

generate_Matrix_CSV_AlisEdits

April 1, 2021

```
[1]: import pandas as pd
import seaborn as sn
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

0.1 Read in Data

```
[2]: matrix_data = pd.read_csv("PROKKA.matrix.data", names=['Gene Count', 'Phage 1', 'Phage 2'], sep=",")
```

```
[3]: unique_genome = matrix_data['Phage 1'].unique().tolist()
```

```
[4]: len(matrix_data)
```

```
[4]: 36481
```

0.2 Generate A Pairwise Count Matrix

To do so, our strategy is the following:

1. Get all the unique phages
2. We get all the unique values of phages, and the corresponding array of counts
3. Get a pandas series
4. Create a numpy matrix from the obtained series (should be size 190x190)

```
[5]: phage_list = matrix_data['Phage 1'].unique()
print(f"Number of unique phages: {len(phage_list)}")
```

Number of unique phages: 191

Now moving on to getting the pairwise counts:

```
[6]: # you want to get all the unique Phage 1 values, and the corresponding gene counts for every other phage, including itself
# here with group by we get, and now we want to get all the values as a list
c_matrix = matrix_data.groupby('Phage 1')['Gene Count'].apply(list)
```

above will result in a pandas series. Now to finally get what we need, we need to turn into an array, but first let's check the type:

```
[7]: type(c_matrix)
```

```
[7]: pandas.core.series.Series
```

```
[8]: # c_matrix.tolist()
numpy_matrix = np.asarray(c_matrix.tolist())
```

Now let's check out the matrix we got:

```
[9]: print(numpy_matrix)
print(f"\n Matrix has the shape: {numpy_matrix.shape}")
```

```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

Matrix has the shape: (191, 191)

```
[10]: np.savetxt("prokka_matrix.csv", numpy_matrix, delimiter=",")
```

which is exactly what we wanted!

0.3 Make a Dataframe with the Count Matrix

```
[11]: pairwise_count = pd.DataFrame(data = numpy_matrix)
```

```
[12]: pairwise_count.columns = list(phage_list)
```

```
[13]: pairwise_count.index = list(phage_list)
```

```
[14]: pairwise_count
```

```
[14]: GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn \
      GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
      0
      GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
      0
      GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
      0
      GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
      0
```

```

GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn

```

```

0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0

```

```

GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_000009005.1_ASM900v1_genomic.gbff_pp15.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_000009005.1_ASM900v1_genomic.gbff_pp16.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn

```

```

0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
7
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_000009005.1_ASM900v1_genomic.gbff_pp17.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
8
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
4

GCA_000009005.1_ASM900v1_genomic.gbff_pp18.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0

```

```

GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_000009005.1_ASM900v1_genomic.gbff_pp19.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn ... \
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn ...
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn ...

```

```

GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn ...
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn ...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn ...
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn ...
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn ...
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn ...
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn ...

GCA_900081425.1_12641_2_75_genomic.gbff_pp21.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_900081425.1_12641_2_75_genomic.gbff_pp22.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn

```



```

0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_900081945.1_12673_2_15_genomic.gbff_pp10.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_900124925.1_17138_2_42_genomic.gbff_pp13.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...

```

```

GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_900126435.1_17175_2_81_genomic.gbff_pp6.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_900251565.1_M3684_genomic.gbff_pp14.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...

```

```

...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_900251895.1_M3925_genomic.gbff_pp17.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
4

GCA_900251895.1_M3925_genomic.gbff_pp18.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0

```

```

...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn \
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn
0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
GCA_000009005.1_ASM900v1_genomic.gbff_pp10.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp11.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp12.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp13.ffn
0
GCA_000009005.1_ASM900v1_genomic.gbff_pp14.ffn

```

```

0
...
...
GCA_900251565.1_M3684_genomic.gbff_pp14.ffn
0
GCA_900251895.1_M3925_genomic.gbff_pp17.ffn
4
GCA_900251895.1_M3925_genomic.gbff_pp18.ffn
0
GCA_900323905.1_24117-WT_genomic.gbff_pp6.ffn
0
GCA_900482705.1_7915_6_43_genomic.gbff_pp16.ffn
0

[191 rows x 191 columns]

```

0.3.1 Example of plotting heatmaps for this dataframe we generated

I think the best way to visualize this data now is to do a heatmap, and pandas provides a very nice way of doing this with the actual numbers in them. Here is how:

```
[15]: ax = sns.heatmap(uniform_data, linewidth=0.5)
      plt.show()
```

```

      □
↳ -----

NameError                                Traceback (most recent call↳
↳ last)

    <ipython-input-15-4fbd8887ca0c> in <module>
----> 1 ax = sns.heatmap(uniform_data, linewidth=0.5)
      2 plt.show()

NameError: name 'uniform_data' is not defined

```

```
[16]: pairwise_count.style.background_gradient(cmap='Blues')
```

```
[16]: <pandas.io.formats.style.Styler at 0x7f8e6767e370>
```

0.4 Optional

0.4.1 Example of getting lower and upper triangular of a matrix

```
[17]: A = np.array([[2,3,4], [3,45,8], [34,7,0.8], [21,31,41]])

print(f'A : \n {A}')
print(f'A^T is : \n {A.T}')

S = np.matmul(A,A.T)

print(f'Symmetric Matrix 1: \n {S}')
```

```
A :
[[ 2.  3.  4. ]
 [ 3. 45.  8. ]
 [34.  7.  0.8]
 [21. 31. 41. ]]
A^T is :
[[ 2.  3. 34. 21. ]
 [ 3. 45.  7. 31. ]
 [ 4.  8.  0.8 41. ]]
Symmetric Matrix 1:
[[ 29.    173.    92.2   299. ]
 [ 173.   2098.   423.4  1786. ]
 [  92.2   423.4 1205.64  963.8 ]
 [ 299.   1786.   963.8 3083. ]]
```

```
[19]: symmetry = pd.DataFrame(data = S)
```

```
[20]: symmetry.style.background_gradient(cmap='Blues')
```

```
[20]: <pandas.io.formats.style.Styler at 0x7f8e66e51b20>
```

Now extract upper and lower tringulars:

```
[18]: upper = np.triu(S)
upper
```

```
[18]: array([[ 29. , 173. ,  92.2 , 299. ],
          [  0. , 2098. , 423.4 , 1786. ],
          [  0. ,   0. , 1205.64,  963.8 ],
          [  0. ,   0. ,   0. , 3083. ]])
```

```
[19]: lower = np.tril(S)
lower
```

```
[19]: array([[ 29.  ,   0.  ,   0.  ,   0.  ],
             [ 173.  , 2098.  ,   0.  ,   0.  ],
             [  92.2 ,  423.4 , 1205.64,   0.  ],
             [ 299.  , 1786.  ,  963.8 , 3083.  ]])
```