Plot clustermap of BLASTP output for GH7 100 sequences

Imports

```
import seaborn as sns
import pandas as pd
import numpy as np
```

Function to create clustermap

```
def plotblastp(nseqs): # Function is called plotblastp.
    """Function takes the output matrix calculated by blastp and plots
them as a cluster map for a
    specified number of sequences (nsqs). """
    infile = f"../Results/BLASTP_out/{nseqs}seqs_blastp_out.tsv"
    # Set correct file with desired number of sequences and speed to
variable infile.

    df = pd.read_table(infile, header=None) # Create a dataframe.
    df.columns = ["qseqid", "sseqid", "qlen", "bitscore"] # Name the
columnes in the dataframe.
    df["normalisedbitscore"]=df["bitscore"]/df["qlen"] # Caculate the
normalised bitscore.
    df = df.drop(["qlen", "bitscore"],axis=1) # Remove uneeded
columnes.

    widedfx = pd.pivot_table(df, index="qseqid", columns= "sseqid",
values="normalisedbitscore")
    # Turn Long dataframe into a wide dataframe.
    widedf = widedfx.fillna(0) # Remove any values NaN and replace
with 0.

    figure=sns.clustermap(widedf, cmap="BuPu_r", figsize=(50, 50));
    # Plot clustermap of the data frame created.
    figure.ax_heatmap.set_xlabel("GH7 Subject Seqeunce
ID",fontsize=40, labelpad=15)
    figure.ax_heatmap.set_ylabel("GH7 Query Sequence ID",fontsize=40,
labelpad=15)
    figure.ax_heatmap.set_title(
        'Clustermap of GH7 CAZymes sequence similarity calculated by
BLASTP',
        fontsize=60,
        pad=80
    )
    figure.savefig(f'../Results/BLASTP_out/{nseqs}_blastp.png') # Save
in results folder.

plot = plotblastp(100) # Run function for 100 sequences.
```

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

```
/home/cjohns/.local/lib/python3.6/site-packages/seaborn/matrix.py:654:
UserWarning: Clustering large matrix with scipy. Installing
`fastcluster` may give better performance.
  warnings.warn(msg)

Help on function plotblastp in module __main__:

plotblastp(nseqs)
    Function takes the output matrix calculated by blastp and plots
them as a cluster map for a
    specified number of sequences (nsqs).
```