Plot clustermaps of Sourmash output for GH7 100 sequences for different kmer sizes

Using distance matrixes created by sourmash

Imports

```
from sourmash import fig
import seaborn as sns
import pandas as pd
```

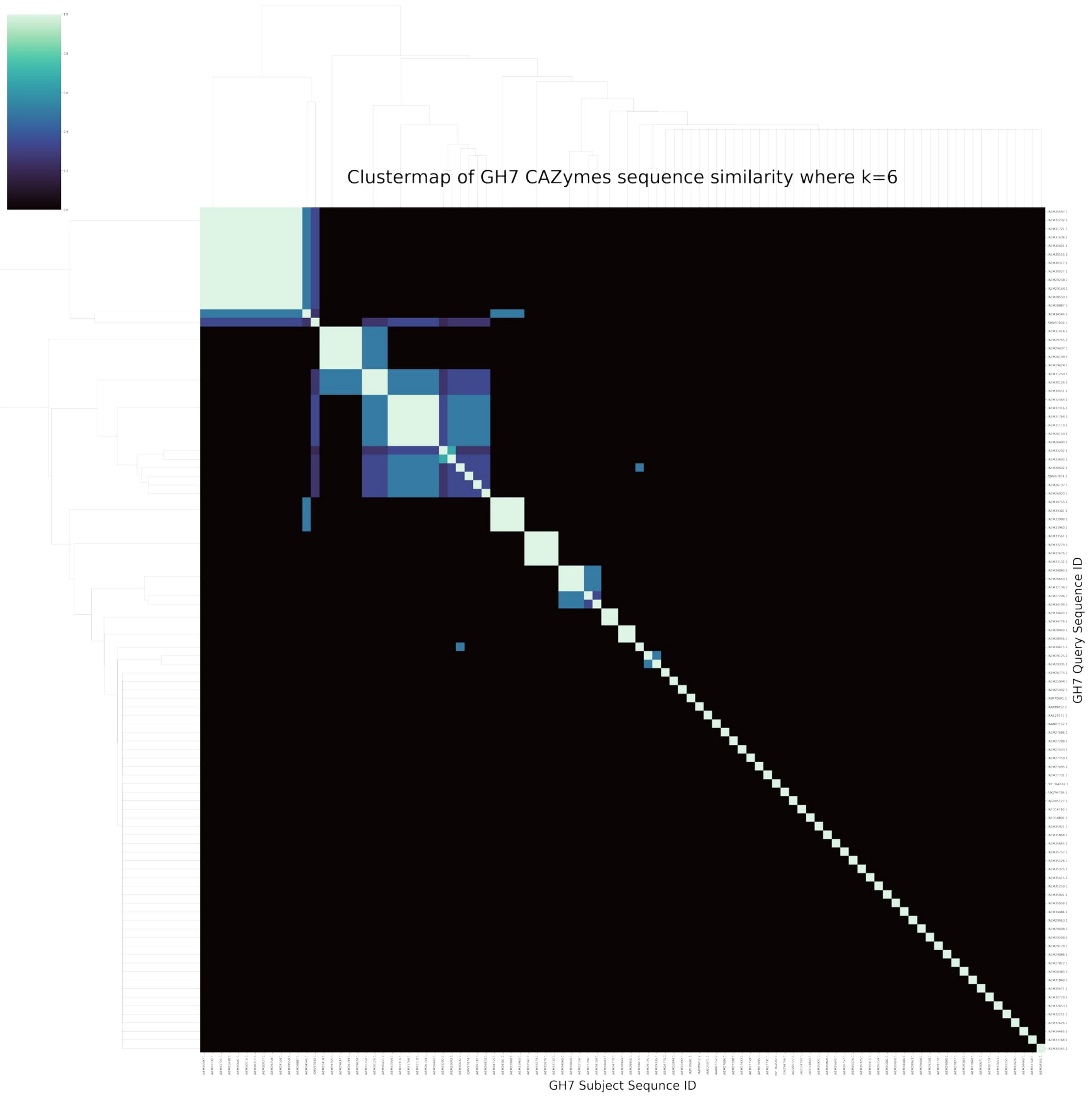
Function to create clustermap

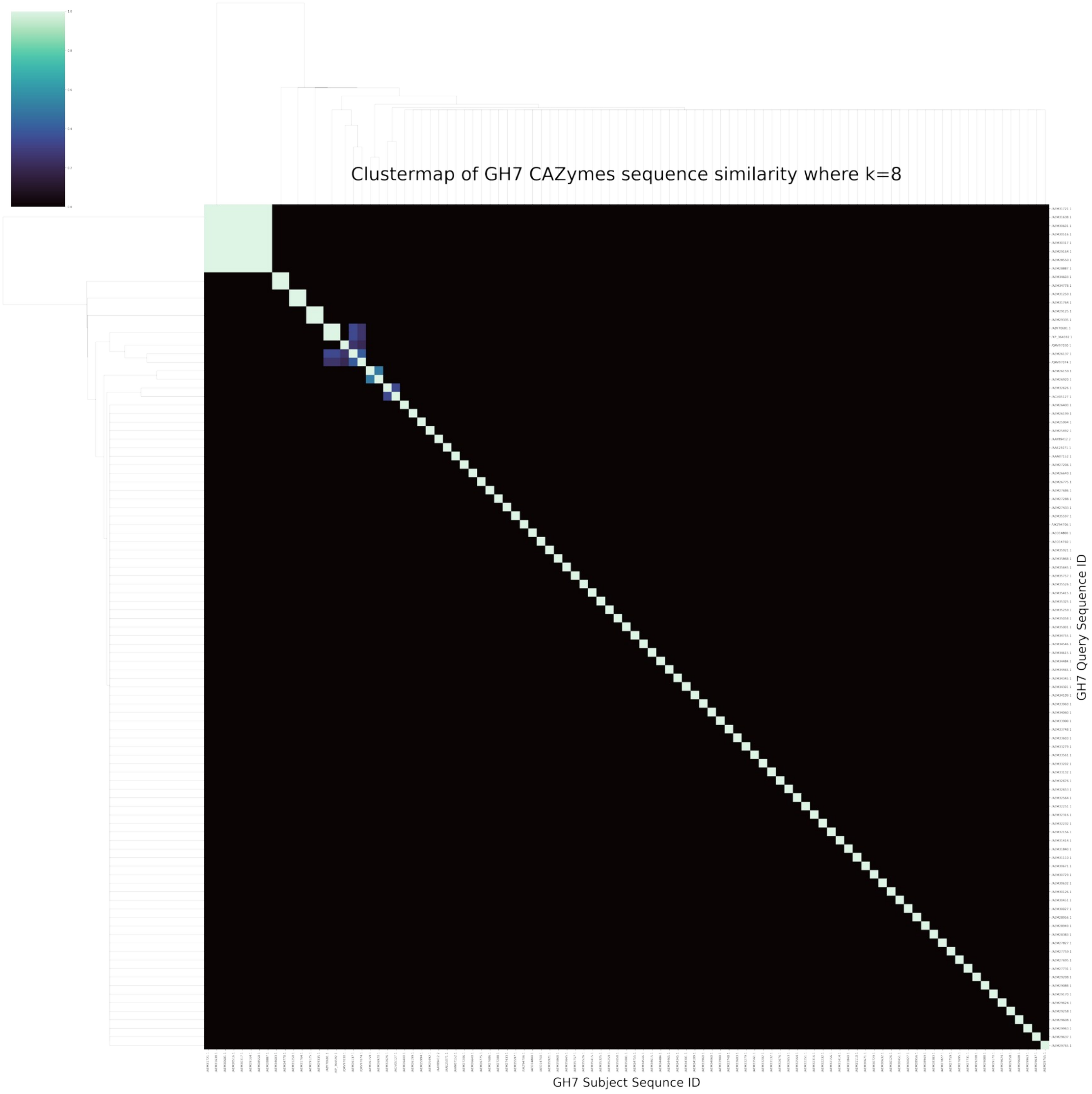
```
def plotsourmash(nseqs, kmer): # Function is called plotsourmash.
    """Function takes the output distance matrix of Jacarrd indexes
calculated by sourmash and plots them as
    a cluster map for a specified number of sequences (nsqs) and kmer
size (kmer). """
    matrix, labels =
fig.load_matrix_and_labels(f'../Results/Sourmash_out/{nseqs}out/all{ns
eqs}k{kmer}.dist')
    # Load matrix with spesified number of sequences and kmer size.
    newlabels=[] # Create empty list called newlables.
    for label in labels: # Cycle through the filenames/lables to:
        label = label.replace('.fasta', '') # Remove .fasta
        newlabel = label.replace(f'Data/Sourmash/sourmash_{nseqs}',
'') # Remove file path
        newlabels.append(newlabel) # Add the new lable of only the
sequence name to the new lables list.

    df = pd.DataFrame(matrix, columns=newlabels, index=newlabels)
    # Create dataframe with the Jaccard index values and sequence IDs
lables.

    figure=sns.clustermap(df, cmap="mako", vmin=0.0, vmax=1.0,
figsize=(50, 50));
    # Plot clustermap of the data frame created.
    figure.ax_heatmap.set_xlabel("GH7 Subject Sequnce ID",fontsize=40,
labelpad=15)
    figure.ax_heatmap.set_ylabel("GH7 Query Sequence ID",fontsize=40,
labelpad=15)
    figure.ax_heatmap.set_title(
        f'Clustermap of GH7 CAZymes sequence similarity where
k={kmer}',
        fontsize=60,
        pad=80
    )
    figure.savefig(f'../Results/Sourmash_out/allk{kmer}.dist.png') #
Save in results folder.
```

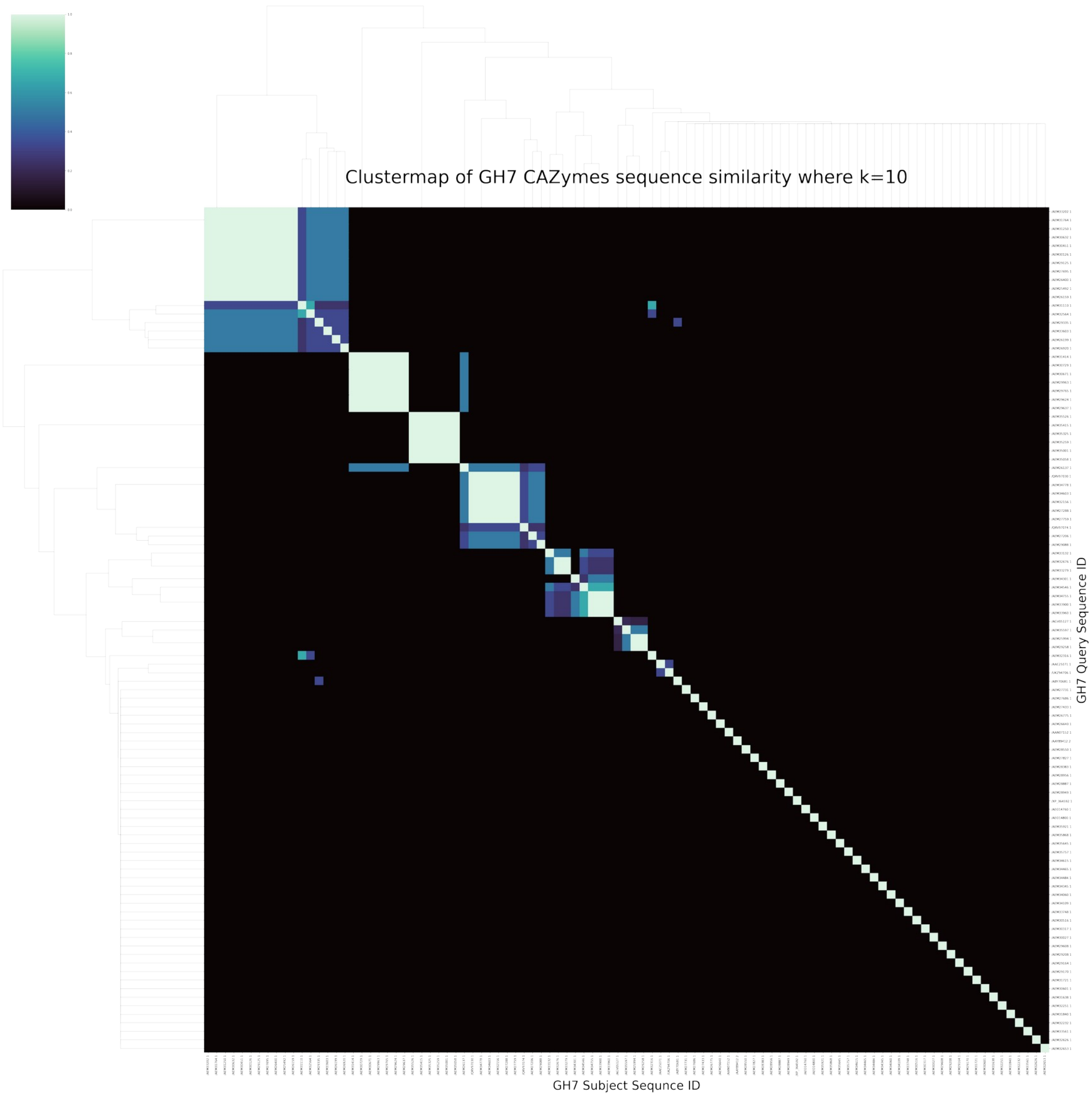
```
plot = plotsourmash(100, 6) # Run function for 100 sequences and kmer
= 6,
plot = plotsourmash(100, 8) # kmer = 8
plot = plotsourmash(100, 10) # kmer = 10
```

```
help(plotsourmash) # Call functions doc string to explain what the
function does.
```

```
Help on function plotsourmash in module __main__:

plotsourmash(nseqs, kmer)
    Function takes the output distance matrix of Jacarrd indexes
calculated by sourmash and plots them as
    a cluster map for a specified number of sequences (nsqs) and kmer
size (kmer).
```


Clustermap of GH7 CAZymes sequence similarity where k=6
1.0
0.8
0.6
0.4
0.2
0.0
/AEM35597.1
/AEM32232.1
/AEM31721.1
/AEM31638.1
/AEM30601.1
/AEM30516.1
/AEM30317.1
/AEM30027.1
/AEM29258.1
/AEM29164.1
/AEM28550.1
/AEM28887.1
/AEM34546.1
/QRV97030.1
/AEM31414.1
/AEM29765.1
/AEM29637.1
/AEM26199.1
/AEM29624.1
/AEM31250.1
/AEM30126.1
/AEM30451.1
/AEM32564.1
/AEM32316.1
/AEM31764.1
/AEM31110.1
/AEM26159.1
/AEM26400.1
/AEM33202.1
/AEM33603.1
/AEM30632.1
/QRV97074.1
/AEM26137.1
/AEM26920.1
/AEM34755.1
/AEM34301.1
/AEM33900.1
/AEM33960.1
/AEM33561.1
/AEM33279.1
/AEM32676.1
/AEM33132.1
/AEM34060.1
/AEM26640.1
/AEM32156.1
/AEM27206.1
/AEM34109.1
/AEM34603.1
/AEM34778.1
/AEM28949.1
/AEM28956.1
/AEM34615.1
/AEM29125.1
/AEM29335.1
/AEM26775.1
/AEM25994.1
/AEM25492.1
/ABY70681.1
/AAY89412.2
/AAE25071.1
/AAN07152.1
/AEM27686.1
/AEM27288.1
/AEM27433.1
/AEM27759.1
/AEM27695.1
/AEM27731.1
/XP_364592.1
/UKZ94706.1
/AGV05127.1
/AEO14760.1
/AEO14800.1
/AEM35921.1
/AEM35868.1
/AEM35645.1
/AEM35757.1
/AEM35526.1
/AEM35325.1
/AEM35415.1
/AEM35259.1
/AEM35001.1
/AEM35058.1
/AEM34484.1
/AEM29963.1
/AEM29608.1
/AEM29208.1
/AEM29170.1
/AEM29088.1
/AEM27827.1
/AEM28383.1
/AEM31840.1
/AEM30671.1
/AEM30729.1
/AEM32653.1
/AEM32251.1
/AEM32626.1
/AEM34465.1
/AEM33748.1
/AEM34345.1
GH7 Query Sequence ID
/AEM32626.1
/AEM34465.1
/AEM33748.1
/AEM34345.1
GH7 Subject Sequnce ID


Clustermap of GH7 CAZymes sequence similarity where k=8
1.0
0.8
0.6
0.4
0.2
0.0
GH7 Query Sequence ID
GH7 Subject Sequnce ID


Clustermap of GH7 CAZymes sequence similarity where k=10
GH7 Query Sequence ID
GH7 Subject Sequnce ID