In [1]:
```
from Bio import SeqIO
import numpy as np
from sklearn.manifold import MDS
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import pandas as pd
```

In [2]:
```
def hamming_distance(s1, s2):
    if len(s1) != len(s2):
        raise ValueError("Strand lengths are not equal!")
    return sum(ch1 != ch2 for ch1,ch2 in zip(s1,s2))
```

In [3]:
```
seqList=[]
```

In [4]:
```
for record in SeqIO.parse("HW2.fas", "fasta"):
    seqList.append(record.seq)
```

In [5]:
```
hdList=[]
```

In [6]:
```
for x in seqList:
    for y in seqList:
        hdList.append(hamming_distance(x,y))
```

In [7]:
```
len(hdList)
```

Out[7]:
```
14400
```

In [8]:
```
matrix=np.array(hdList)
matrix
```

Out[8]:
```
array([ 0,  1,  1, ..., 28, 28,  0])
```

In [9]:
```
matrix.shape
```

Out[9]:
```
(14400,)
```

In [10]:
```
matrix=matrix.reshape(120,120)
```

In [11]:
```
matrix
```

Out[11]:
```
array([[ 0,  1,  1, ..., 76, 74, 79],
       [ 1,  0,  2, ..., 77, 75, 80],
       [ 1,  2,  0, ..., 77, 75, 80],
       ...,
       [76, 77, 77, ...,  0,  4, 28],
       [74, 75, 75, ...,  4,  0, 28],
       [79, 80, 80, ..., 28, 28,  0]])
```

In [12]:
```
embedding=MDS(n_components = 2)
```

In [13]:
```
xtransform=embedding.fit_transform(matrix[:120])
```

```
C:\Users\Juney\Anaconda3\lib\site-packages\sklearn\manifold\mds.py:42
1: UserWarning: The MDS API has changed. ``fit`` now constructs an di
ssimilarity matrix from data. To use a custom dissimilarity matrix, s
et ``dissimilarity='precomputed'``.
  warnings.warn("The MDS API has changed. ``fit`` now constructs an"
```

In [14]:
```
xtransform.shape
```

Out[14]:
```
(120, 2)
```

In [15]:
```
xtransform
```

Out[15]:
```
array([[ 336.21717018,  139.40752686],
       [ 338.25587276,  141.05911055],
       [ 338.49797387,  140.71737312],
       [ 342.41338996,  140.03250948],
       [ 342.27105178,  141.01730558],
       [ 327.65043255,  135.70759135],
       [ 338.61807982,  140.5400141 ],
       [ 334.93867641,  144.37553733],
       [ 338.10443632,  141.5667336 ],
       [ 334.9887192 ,  149.65042408],
       [ 337.50040771,  136.34222127],
       [ 341.83227748,  141.19036268],
       [ 332.16003933,  135.74159982],
       [ 338.12264207,  141.22257478],
       [ 338.37795136,  140.89058049],
       [ 332.4019961 ,  137.73938878],
       [ 318.17774657,  279.15373086],
       [ 315.40943539,  284.79806639],
       [ 310.49883246,  279.81182995],
       [ 318.91047806,  275.56563622],
       [ 314.86295133,  285.16771371],
       [ 315.70391649,  284.46043971],
       [ 315.57390027,  277.63251426],
       [ 315.3587506 ,  281.29820653],
       [ 308.43046049,  286.96651924],
       [ 315.34355635,  284.68094408],
       [ 316.08228527,  288.91389387],
       [ 314.44742174,  275.08415142],
       [ 305.63688618,  278.79904357],
       [ 314.13879099,  270.57893388],
       [ 307.69161122,  270.7514105 ],
       [ 310.30129998,  280.19110367],
       [ 315.58847559,  279.8325568 ],
       [ 301.31345255,  277.60056464],
       [ 315.64858003,  282.60461182],
       [ 313.56736905,  285.44175016],
       [ 305.26911398,  279.20891275],
       [ 315.6293873 ,  284.8422978 ],
       [ 307.62770316,  286.88879029],
       [ 313.72461519,  261.4947599 ],
       [ 336.01295514,  190.77543241],
       [ 311.70317345,  283.3179049 ],
       [-123.5723512 , -229.61648994],
       [-129.09383428, -220.11025011],
       [-128.8224948 , -228.790318  ],
       [-113.90864987, -204.93896562],
       [-120.75632486, -207.50022666],
       [-130.47997914, -214.9822144 ],
       [-123.40428074, -216.85874697],
       [-124.5115173 , -219.01343547],
       [-128.5570011 , -224.02972908],
       [-108.67493099, -199.64090883],
       [-135.94733364, -226.28340334],
       [-126.88167566, -225.21599934],
       [-135.24296372, -223.05203203],
       [-118.54264872, -222.70795129],
```

```
[-129.88660142, -222.39447464],
[-122.01969491, -225.90294984],
[-121.67566551, -225.94765277],
[-128.23393852, -233.95086324],
[-121.79652331, -215.19648804],
[-124.71748593, -214.60268979],
[ -35.33283015, -260.62737865],
[ -37.81782536, -257.99266893],
[ -33.13204313, -259.9517154 ],
[ -32.99273649, -257.96830219],
[ -32.90593474, -251.97762746],
[ -30.77897512, -255.90502253],
[ -33.97754007, -246.59079251],
[ -30.73427791, -242.48236534],
[ -40.10975867, -262.66286476],
[ -35.64925226, -244.90499571],
[ -39.15284629, -256.56457349],
[ -28.28570263, -240.54092306],
[ -46.51015602, -262.9642177 ],
[ -33.05563327, -258.09690709],
[ -43.71521827, -260.12328634],
[ -30.68934992, -257.2673507 ],
[ -42.2220171 , -261.31481456],
[ -38.47307636, -251.41752909],
[ -44.80431592, -254.57002215],
[-280.98089883,  -30.48610316],
[-285.84677617,  -36.02708279],
[-272.76357818,   76.70745516],
[-286.05321502,  -30.92741906],
[-248.52987944,   33.93595905],
[-282.92009272,  -40.92582843],
[-283.30098741,  -45.63403637],
[-279.43770068,   52.3420009 ],
[-261.96681492,  -41.40782028],
[-261.90167823,   -0.5215357 ],
[-276.4313365 ,   57.23683836],
[-253.68368475,   37.33890918],
[-250.12749351,   37.74605668],
[-277.45909425,  -40.47305586],
[-239.57540532,  -35.19958962],
[-287.23797854,   53.07957868],
[-279.54092839,  -40.39365986],
[-246.83069795,  -61.37865284],
[-275.92682189,  -55.66088314],
[-255.75608146,   43.02789248],
[-276.49148238,   63.64688616],
[-276.21564329,  -34.83765091],
[-263.9486101 ,  -61.42460551],
[-292.79098421,   41.29792444],
[-218.82992818,   -0.78600066],
[-286.07784944,  -51.27146284],
[-279.54839198,  -46.610564  ],
[-279.05420631,   46.98977518],
[-281.83578409,  -46.83950173],
[-224.82869204,  -10.16191534],
[-286.25797858,   45.56313358],
```

```
       [-249.0387382 ,   32.22889082],
       [-222.88634477,   -4.70765535],
       [-272.14795226,  -50.89606196],
       [-255.68219311,   30.25645086],
       [-249.80029918,   46.61044951],
       [-276.83166066,  -55.96650298],
       [-258.82907605,  -61.36553863],
       [-220.57192144,  -12.50850017]])
```

In [16]:
```
df=pd.DataFrame(xtransform)
df
```

Out[16]:

| | 0 | 1 |
|---|---|---|
| 0 | 336.217170 | 139.407527 |
| 1 | 338.255873 | 141.059111 |
| 2 | 338.497974 | 140.717373 |
| 3 | 342.413390 | 140.032509 |
| 4 | 342.271052 | 141.017306 |
| ... | ... | ... |
| 115 | -255.682193 | 30.256451 |
| 116 | -249.800299 | 46.610450 |
| 117 | -276.831661 | -55.966503 |
| 118 | -258.829076 | -61.365539 |
| 119 | -220.571921 | -12.508500 |

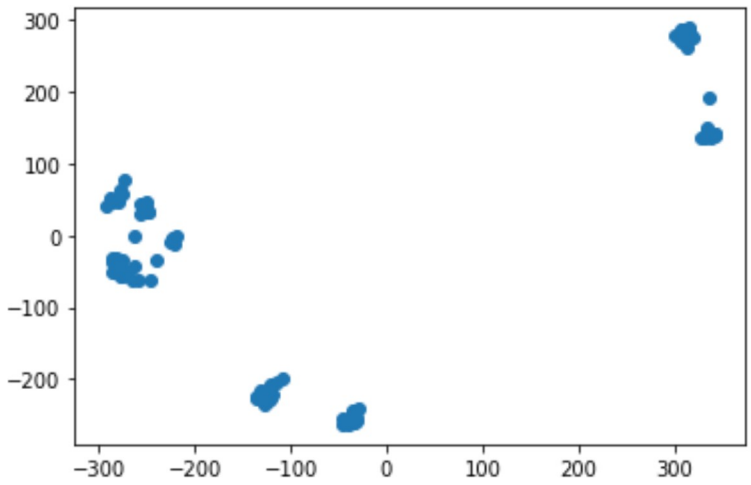
120 rows × 2 columns

In [17]:
```
x=df[0]
```

In [18]:
```
y=df[1]
```

In [19]:
```
plt.scatter(x,y)
```

Out[19]:
```
<matplotlib.collections.PathCollection at 0x275504c5548>
```




Estimate: K=3

In [20]:
```
kmeans=KMeans(n_clusters=3, random_state=0).fit(df)
```

In [21]:
```
kmeans.labels_
```

Out[21]:
```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 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, 0, 0, 2, 2, 2, 2, 2,
2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

In [22]:
```
kmeans=kmeans.cluster_centers_
```

In [23]:
```
df=pd.DataFrame(kmeans)
df
```

Out[23]:

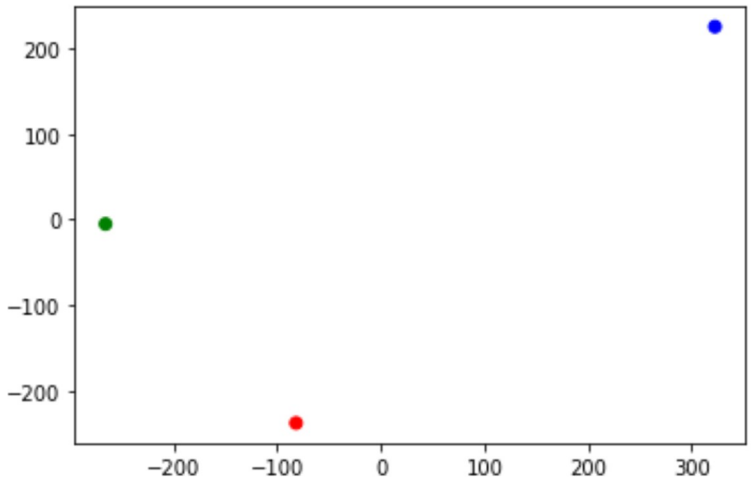
| | 0 | 1 |
|---|---|---|
| 0 | -81.719625 | -237.042542 |
| 1 | 322.500102 | 224.834823 |
| 2 | -265.588176 | -5.087267 |

In [24]:
```
x=df[0]
```

In [25]:
```
y=df[1]
```

In [26]:
```
plt.scatter(x,y, c=['red','blue','green'])
```

Out[26]:
```
<matplotlib.collections.PathCollection at 0x2755057a9c8>
```




In [ ]:
```

```