
"['CBM6', 'GH27']": 7,
"['CBM6', 'GH59']": 3,
"['CBM32', 'CBM51']": 165,
"['CBM51', 'GH95']": 87,
"['CBM32', 'GH36']": 45,
"['CBM32', 'GH20']": 342,
"['CBM32', 'GH89']": 93,
"['CBM32', 'CBM40', 'GH33']": 74,
"['CBM51', 'GH2']": 61,
"['CBM2', 'CBM3', 'GH12', 'GH6']": 3,
"['GH130', 'GT81']": 14,
"['CBM32', 'CBM4', 'CBM54', 'GH16']": 4,
"['CBM2', 'CBM3', 'GH5']": 2,
"['CE12', 'PL10']": 2,
"['CBM85', 'GH10', 'GH16']": 1,
"['GT2', 'GT26']": 68,
"['CBM2', 'CBM3', 'CE1']": 1,
"['CBM2', 'CBM3', 'GH10']": 1,
"['CBM2', 'CBM3', 'GH48']": 2,
"['CBM2', 'CBM3', 'GH74']": 1,
"['CBM16', 'GH18']": 872,
"['CBM16', 'CBM5', 'GH18']": 2,
"['CBM51', 'GH31']": 105,
"['CBM20', 'GH31']": 38,
"['GT102', 'GT103']": 52,
"['CBM51', 'GH101']": 94,
"['CBM2', 'GH18', 'GH5']": 15,
"['CBM2', 'CBM63']": 23,
"['CBM35', 'PL11']": 23,
"['CBM3', 'CBM4', 'GH9']": 11,
"['CBM30', 'CBM44', 'GH44', 'GH9']": 6,
"['CBM32', 'GH5']": 47,
"['CBM3', 'CBM4']": 5,
"['CBM35', 'GH39']": 237,
"['CBM42', 'GH30', 'GH43']": 5,
"['CBM35', 'PL1', 'PL9']": 4,
"['CBM13', 'CBM6', 'CBM62', 'GH5']": 6,
```

```
"['CBM6', 'CE0']": 17,
"['CBM6', 'GH141']": 16,
"['CBM6', 'GH2']": 32,
"['CBM4', 'CBM54', 'GH16']": 20,
"['CBM6', 'GH30']": 31,
"['CBM0', 'GT39']": 27,
"['CBM35', 'CE12']": 19,
"['CBM2', 'PL14']": 3,
"['CBM11', 'GH3']": 140,
"['CBM32', 'GH92']": 88,
"['CBM12', 'CBM5', 'GH19']": 60,
"['CBM13', 'GH12', 'GH5']": 1,
"['CBM28', 'GH5']": 15,
"['CBM20', 'CBM41', 'CBM48', 'GH13']": 14,
"['CBM32', 'GH173']": 6,
"['CBM4', 'CBM6', 'GH16']": 10,
"['CBM0', 'CBM13', 'GH43']": 3,
"['CBM77', 'PL1']": 85,
"['CE0', 'CE6']": 3,
"['CBM0', 'CBM57', 'GH2']": 65,
"['CBM13', 'CBM6']": 228,
"['CBM13', 'CBM6', 'GH64']": 16,
"['CE8', 'PL10']": 45,
"['GT116', 'GT4']": 2,
"['CBM48', 'GH57']": 3,
"['CBM32', 'GH55']": 317,
"['CBM56', 'GH64']": 59,
"['CBM12', 'GH16']": 14,
"['CBM5', 'GH23']": 8,
"['CBM48', 'CBM6', 'CE0', 'GH43']": 9,
"['GH106', 'GH28']": 26,
"['GH105', 'GH154']": 33,
"['CE0', 'PL10']": 41,
"['GH28', 'GH43']": 18,
"['GT0', 'GT32']": 70,
"['CBM35', 'GH5']": 1049,
"['CBM57', 'GH26']": 5,
"['CBM35', 'GH66']": 29,
"['CBM35', 'GH15']": 3,
"['CBM10', 'CBM5', 'GH5', 'GH6']": 4,
"['CBM41', 'CBM48', 'CBM69', 'GH13']": 24,
"['CBM32', 'GH64']": 59,
"['CBM13', 'GH12']": 19,
"['CBM6', 'GH45']": 1,
"['CBM4', 'GH5']": 21,
"['CBM2', 'CBM59', 'GH26']": 1,
"['CBM8', 'GH51']": 6,
"['CBM13', 'GH81']": 1,
"['CBM92', 'GH55']": 6,
"['CBM12', 'CBM32', 'CBM63']": 1,
```

```
"['CBM12', 'GH19']": 246,
"['CBM32', 'GH18']": 6,
"['CBM32', 'CBM92', 'GH5']": 5,
"['CBM12', 'GH8']": 1,
"['CBM2', 'CBM32', 'GH16']": 1,
"['CBM2', 'PL1']": 22,
"['CBM50', 'PL7']": 1,
"['CBM3', 'CBM35', 'GH26']": 32,
"['CBM86', 'GH10']": 3,
"['CBM77', 'PL1', 'PL9']": 3,
"['CBM36', 'CE4']": 52,
"['CBM32', 'CBM54', 'GH55']": 1,
"['CBM35', 'CBM6', 'GH87']": 2,
"['CBM2', 'CBM46', 'GH5']": 58,
"['CE15', 'GH10']": 4,
"['GH13']": 113,
"['GT2', 'GT32']": 40,
"['CBM36', 'CBM6', 'GH43']": 56,
"['GH57', 'GT4']": 82,
"['CBM57', 'GH16']": 14,
"['CBM50', 'CE4']": 76,
"['CBM4', 'GH148']": 25,
"['GH43', 'GH95']": 1,
"['GH105', 'PL42']": 2,
"['CE15', 'GH106']": 1,
"['CBM6', 'GH29']": 3,
"['CBM91', 'GH43', 'GH62']": 1,
"['GH105', 'GH28']": 20,
"['CBM34', 'GH13', 'GH77']": 47,
"['GH10', 'GT4']": 1,
"['GT4', 'GT41']": 18,
"['CBM51', 'GH110']": 122,
"['CBM32', 'GH123']": 98,
"['CBM2', 'GH30']": 52,
"['AA10', 'CBM5', 'CBM73']": 22,
"['CBM10', 'CBM60', 'CE15']": 5,
"['CBM2', 'CBM35', 'GH10']": 5,
"['CBM2', 'CBM35', 'GH98']": 4,
"['CBM10', 'CBM2', 'GH74']": 6,
"['AA10', 'CBM10']": 11,
"['CBM15', 'GH10']": 10,
"['CBM2', 'CBM35', 'CE0']": 6,
"['CBM2', 'CBM35', 'GH62']": 6,
"['CBM6', 'GH19']": 6,
"['CBM10', 'CBM2', 'GH6']": 13,
"['CBM88', 'GH5']": 2,
"['CBM32', 'CBM51', 'GH31']": 129,
"['CBM40', 'GH16', 'GH33']": 48,
"['CBM32', 'GH28']": 6,
"['CBM0', 'GH16']": 7,
```

```
"['CBM0', 'CBM14', 'GH18']": 1,
"['CBM59', 'GH10']": 3,
"['GT111', 'GT2', 'GT8']": 40,
"['CBM66', 'GH101']": 12,
"['CBM6', 'CE3']": 9,
"['CE1', 'GH11']": 1,
"['CBM4', 'GH16', 'GH17']": 1,
"['GT30', 'GT9']": 3,
"['CBM44', 'GH44']": 6,
"['CBM11', 'GH5']": 4,
"['CBM9', 'GH141']": 2,
"['CBM6', 'CBM91', 'GH43']": 67,
"['CBM6', 'GH62']": 3,
"['CBM6', 'CE6', 'GH62']": 3,
"['CBM32', 'CBM6', 'GH95']": 3,
"['CBM9', 'CE1']": 1,
"['CBM35', 'GH43']": 229,
"['CBM9', 'CE0']": 12,
"['CBM3', 'PL11']": 5,
"['CBM66', 'PL3']": 21,
"['CBM66', 'PL9']": 24,
"['CBM3', 'GH10', 'GH48']": 4,
"['CBM3', 'GH48', 'GH74']": 5,
"['CBM3', 'GH5', 'GH9']": 4,
"['CBM66', 'PL0']": 7,
"['CBM10', 'CBM3', 'CBM5', 'GH9']": 2,
"['CBM48', 'CE0', 'GH10']": 31,
"['CE8', 'GH28']": 27,
"['CBM13', 'GH51']": 8,
"['GH26', 'GH8']": 2,
"['CBM57', 'CE0', 'GH10']": 2,
"['GH10', 'GH9']": 4,
"['CE0', 'GH10']": 2,
"['CBM9', 'GH10']": 23,
"['CBM87', 'CE18']": 41,
"['CBM66', 'GH93']": 9,
"['CBM0', 'CBM22', 'CBM9', 'GH10']": 1,
"['CBM5', 'CBM66', 'PL0']": 1,
"['CBM32', 'PL35']": 8,
"['CBM22', 'CBM6', 'GH10']": 8,
"['CBM35', 'CBM5', 'PL1']": 1,
"['CBM13', 'CBM2', 'GH16']": 1,
"['CBM2', 'CBM35', 'CE12']": 2,
"['CBM2', 'CBM35', 'CE8']": 4,
"['CBM10', 'CBM5', 'CBM60', 'CE6', 'GH10']": 1,
"['CBM57', 'CE15']": 3,
"['CBM10', 'CBM88', 'GH53']": 1,
"['CBM10', 'CBM5', 'GH6']": 1,
"['CBM10', 'CBM5', 'GH16']": 1,
"['CBM10', 'CBM5', 'GH11', 'GH5']": 1,
```

```
"['CBM2', 'CBM5', 'GH44']": 1,
"['CBM0', 'CBM10', 'CBM2', 'GH26']": 1,
"['CBM2', 'CBM35']": 1,
"['CBM2', 'CBM6', 'GH62']": 1,
"['CBM10', 'CBM2', 'GH16']": 2,
"['CBM10', 'CBM2', 'CE3']": 2,
"['CBM10', 'CBM5', 'GH5']": 1,
"['CBM10', 'CBM5']": 1,
"['CBM5', 'CBM57', 'CBM60', 'CE15', 'GH11']": 1,
"['CBM10', 'GH30']": 1,
"['CBM13', 'CBM32', 'GH29']": 50,
"['CBM32', 'GH158']": 197,
"['PL1', 'PL9']": 20,
"['CBM82', 'CBM83', 'GH13']": 2,
"['CBM9', 'GH166']": 2,
"['CBM20', 'GH57']": 2,
"['GH77', 'GT4']": 18,
"['GH77', 'GT35']": 18,
"['CBM6', 'GH76']": 18,
"['CBM32', 'CBM35', 'GH87']": 59,
"['CBM23', 'GH5']": 40,
"['CBM61', 'GH43']": 59,
"['CBM4', 'CBM54', 'CBM6', 'GH16']": 24,
"['CBM35', 'GH87']": 45,
"['CBM0', 'CBM66', 'GH136']": 2,
"['CBM35', 'GH36']": 30,
"['GH5', 'GH93']": 2,
"['CBM32', 'GH136']": 37,
"['CBM13', 'GH39']": 104,
"['CBM6', 'CBM61', 'GH30']": 4,
"['CBM32', 'CBM6', 'GH92']": 10,
"['CBM32', 'GH120', 'GH95']": 2,
"['CBM32', 'GH81']": 32,
"['CE12', 'PL11']": 55,
"['GT0', 'GT41']": 4,
"['CE0', 'GH9']": 80,
"['CBM57', 'GH2']": 221,
"['CBM13', 'CBM35', 'GH31']": 12,
"['CBM42', 'GH2']": 174,
"['CBM2', 'GH51']": 107,
"['CBM13', 'GH146']": 108,
"['CBM13', 'CE3']": 103,
"['CBM2', 'GH62']": 64,
"['CBM2', 'CE0']": 95,
"['CBM2', 'CE2']": 30,
"['CBM2', 'CE1']": 216,
"['CBM2', 'CE15']": 48,
"['CBM13', 'CE8']": 23,
"['CBM13', 'GH46']": 17,
"['CBM13', 'PL9']": 27,
```

```
"['CBM13', 'CE8', 'PL1']": 2,
"['CBM35', 'PL9']": 92,
"['GH106', 'GH137']": 2,
"['CBM92', 'GH18', 'GH5']": 1,
"['CBM13', 'GH76']": 30,
"['CBM6', 'GH64']": 54,
"['CBM6', 'CBM92']": 20,
"['CBM32', 'CBM6', 'GH18']": 2,
"['CBM13', 'CE2']": 17,
"['CBM9', 'GH87']": 10,
"['CBM92', 'GH16']": 51,
"['CE12', 'GH28']": 29,
"['CBM13', 'CBM67', 'GH78']": 1,
"['CBM13', 'GH55']": 155,
"['CBM13', 'GH18']": 198,
"['CBM32', 'CBM6', 'GH3']": 6,
"['CBM2', 'GH54']": 20,
"['CBM13', 'GH0']": 49,
"['CBM13', 'CBM16', 'CBM84']": 1,
"['CBM35', 'GH28']": 18,
"['CBM13', 'GH59']": 124,
"['CBM51', 'GH35']": 50,
"['CBM32', 'GH30']": 41,
"['CBM13', 'GH95']": 59,
"['CBM13', 'GH3']": 16,
"['CBM13', 'CBM32', 'GH16']": 3,
"['CBM32', 'GH158', 'GH16']": 30,
"['CBM32', 'CBM6', 'GH87']": 16,
"['CBM13', 'GH54']": 169,
"['CBM13', 'GH141']": 93,
"['CBM13', 'GH29']": 106,
"['CBM13', 'GH11']": 17,
"['CBM32', 'GH29']": 294,
"['CBM13', 'CE1']": 30,
"['CBM92', 'GH54']": 6,
"['CBM32', 'CBM51', 'GH27']": 2,
"['CBM13', 'GH79']": 2,
"['CBM92', 'GH3']": 4,
"['CBM13', 'CBM32', 'GH20']": 8,
"['CBM13', 'GH92']": 25,
"['CBM35', 'GH18']": 8,
"['CBM32', 'CBM56', 'GH55']": 1,
"['CBM32', 'CBM56', 'GH16']": 4,
"['CBM35', 'CBM61', 'GH16']": 1,
"['CBM2', 'GH66']": 1,
"['CBM2', 'CBM35', 'GH27']": 1,
"['CBM32', 'PL31']": 9,
"['CBM3', 'GH0']": 165,
"['CBM6', 'GH55']": 6,
"['CBM13', 'PL7']": 12,
```

```
"['CBM35', 'GH75']": 4,
"['AA10', 'CBM5', 'GH18']": 2,
"['CBM5', 'CE6']": 1,
"['CBM88', 'PL1']": 2,
"['CBM85', 'GH43']": 2,
"['CBM5', 'GH20']": 29,
"['CBM12', 'GH23']": 28,
"['GT116', 'GT2']": 3,
"['CBM32', 'CBM70', 'PL35']": 9,
"['CBM66', 'GH28']": 4,
"['CBM56', 'GH16']": 60,
"['CBM6', 'CBM66', 'GH43']": 31,
"['CBM13', 'CBM42', 'GH93']": 6,
"['CBM6', 'GH54']": 1,
"['CBM35', 'CE8']": 4,
"['CBM6', 'GH39']": 3,
"['GH5', 'GT2']": 2,
"['CBM4', 'CE1']": 4,
"['CBM6', 'GH95']": 4,
"['CBM35', 'CBM61', 'GH43']": 7,
"['CBM35', 'GH53']": 2,
"['GH0', 'GT2']": 13,
"['CBM88', 'GH43']": 2,
"['CBM88', 'GH30']": 2,
"['CBM50', 'GH46']": 2,
"['GH0', 'GH23']": 8,
"['GH130']": 9,
"['GT2', 'GT9']": 25,
"['CBM9', 'GH39']": 14,
"['CBM32', 'GH8']": 30,
"['CBM20', 'CBM25', 'GH13']": 42,
"['CBM5', 'CE4']": 2,
"['CBM32', 'GH99']": 5,
"['CBM47', 'GH29']": 28,
"['CBM57', 'GH18']": 3,
"['CBM2', 'CBM65', 'GH5']": 3,
"['CBM13', 'CBM2', 'GH10']": 5,
"['CBM13', 'CBM2', 'GH30']": 4,
"['CBM13', 'CBM2', 'CE0']": 3,
"['CBM3', 'CBM46', 'GH5']": 1,
"['CBM13', 'CBM2', 'CBM91', 'GH43']": 1,
"['CBM3', 'GH26']": 1,
"['CBM65', 'CE2']": 1,
"['CBM0', 'CBM41', 'CBM48', 'GH13']": 1,
"['GH0', 'GH33']": 5,
"['CBM13', 'GH93']": 130,
"['CBM51', 'GH97']": 40,
"['CBM16', 'CBM6', 'GH0']": 1,
"['CBM38', 'CBM66']": 1,
"['GH76', 'GT4']": 1,
```

```
"['CBM35', 'GH97']": 37,
"['CBM9', 'GH5']": 3,
"['CBM13', 'GH32']": 5,
"['CBM32', 'CBM35', 'GH29']": 33,
"['CBM22', 'GH11']": 2,
"['CE15', 'GH8']": 12,
"['CE6', 'GH95']": 19,
"['CBM4', 'GH11']": 4,
"['CBM57', 'CBM6']": 9,
"['CBM92', 'GH86']": 1,
"['CBM92', 'GH118']": 1,
"['CBM6', 'GH120']": 8,
"['CBM48', 'CE0', 'CE6']": 71,
"['CE7', 'GH26']": 2,
"['GH18', 'GH78']": 2,
"['CBM91', 'GH28', 'GH43']": 29,
"['CE0', 'GH26']": 1,
"['GH35', 'GH43']": 27,
"['CBM51', 'GH0']": 2,
"['CBM2', 'GH43']": 7,
"['CBM2', 'CE4', 'GH10', 'GH11']": 1,
"['CBM22', 'CBM9', 'CE4', 'GH10']": 29,
"['CBM2', 'CBM22', 'GH10']": 31,
"['CBM13', 'CBM91', 'GH43']": 77,
"['CBM2', 'GH10', 'GH62']": 11,
"['CBM2', 'GH26']": 5,
"['CBM6', 'GH99']": 33,
"['CBM13', 'GH97']": 17,
"['CBM0', 'GH92']": 239,
"['GT2', 'GT32', 'GT62']": 14,
"['GH10', 'GH62']": 27,
"['CBM2', 'GH44']": 30,
"['CBM48', 'GH13', 'GT5']": 2,
"['CBM91', 'CE6', 'GH43']": 5,
"['CBM13', 'CBM92']": 14,
"['CBM13', 'CBM32']": 17,
"['CBM2', 'CBM91', 'GH43']": 22,
"['CBM5', 'GH0']": 25,
"['CBM35', 'GH146']": 7,
"['CBM2', 'PL31']": 31,
"['CBM2', 'GH2']": 6,
"['CBM2', 'CE1', 'GH10']": 6,
"['CBM2', 'CE3', 'GH5']": 9,
"['CBM16', 'GH136']": 51,
"['CBM13', 'GH75']": 10,
"['CBM13', 'GH74']": 65,
"['CBM56', 'GH55']": 39,
"['CBM48', 'GT2']": 2,
"['CBM72', 'GH16']": 18,
"['GH30', 'GH43']": 12,
```

```
"['CBM86', 'CBM9', 'GH10']": 1,
"['CBM13', 'CBM2', 'CBM36', 'CBM6', 'GH10', 'GH43']": 1,
"['CBM13', 'CBM2', 'CE0', 'GH10']": 1,
"['CBM2', 'CBM6']": 6,
"['CE8', 'PL9']": 17,
"['CBM32', 'CBM61', 'GH53']": 4,
"['CBM66', 'PL11']": 3,
"['CBM2', 'CBM6', 'GH43']": 19,
"['CBM2', 'PL9']": 17,
"['CBM2', 'GH95']": 19,
"['CBM2', 'CE8']": 8,
"['CBM13', 'GH23']": 20,
"['CBM77', 'PL9']": 2,
"['CBM13', 'CBM35', 'GH98']": 2,
"['GH43', 'GH51']": 36,
"['CBM27', 'CE20']": 3,
"['CBM0', 'CBM23', 'CBM54', 'CBM59', 'GH26']": 1,
"['CBM3', 'GH6']": 141,
"['CBM64', 'GH12']": 4,
"['CBM56', 'CBM6', 'GH64']": 2,
"['CBM64', 'GH5']": 11,
"['CBM64', 'GH10']": 4,
"['CBM3', 'CBM64', 'GH9']": 11,
"['CBM48', 'GH10']": 3,
"['CBM13', 'GH25']": 44,
"['CBM13', 'CBM32', 'GH5']": 27,
"['CBM42', 'GH76']": 3,
"['CBM20', 'CBM25', 'GH119']": 28,
"['CBM61', 'GH99']": 5,
"['CBM32', 'GH128']": 14,
"['CBM42', 'CE2']": 1,
"['CBM61', 'GH30']": 5,
"['CBM61', 'GH18']": 3,
"['CBM66', 'GH43']": 65,
"['CBM35', 'CBM66', 'GH142']": 2,
"['CBM32', 'GH53']": 17,
"['CBM35', 'CBM9', 'GH87']": 2,
"['CBM54', 'GH43']": 9,
"['GH16', 'GH50']": 1,
"['CE12', 'GH105']": 11,
"['CBM3', 'GH44', 'GH74']": 2,
"['CBM22', 'CBM9', 'CE15', 'GH10']": 2,
"['CBM32', 'CBM54', 'CBM92', 'GH16', 'GH55']": 2,
"['CBM32', 'CBM6', 'GH81']": 6,
"['GH16', 'GT25']": 46,
"['CBM35', 'GH121']": 3,
"['CBM35', 'CBM37', 'CE3', 'GH26']": 1,
"['CBM37', 'CE12']": 1,
"['CBM22', 'CBM37', 'GH30']": 3,
"['CBM35', 'CBM37', 'GH26']": 1,
```

```
"['CBM35', 'CBM37', 'GH98']": 1,
"['CBM13', 'CBM35', 'CE12']": 2,
"['CBM37', 'GH11']": 1,
"['CBM22', 'CBM37', 'CE4', 'GH11']": 3,
"['CBM37', 'CE8', 'PL10']": 1,
"['CBM22', 'CE0']": 1,
"['CBM37', 'PL11']": 1,
"['CBM22', 'CBM37', 'GH10']": 4,
"['CBM22', 'CBM91', 'CE0', 'GH43']": 2,
"['CBM37', 'GH9']": 1,
"['CBM13', 'CBM37', 'GH43']": 1,
"['CBM37', 'PL1']": 1,
"['CBM3', 'CBM37', 'GH9']": 2,
"['CBM13', 'CBM37', 'PL11']": 1,
"['CBM13', 'CBM37', 'CE12']": 1,
"['CBM37', 'CBM77', 'PL1']": 1,
"['CBM22', 'CBM37', 'CE0', 'GH11']": 2,
"['CBM37', 'GH74']": 1,
"['CBM13', 'CBM37']": 1,
"['CBM37', 'CBM62', 'GH30']": 1,
"['CBM22', 'CBM37', 'GH10', 'GH11']": 1,
"['CBM4', 'GH81']": 1,
"['CBM23', 'CBM27', 'CBM59', 'GH26']": 3,
"['CBM47', 'CBM61', 'GH53']": 1,
"['CBM32', 'GH95']": 42,
"['CBM32', 'PL0']": 7,
"['GH88', 'PL38']": 18,
"['CBM47', 'CBM6', 'PL7']": 10,
"['GH13', 'GH57']": 1,
"['GH13', 'GT4']": 1,
"['CBM6', 'GH81']": 50,
"['CBM92', 'GH18']": 27,
"['CBM32', 'GH2', 'GH64']": 31,
"['CBM57', 'GH55']": 6,
"['CBM32', 'CBM57']": 8,
"['CBM35', 'CBM57']": 3,
"['GT23', 'GT41']": 7,
"['GH0', 'GH144']": 9,
"['GH5']": 2,
"['CBM35', 'GH26', 'GH5']": 1,
"['AA1', 'CE4']": 1,
"['GH154', 'GH16']": 3,
"['CE8', 'GH105']": 2,
"['CE12', 'GH106']": 1,
"['CBM0', 'GH10']": 13,
"['GH0', 'GH28']": 4,
"['CBM22', 'GH5']": 4,
"['GH2', 'GH43']": 11,
"['CBM47', 'GT2']": 78,
"['CBM40', 'GH13']": 3,
```

```
"['CBM3', 'GH10']": 12,
"['CBM3', 'CBM32', 'GH5']": 3,
"['CBM3', 'CE0']": 3,
"['CBM13', 'CBM2', 'PL11']": 2,
"['CBM13', 'CBM2', 'CBM6', 'GH43']": 4,
"['GT101', 'GT113']": 5,
"['CBM13', 'CE4']": 3,
"['CBM13', 'GH128']": 12,
"['CBM2', 'CBM22', 'CBM9', 'GH10']": 9,
"['GH19', 'GH23']": 5,
"['CBM6', 'CBM91']": 48,
"['CBM32', 'CE0']": 1,
"['CBM0', 'CBM22', 'CBM9', 'CE4', 'GH10']": 3,
"['CBM11', 'CBM32', 'GH3']": 11,
"['GH2', 'GH53']": 16,
"['CBM64', 'GH48']": 3,
"['CBM46', 'CBM64', 'GH5']": 1,
"['GH148', 'GH30']": 1,
"['CBM35', 'CBM4', 'GH16']": 1,
"['CE4', 'GH39']": 1,
"['CBM22', 'GH30']": 2,
"['CBM5', 'CBM61', 'GH53']": 2,
"['GT107', 'GT8']": 8,
"['GT45', 'GT93']": 13,
"['CBM35', 'GH8']": 3,
"['GH18', 'GH8']": 2,
"['CBM0', 'PL11']": 1,
"['GT32', 'GT62']": 36,
"['CBM9', 'GH50']": 15,
"['CBM23', 'CBM59', 'GH26']": 5,
"['CBM16', 'PL0']": 8,
"['CBM35', 'CBM9', 'PL0']": 5,
"['CBM23', 'CBM27', 'CBM54', 'CBM59', 'GH26']": 12,
"['CBM22', 'CBM6', 'GH30']": 4,
"['CBM0', 'CBM6', 'GH3', 'GH30']": 1,
"['CBM46', 'GH5', 'GH74']": 3,
"['CBM6', 'PL1']": 6,
"['CE0', 'PL9']": 3,
"['CBM66', 'GH5']": 41,
"['GH29', 'GH33']": 3,
"['CBM8', 'GH18']": 38,
"['CBM48', 'GH13', 'GH133']": 3,
"['CBM13', 'GH35']": 11,
"['CE0', 'CE15']": 6,
"['GT0', 'GT9']": 4,
"['CBM3', 'GH44', 'PL11']": 1,
"['CBM3', 'CBM66', 'PL9']": 1,
"['CBM13', 'GH16', 'GH18']": 1,
"['CBM90', 'PL28']": 15,
"['CBM1', 'CE5']": 20,
```

```
"['CBM1', 'GH62']": 27,
"['CBM1', 'GH131']": 23,
"['CBM1', 'GH16']": 2,
"['CBM1', 'CE2']": 2,
"['CBM18', 'CBM50']": 12,
"['CBM1', 'GH28']": 2,
"['CBM1', 'CE16']": 20,
"['GH12', 'GT2']": 2,
"['CBM32', 'GH16', 'GH20']": 24,
"['CBM66', 'PL1']": 81,
"['CBM13', 'CBM57', 'GH87']": 1,
"['CBM13', 'CBM6', 'GH55']": 1,
"['CBM78', 'GH5']": 5,
"['CBM6', 'CE6', 'GH11']": 1,
"['CBM22', 'CBM9', 'CE0', 'GH10']": 1,
"['CBM9', 'CE0', 'GH10']": 2,
"['CBM6', 'GH10', 'GH11']": 2,
"['CBM36', 'CBM6', 'GH11']": 1,
"['CE4', 'GT51']": 9,
"['CBM13', 'GH26', 'GH75']": 3,
"['CBM32', 'GH46']": 24,
"['GH88', 'PL8']": 1,
"['CBM3', 'GH128']": 2,
"['GH29', 'GH89']": 1,
"['GH0', 'GH104']": 1,
"['CBM23', 'CBM27', 'GH26']": 39,
"['CBM32', 'GH141']": 13,
"['CBM2', 'GH3']": 37,
"['CBM22', 'CBM4', 'GH10']": 7,
"['CBM35', 'CBM6', 'GH30']": 1,
"['CBM9', 'GH9']": 3,
"['CBM35', 'CBM6', 'GH3', 'GH30']": 3,
"['CBM25', 'CBM41', 'GH13']": 3,
"['CBM6', 'GH92']": 42,
"['CBM53', 'GH0']": 1,
"['CE13', 'GT1']": 1,
"['CBM88', 'GH16']": 1,
"['GT4', 'GT9']": 30,
"['GT107', 'GT2']": 10,
"['CBM16', 'PL7']": 17,
"['CBM4', 'GH0']": 10,
"['CBM60', 'GH0']": 2,
"['GH1', 'GH5']": 1,
"['GT116', 'GT41']": 1,
"['GT11', 'GT41']": 1,
"['CBM84', 'GH16']": 2,
"['CBM57', 'GH5']": 1,
"['CBM32', 'GH0']": 42,
"['CBM17', 'CBM28']": 1,
"['CBM67', 'GH140']": 3,
```

```
"['GH36', 'GH38']": 1,
"['CBM35', 'CE20']": 3,
"['CBM3', 'PL1']": 5,
"['GH173', 'GH36']": 25,
"['GH35', 'GH53']": 1,
"['GH28', 'GH88']": 1,
"['CBM4', 'GH3']": 1,
"['CBM22', 'CE4', 'GH10']": 3,
"['GH25', 'GH73']": 8,
"['CE15', 'GH78']": 2,
"['GT111', 'GT29', 'GT8']": 2,
"['CBM2', 'GH27']": 10,
"['CE12', 'CE2']": 2,
"['CBM2', 'GH141']": 1,
"['CBM20', 'GH97']": 3,
"['CBM84', 'GH43']": 2,
"['CBM84', 'PL3']": 8,
"['CBM8', 'GH44']": 13,
"['CBM42', 'GH62']": 2,
"['GH0', 'GH172']": 3,
"['CBM0', 'GH93']": 7,
"['CBM13', 'GH47']": 4,
"['CBM20', 'GT20']": 9,
"['CBM38', 'GH116']": 42,
"['CBM32', 'GH2', 'GH20']": 13,
"['CBM13', 'GH142']": 20,
"['CBM13', 'CBM35']": 14,
"['CBM2', 'CE4', 'GH10']": 4,
"['CBM32', 'CE3']": 6,
"['CBM12', 'CBM32', 'GH81']": 3,
"['CBM2', 'GH64']": 8,
"['CBM81', 'GH5']": 1,
"['CE4', 'GH0', 'GT2']": 6,
"['GH3', 'GH43']": 3,
"['CBM13', 'CBM26', 'GH13']": 1,
"['GH81', 'GT2']": 2,
"['CBM16', 'PL12']": 1,
"['CBM67', 'GH0']": 2,
"['CE12', 'PL22']": 1,
"['GH141', 'GH2']": 4,
"['CBM4', 'CE15']": 4,
"['CBM32', 'CBM6', 'GH16']": 5,
"['GT2', 'PL38']": 3,
"['GH158', 'GH26', 'GT2']": 6,
"['CBM60', 'GH10']": 3,
"['CBM88', 'GH74']": 8,
"['CBM2', 'CBM60', 'GH30']": 2,
"['CBM66', 'GH136']": 15,
"['CBM35', 'CBM61', 'GH27', 'GH31']": 13,
"['CE6', 'GH43']": 2,
```

```
        "['CE0', 'CE20']": 1,
        "['CBM35', 'GH93']": 14,
        "['GH26', 'GH3']": 1,
        ...})
```

The `Counter` object is not easy to read. Therefore, parse the `Counter` object into a dataframe, listing the group of CAZy families and the incidence (specifically, the number of protein accessions associated with this group in the local CAZyme database).

```
# sort the fam pairs into descending order
ordered_counter = {key: val for key, val in sorted(counter.items(),
key = lambda ele: ele[1], reverse = True)}

# create a df of fam1, fam2, freq
cooccurring_fam_data = []

for families in tqdm(ordered_counter, desc='Building coccurring df'):
    freq = ordered_counter[families]
    fams = families.replace('[','').replace(']','').replace("'",
"").replace(',','').replace(' ', '+').strip()

    cooccurring_fam_data.append([fams, freq])

# write to csv
cooccur_fams_df = pd.DataFrame(cooccurring_fam_data,
columns=['Families', 'Incidence'])
cooccur_fams_df.to_csv('../Data/cooccuring_families/cooccurring_fams.c
sv')
cooccur_fams_df
```

{"model_id":"86718dc63b2344908ab6371af34e9043","version_major":2,"vers
ion_minor":0}

	Families	Incidence
0	CBM48+GH13	45737
1	CBM50+GH23	14178
2	CBM91+GH43	9730
3	CBM34+GH13	8244
4	CBM5+GH18	5750
...	...	...
2229	CBM0+CBM35+GH39	1
2230	CE3+GH43	1
2231	CBM5+GH59	1
2232	CBM35+CBM47+GH107	1
2233	GH125+GH16	1

[2234 rows x 2 columns]

In the above dataframe, `Families` lists the CAZy families that appear together in the same CAZyme (each CAZyme was identified by its unique NCBI protein version accession). The

`Incidence` is the number of CAZymes (specifically, the number of unique NCBI protein version accessions) that contained all CAZy families in the group of families.

Identify CAZymes with multiple catalytic domains

Most of the groups of CAZy families contained a Carbohydrate Binding Module (CBM) and a catalytic domain. CBMs are non-catalytic domains that facilitate the enzyme targeting and/or binding its substrate.

The interest was in in CAZymes containing more than one catalytic domain. Therefore, iterate through `coocur_fams_df` and identify rows containing CAZymes with multiple catalytic domains.

```
multi_cat_domains_data = []

for ri in tqdm(range(len(cooccur_fams_df)), desc="Identify groups with multiple catalytic domains"):
    families = cooccur_fams_df.iloc[ri]['Families'].split('+')

    num_of_fams_in_group = len(families)

    num_of_catalytic_domains = 0
    for fam in families:
        if fam.startswith('CBM') is False:  # Does not start with CBM, therefore Not a CBM domain
            # CBM domains are non-catalytic
            # all other domains are catalytic domains
            num_of_catalytic_domains += 1

    multi_cat_domains_data.append( [
        cooccur_fams_df.iloc[ri]['Families'],
        cooccur_fams_df.iloc[ri]['Incidence'],
        num_of_fams_in_group,
        num_of_catalytic_domains,
    ] )

cat_domains_df = pd.DataFrame(multi_cat_domains_data,
                                    columns = [
                                        'Families',
                                        'Incidence',
                                        'Num_of_fams',
                                        'Num_of_catalytic_domains'
                                    ])
cat_domains_df
```

{"model_id":"46453c7222d54b0391f792d543c074c2","version_major":2,"version_minor":0}

	Families	Incidence	Num_of_fams	Num_of_catalytic_domains

```
0            CBM48+GH13      45737            2
1
1            CBM50+GH23      14178            2
1
2            CBM91+GH43       9730            2
1
3            CBM34+GH13       8244            2
1
4             CBM5+GH18       5750            2
1
...                ...        ...          ...
...
2229    CBM0+CBM35+GH39          1            3
1
2230           CE3+GH43          1            2
2
2231          CBM5+GH59          1            2
1
2232  CBM35+CBM47+GH107          1            3
1
2233         GH125+GH16          1            2
2

[2234 rows x 4 columns]
```

Write out the groups of CAZy families that contain more that one catalytic family, as well the number of CAZymes (identified as the number of unqiue NCBI protein accessions) annotated with the corresponding group of families, the total number of families in the group and the number of catalytic families in the group.

```
multi_cat_domains_df =
cat_domains_df[cat_domains_df['Num_of_catalytic_domains'] > 1]
multi_cat_domains_df.to_csv("../Data/cooccuring_families/multi_cat_dom
ains_groups.csv")
multi_cat_domains_df
```

```
                Families  Incidence  Num_of_fams
Num_of_catalytic_domains
5              CE4+GH153       4586            2
2
6                GT2+GT4       4244            2
2
10             GH94+GT84       2266            2
2
16              GH17+GT2       1412            2
2
18              GT0+GT2        1305            2
2
...                 ...        ...          ...
...
```

```
2225        CBM3+GH5+PL9          1            3
2
2226  CBM3+GH12+GH48+GH5          1            4
3
2227       CBM3+PL11+PL9          1            3
2
2230           CE3+GH43           1            2
2
2233          GH125+GH16          1            2
2

[741 rows x 4 columns]
```

Identify families that are subdivided

The analysis above looks at all CAZy families listed in CAZy.

To identify pairs of CAZy families whose co-occurrence could be represented by a CAZy subfamily, we could filter out CAZy families that are already subdivided into CAZy subfamilies.

The local CAZyme database `cazycsj` was queried using the bash script `get_fam_subfams.sh`, which generated the CSV file `cazy_fam_subfams.csv`, which lists every pair of CAZy family and associated CAZy subfamily.

```
cazy_fam_subfams_df =
pd.read_csv('../Data/cooccuring_families/cazy_fam_subfams.csv')
cazy_fam_subfams_df.head(3)
```

```
  family subfamily
0   AA10       NaN
1   CBM9       NaN
2   GH16   GH16_24
```

Itereate through `cazy_fam_subfams_df` and aggregate together all subfamilies for each CAZy family, then identify which families are not associated with any subfamilies. Write the names of these families to a new list.

```
cazy_fam_dict = {}  # {fam: {subfams}}

for ri in tqdm(range(len(cazy_fam_subfams_df)), desc='Gathering
subfamilies'):
    row = cazy_fam_subfams_df.iloc[ri]  # select the row of the
dataframe with the rowindex (ri)

    fam = row['family']
    subfam = row['subfamily']

    # check if the family is already in the dict
    try:
        cazy_fam_dict[fam].add(subfam)  # if present add subfam to set
```

```
of subfams
    except KeyError:
        # add fam to the cazy_fam_dict
        cazy_fam_dict[fam] = {subfam}

# print example output
print('Subfamilies for AA10:', cazy_fam_dict['AA10'])
print('Subfamilies for GH16:', cazy_fam_dict['GH16'])
```

{"model_id":"81115dd8e82e4e5b8189e71c0c4b460a","version_major":2,"vers
ion_minor":0}

```
Subfamilies for AA10: {nan}
Subfamilies for GH16: {nan, 'GH16_27', 'GH16_13', 'GH16_12', 'GH16_1',
'GH16_4', 'GH16_15', 'GH16_7', 'GH16_6', 'GH16_18', 'GH16_20',
'GH16_19', 'GH16_11', 'GH16_23', 'GH16_16', 'GH16_14', 'GH16_24',
'GH16_3', 'GH16_8', 'GH16_21', 'GH16_26', 'GH16_10', 'GH16_9',
'GH16_2', 'GH16_25', 'GH16_22', 'GH16_5', 'GH16_17'}
```

Identify families that have and have not been divided into subfamilies.

```
fams_wo_subfams = []  # families without subfams
fams_with_subfams = []  # families with subfams

for fam in cazy_fam_dict:
    if len(cazy_fam_dict[fam]) == 1:
        if (str(cazy_fam_dict[fam])) == '{nan}':
            fams_wo_subfams.append(fam)
    else:
        fams_with_subfams.append(fam)


fams_with_subfams = set(fams_with_subfams)  # remove duplicates from
list if present
with open('fams_with_subfams', 'w') as fh:
    for fam in fams_with_subfams:
        fh.write(f"{fam}\n")

fams_wo_subfams = set(fams_wo_subfams)  # remove duplicates from list
if present
with open('fams_withOUT_subfams', 'w') as fh:
    for fam in fams_wo_subfams:
        fh.write(f"{fam}\n")

print(
    f"{len(fams_with_subfams)} families are subdivided into
subfamilies and "
    f"{len(fams_wo_subfams)} are NOT divided into subfamilies"
)
```

33 families are subdivided into subfamilies and 420 are NOT divided into subfamilies