# Desktop Genetics

# Clustering CRISPR guide RNAs and Patterns recognition.

Presentation:
http://deskgen.alwaysdata.net

Mohamed Ali Nachi
Mail : nachi.med.ali@gmail.com
Tel: +216 55 458 723

# Summary

# Approach

## Data:

Data provided for the homework contains 12 columns:

1. column 1 : Gene Name
2. Spacer ID
3. Spacer Sequence
4. Delivery Plasmid immediately upon construction
5. Column 5,6 : Evolution 7 days without drugs
6. Column 7,8 : Evolution 14 days without drugs
7. Column 9,10 : Evolution 7 days with drugs
8. Column 11,12 : Evolution 14 Days with Drugs

## Development Tool

To Cluster the Data and to proceed, after that, to pattern recognition of different genes, We have used **R programming Language**[1].

During the development of our solution, we did not used any third party code as we turned into a purely statistical analysis to these data. Many libraries provide a statistical approach for bioinformatics as R bioconductor, but, we would turn to these solutions in the next step as we wanted, at this moment, use an approach purely statistical.

But, three librairies were used to generate dendrogram and strong representation of our clusters:

– ade4
– FactoClass
– FactoMiner

# General Description

## First Step:

Data Provided contains two values for each period, in our calcul, we used the "median" of the two values to proceed to the clustering.

### Clustering

[1] www.rstudio.com

Clustering CRISPR guide RNAs and Patterns recognition.

We have turned to Hierarchical Clustering as this would distinguish automatically the number of clusters. And to do this, we proceeded by:

1. Importing the Database.
2. Calculate the median of the four columns (7 days without drugs, 14 days without drugs, 7 days with drugs, 14 days without drugs)
3. Generating a new database containing 6 columns:
   - Spacer_id : Needed to identify genes.
   - Value_t0 : Delivery plasmid immediately upon construction
   - Value_t1_nodrugs : Median of the two corresponding values
   - Value-t1_drugs : Median of the two corresponding values
   - Value_t1_nodrugs : Median of the two corresponding values
   - Value_t2_drugs  Median of the two corresponding values

   with:

   t0 : immediately after plasmid delivery

   t1 : 7 days

   t2 : 14 days
4. We used a random sampling of the data, we took 20000 samples randomly from the data as our local computer cannot perform a clustering of 64000 rows.
5. With Euclidian distance fixed, we have plotted a dendrogram to know what number of clusters we should use.

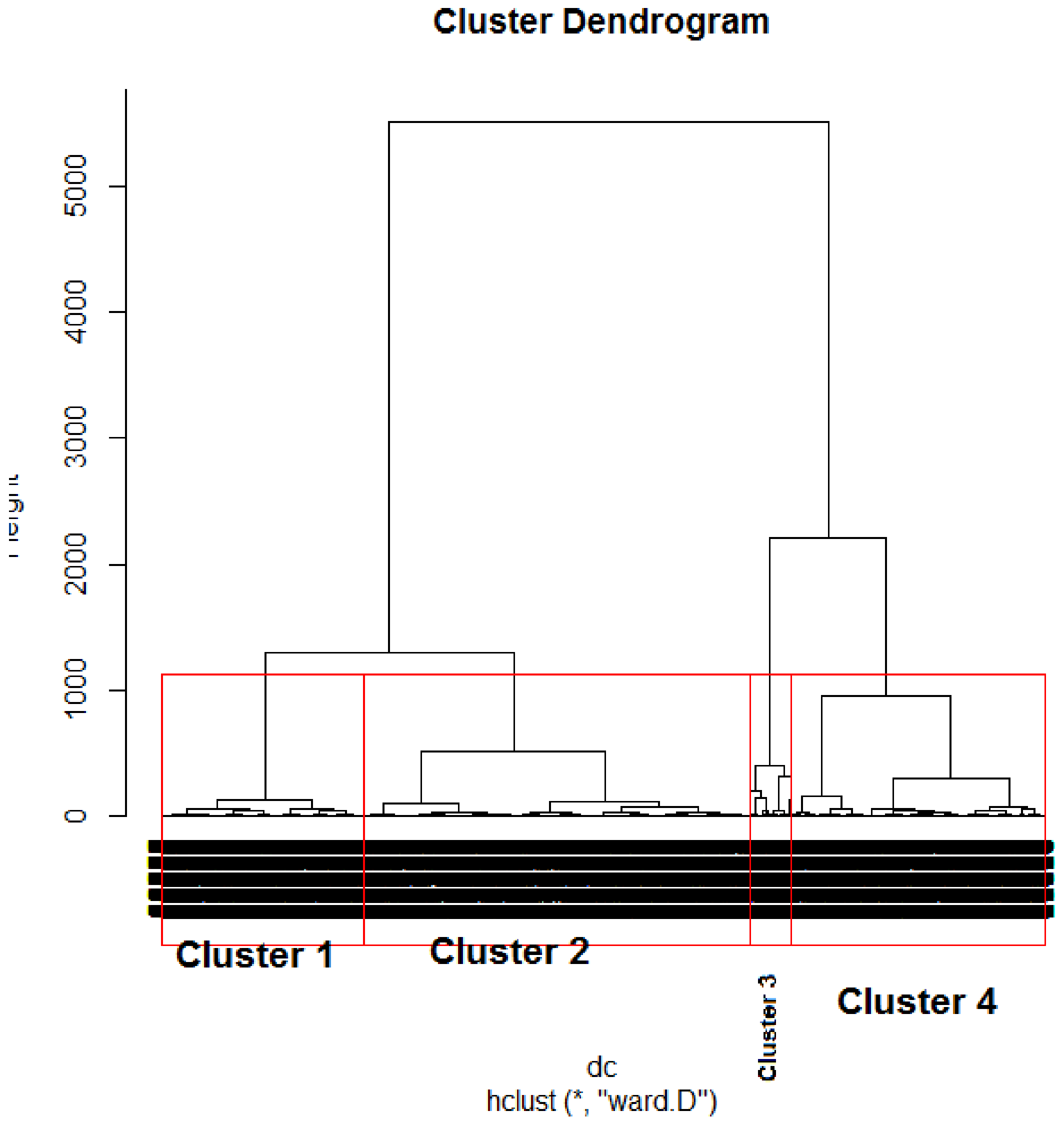
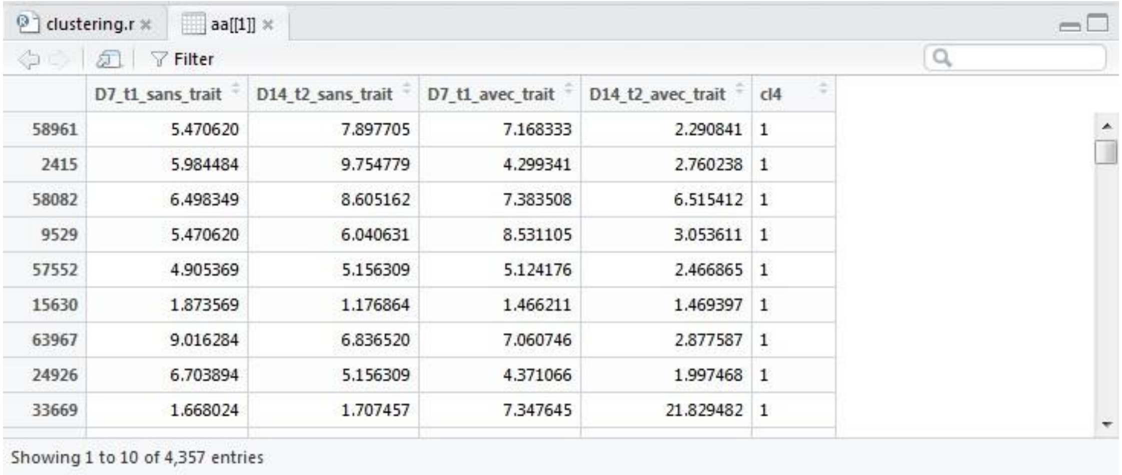


**Figure 1 : Clustering Dendrogram**

The dendrogam showed that data can be clustred in 4 clusters depending on the evolution of their corresponding values.

6. Having the data clustered, we need to extract genes in every cluster to be able to proceed to the pattern recognition of genes, this would provide us with the similarity between genes. In other words, by extracting genes present in every cluster, and with PCA statistical approach, we would be able to extract rules explaining the evolution of genes with or without drugs.

   Here is an example of data extracted:



| | D7_t1_sans_trait | D14_t2_sans_trait | D7_t1_avec_trait | D14_t2_avec_trait | cl4 |
|---|---|---|---|---|---|
| 58961 | 5.470620 | 7.897705 | 7.168333 | 2.290841 | 1 |
| 2415 | 5.984484 | 9.754779 | 4.299341 | 2.760238 | 1 |
| 58082 | 6.498349 | 8.605162 | 7.383508 | 6.515412 | 1 |
| 9529 | 5.470620 | 6.040631 | 8.531105 | 3.053611 | 1 |
| 57552 | 4.905369 | 5.156309 | 5.124176 | 2.466865 | 1 |
| 15630 | 1.873569 | 1.176864 | 1.466211 | 1.469397 | 1 |
| 63967 | 9.016284 | 6.836520 | 7.060746 | 2.877587 | 1 |
| 24926 | 6.703894 | 5.156309 | 4.371066 | 1.997468 | 1 |
| 33669 | 1.668024 | 1.707457 | 7.347645 | 21.829482 | 1 |

Showing 1 to 10 of 4,357 entries

**Figure 2 : Sample of data in a cluster**

As explained above, we clustered data with a new database containing medians. We extract the data with cluster_id that we can distinguish data in every cluster.

7. Before we proceed to pattern recognition between genes in the same cluster and comparing them, we plotted, in a second time, two plots:

   - The first one with centers only:

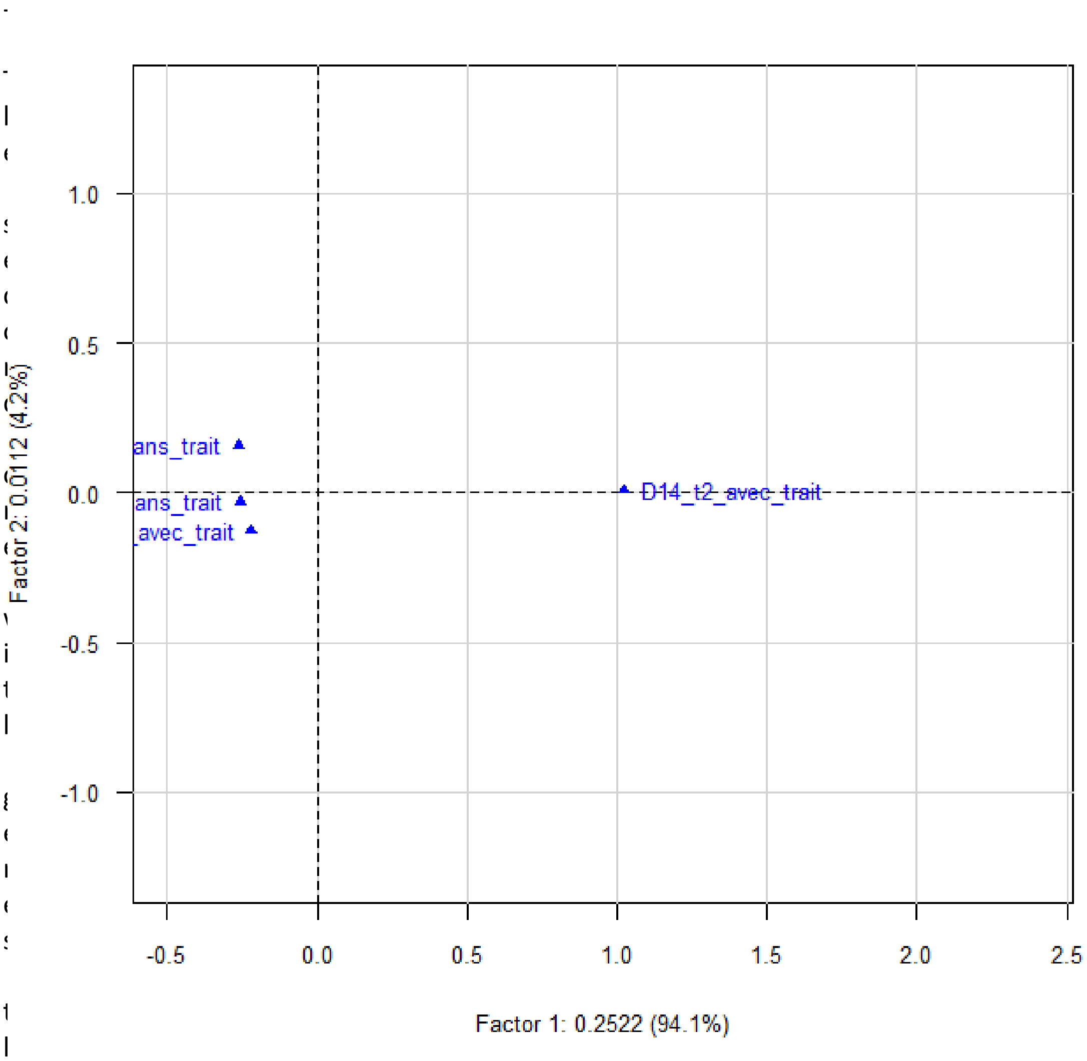


**Figure 3 : Plot of clusters centers**

t we see relations between genes and centers:

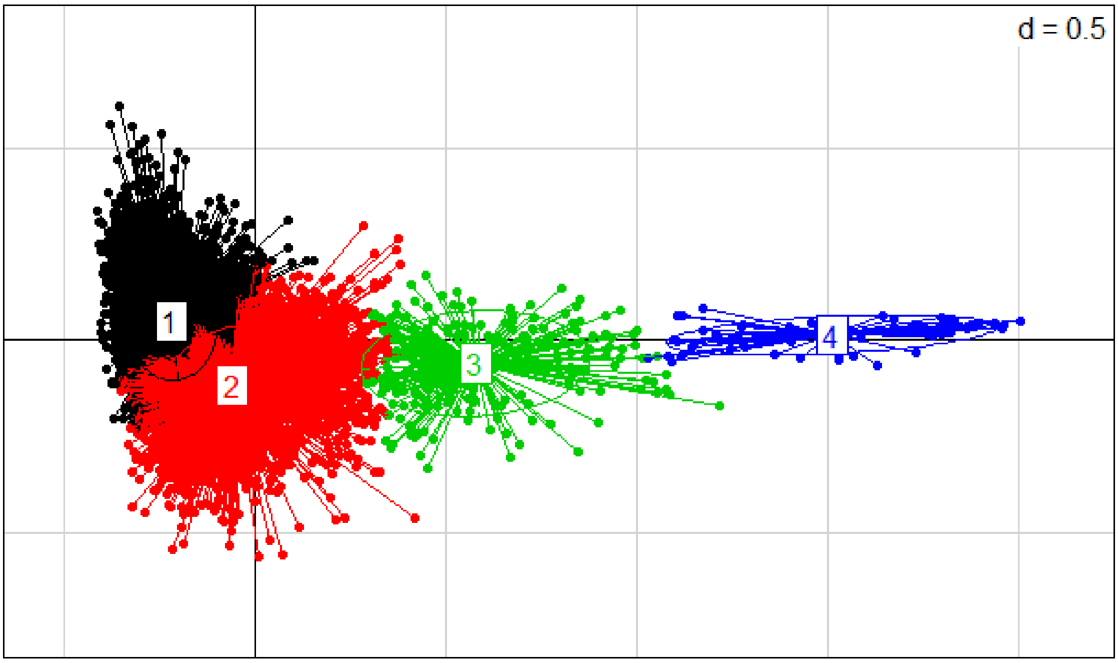


**Figure 4 : Plot of 4 clusters**

Now, we can see the 4 clusters. The numbers in the plot refers to:

1. Evolution of genes that represents the same evolution 7 days without drugs
2. Evolution of genes that represents the same evolution 7 days with drugs
3. Evolution of genes that represents the same evolution 14 days without drugs
4. Evolution of genes that represents the same evolution 14 days with drugs

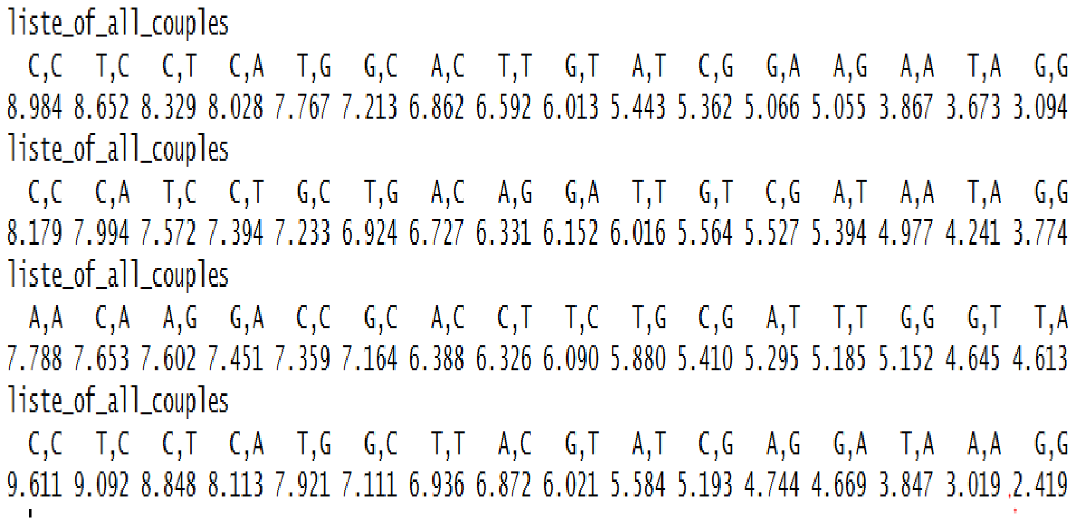
## Conclusion

We can assume now that the evolution of different genes with/without drugs can be conducted to 4 possible results. We extracted, from every cluster, genes that had the same evolution with and without drugs. Now, we need to proceed to the pattern recognition of genes present in the same cluster to be able to extract "rules" that had, eventually, reacted with the drugs or that do not have any effect from these drugs.

# Pattern recognition

In this part, we assume that we have 4 clusters of genes, and, for every cluster, genes sequences are available in our database. We would proceed to a pattern recognition to extract rules that drive genes to react or not to the drugs.

Our first approach in this case is to design a PCA, that permit to calculate the correlation degree between a couple of letters. The script is designed also with R programming language and available in the file "pca.R". It consists in splitting all gene sequences in the same cluster and then, calculate the correlation degree between them, this would reflect dependencies between different nucleic bases in the same cluster. Knowing that, for example 'A' and 'T' are highly correlated, we can proceed by finding patterns between the two nucleic bases. By executing the script, we got this result:



```
liste_of_all_couples
  C,C   T,C   C,T   C,A   T,G   G,C   A,C   T,T   G,T   A,T   C,G   G,A   A,G   A,A   T,A   G,G 
8.984 8.652 8.329 8.028 7.767 7.213 6.862 6.592 6.013 5.443 5.362 5.066 5.055 3.867 3.673 3.094 
liste_of_all_couples
  C,C   C,A   T,C   C,T   G,C   T,G   A,C   A,G   G,A   T,T   G,T   C,G   A,T   A,A   T,A   G,G 
8.179 7.994 7.572 7.394 7.233 6.924 6.727 6.331 6.152 6.016 5.564 5.527 5.394 4.977 4.241 3.774 
liste_of_all_couples
  A,A   C,A   A,G   G,A   C,C   G,C   A,C   C,T   T,C   T,G   C,G   A,T   T,T   G,G   G,T   T,A 
7.788 7.653 7.602 7.451 7.359 7.164 6.388 6.326 6.090 5.880 5.410 5.295 5.185 5.152 4.645 4.613 
liste_of_all_couples
  C,C   T,C   C,T   C,A   T,G   G,C   T,T   A,C   G,T   A,T   C,G   A,G   G,A   T,A   A,A   G,G 
9.611 9.092 8.848 8.113 7.921 7.111 6.936 6.872 6.021 5.584 5.193 4.744 4.669 3.847 3.019 2.419 
```

**Figure 5 : Correlation degree between two nucleic bases**

These results reflect that there is no "REAL" correlation between different nucleic bases in different clusters. So, depending on the sample of data that we have used, we did not found any correlation between them. This let us conclude that there is no dependency between two nucleic bases. This is also demonstrated by the next PCA figure that sows that all 4 nucleic bases present the same correlation.

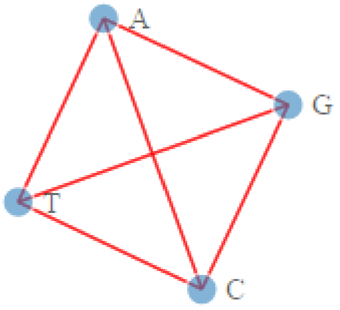
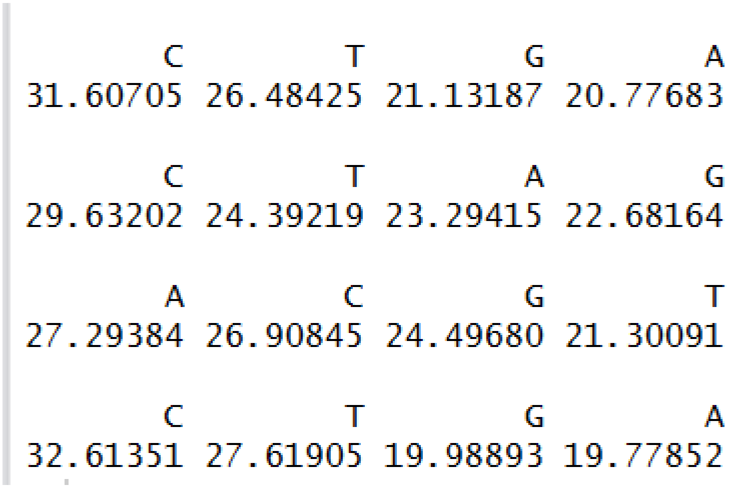


**Figure 6 : 2d Plot of correlation degree between nucleic bases**

## 2nd approach:

As the previous study showed as that there is no consisting result between couples of nucleic bases, we started have used the same approach to see the correlation between nucleic bases and not with a "couple" of nucleic bases. the script is available also in the pca.R file. After the execution of the script, we got this table:



```
       C        T        G        A
31.60705 26.48425 21.13187 20.77683

       C        T        A        G
29.63202 24.39219 23.29415 22.68164

       A        C        G        T
27.29384 26.90845 24.49680 21.30091

       C        T        G        A
32.61351 27.61905 19.98893 19.77852
```

**Figure 7 : correlation degree of single nucleic bases**

Based on this table, we cannot be affirmative but showed us that "C" Nucleic is more reactive in cluster "1", "2" and "4", corresponding to the clusters:

- Gene Evolution without drugs after 7 days
- Gene Evolution with drugs after 7 days
- Gene Evolution with drugs after 14 days

And another consideration can be noted, is that the "A" nucleic base has the biggest value in the cluster "3" and the least value in the cluster "4". That means that "A" nucleic base has, eventually, an impact in "Gene evolution after 14 days without drugs" and a "bad" impact in the "Gene evolution after 14 Days with drugs".

## 3rd Approach

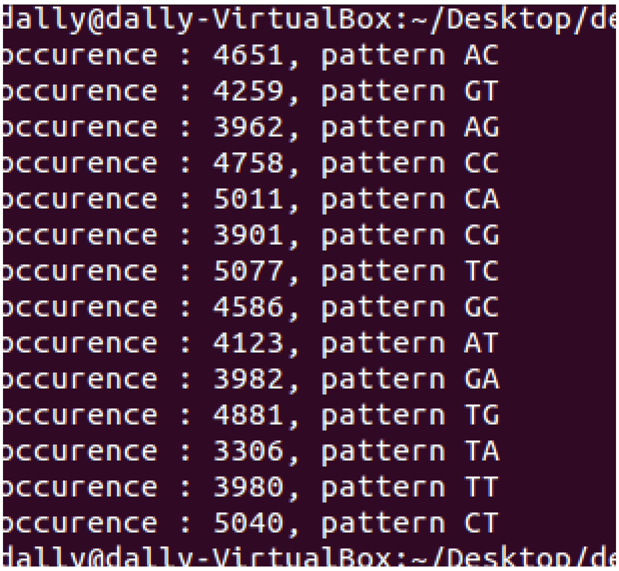
Conventional Data analysis methods did not mentioned any consisting patterns between genes in the same cluster. We need, at this point, to adapt "*a priori*[2]" algorithm to our problem and proceed to pattern recognition of our genes. This algorithm, from a transactional database, it permits to determine the nucleic bases that are most repeated in every cluster.

This module is developed with **python** and permit to generate 4 JSON files in the data folder. Every JSON file is structured in this way:

- Patterns founds in the cluster
- Number of occurrence of every pattern. This reflects how much the pattern has been repeated in genes sequences in a specific cluster.
- The places where these nucleic bases were found. It would let us know if the position of the pattern has an influence in the gene evolution with/without drugs.

[2] Wikipedia Definition of Apriori algorithm : en.wikipedia.org/wiki/Apriori_algorithm

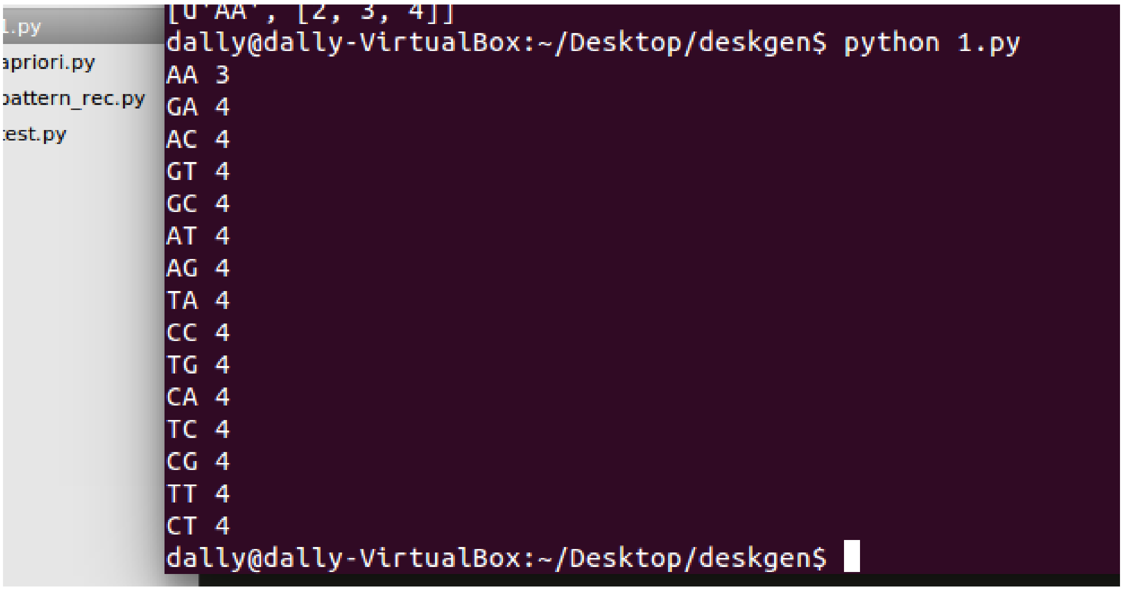
*Kindly note that these results are generated depending on our sample of 2000 gene from the .tsv file, so this analysis is not representative.*



```
dally@dally-VirtualBox:~/Desktop/de
occurence : 4651, pattern AC
occurence : 4259, pattern GT
occurence : 3962, pattern AG
occurence : 4758, pattern CC
occurence : 5011, pattern CA
occurence : 3901, pattern CG
occurence : 5077, pattern TC
occurence : 4586, pattern GC
occurence : 4123, pattern AT
occurence : 3982, pattern GA
occurence : 4881, pattern TG
occurence : 3306, pattern TA
occurence : 3980, pattern TT
occurence : 5040, pattern CT
dally@dally-VirtualBox:~/Desktop/de
```

**Figure 8 : returning pattern with number of occurence**

After generating patterns of our 4 clusters, we have discovered that these patterns are similar for all cluster except for the "AA" pattern that is available only for the cluster 2, 3 and 4. So we need to proceed with a study of the position of every pattern in the gene sequence. This would reflect us if the position of the pattern has an influence or these are only fuzzy data. The function clean_data() returns how much cluster do present a pattern.



```
[u'AA', [2, 3, 4]]
dally@dally-VirtualBox:~/Desktop/deskgen$ python 1.py
AA 3
GA 4
AC 4
GT 4
GC 4
AT 4
AG 4
TA 4
CC 4
TG 4
CA 4
TC 4
CG 4
TT 4
CT 4
dally@dally-VirtualBox:~/Desktop/deskgen$
```

**Figure 9 : Occurence of patterns in clusters**

Now, we have to test, for every pattern, the position in which it is present. Using Regular expression library with python, we have generated corresponding places of every pattern in a gene sequence.

After plotting different position of patterns with their corresponding average, we have seen that the majority of the data can be considered as a "fuzzy data" and cannot lead us to a concrete conclusion. Except for 3 rules data that can be considered as meaningful compared to other data. And all turns around the 4th cluster:

I - The fourth cluster contains 761 gene. We have discovered that 115 out of these 761 genes have "CC" subsequence in the 18 place

II - 111 genes out of 761 genes have "CT" subsequence in the 16th position in the gene sequence

III - Only 12 out of 761 genes have the subsequence "TA" in the 18th position of the gene sequence.

# Conclusion

We were able to cluster the data into 4 clusters representing:

- Genes that have the meaningful evolution in 7 days without drugs.
- Genes that have a meaningful evolution 14 days without drugs.
- Genes that have a meaningful evolution 7 days with drugs.
- Genes that have a meaningful evolution 14 days with drugs.

To proceed to this classification, we have taken a sample of 10000 genes out of the 64000 genes due to the performance of our computer.

After this classification, we have turned to PCA analysis to measure the correlation degree of different nucleic bases, in couple or in a simple, and this was not really concrete as we got only these information:

1. "C" nucleic base is the most involved in clusters 1,2 and 4. That means that genes that have the highest number of "C" nucleic base have a good chance to evolve in 7 or 14 days without drugs, or 14 days with drugs.
2. "A" nucleic base is the most involved in the cluster number 3 and "C" nucleic base do have the worst importance in this cluster, that means that "A" and "C" have an "opposite correlation". In other words, genes sequences with a high number of "A" and least number of "C" have a good chance to evolve in the cluster 3 : to have a high average in 7 days with drugs.

As PCA analysis was not consisting, we tried to find patterns in genes sequences using "apriori" algorithm, that consists in generating a tree of all possibilities, and to return only, sequences that have a probability (P (x>x0)). To have a meaningful analysis, x0 (the support) should be fixed to a minimum of 75%. But in our analysis, we have used 50% as it is only representative. After generating patterns, we have discovered that the subsequence "AA" is only present in clusters 2,3 and 4. We concluded that genes sequences with "AA" subsequence have a little chance to be present in the cluster 1, that means a little chance to evolve in 7 days without drugs.

For the other patterns, we proceeded to their analysis depending on their position. No real "Meaningful" or "consisting" rule has been deduced except for 3 rules that can be taken in consideration in the 4th cluster :

I - The fourth cluster contains 761 gene. We have discovered that 115 out of these 761 genes have "CC" subsequence in the 18 place

II - 111 genes out of 761 genes have "CT" subsequence in the 16th position in the gene sequence

III - Only 12 out of 761 genes have the subsequence "TA" in the 18th position of the gene sequence.

# CONCLUSION

In other words, gene with as sequence containing "CT" subsequence in the 16th position and having "CC" in the 18th position have the highest chance to evolve in 14 days with drugs. And genes that have "TA" subsequence in the 18th position do have a little chance to have a meaningful evolution after 14 days with drugs.