WORK EXPLORATION ON GENE SEQUENCING WITH SINGLE CELL

Qin He

272486
qin.he@tuni.fi

ABSTRACT

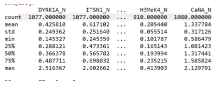
Analysis on single-cell usually combines the process of RNA sequences with ignal processing pipelines, including prepcrocessing with normalization, correction, feature selection, dimension reduction and transformation, cellular and genetic scale analysis. In this exploration, I started from the basic signal preprocessing skills conduction and later applied the clustering, PCA analysis, t-NSE to have a better understanding of the cellular data.Finally, more works are on basis of single cell sequencing techniques, on temporal scale, pseudotime analysis is analysed. Meanwhile, the prediction of the cells helping with the detection of cell-line of differentiation and reprogramming is also conducted with sigmoid regression.

1. PREPROCESSING

The processing of the genetic and cellular data is conducted on python. THe dataset are downloaded from some open database. FOr instance, the UCI online sites and UK gene bank. In this work, major work are conducted on the gene espressions, TF and micro-array data.

1.1. Data obtain and cleaning with descriptive statistics

As the genetic and cellular data are downloaded form the open database recorded by institutions wither clinically or with research aim in the lab. The error introduced by device and measurements cannot be ignored. THe first step as usual is to have a general understanding of its descriptive information. FOr instances, the dimensions,mean and variance and counts. Such descriptive statistics nowadays can be achieved easily using statistical functions. I used pandas as the tool to start the clean of the data. After replacing the NAN and removing some outliers, the data is again checked with desciptive statistics which leads to normalization and dimension reduction. TF is computed as well to track the gene expression over time and profile the cell fate of different cells from the ESC state. Usually, the predictivity of TFs cn be used as markers of a cell fate and are potential candidates for reprogramming protocols.



```
       DYRK1A_N     ITSN1_N  ...    H3MeK4_N       CaNA_N
count  1077.000000  1077.000000  ...  810.000000  1080.000000
mean      0.425810     0.617102  ...    0.205440     1.337784
std       0.249362     0.251640  ...    0.055514     0.317126
min       0.145327     0.245359  ...    0.101787     0.586479
25%       0.288121     0.473361  ...    0.165143     1.081423
50%       0.366378     0.565782  ...    0.193994     1.317441
75%       0.487711     0.698032  ...    0.235215     1.585824
max       2.516367     2.602662  ...    0.413903     2.129791
```

Fig. 1. Descriptive statistics

1.2. dimension reduction and feature selection

Genetic data are usually of high dimension since one groups of genes are usually related with not only one cell but also other cells and other groups of gene sequences, the reduction of the dimensions are required as a common process. For this work, the dimension reduction is conducted through the PCA analysis. As the statsitics under PCA is also based on variance, and the the choice of the principal component is computed with statistics which is also related to time-bins and fft process as well. In my work the analysis conducted combining cross interation analysis as well. DIfferent algorithm usually use different methods to decide the dimension and features for further process. As for some methods based on deterministant matrices, the important factors are kept with deterministant quatities and statistics. Of course, even for stocastics method, the reduce of the factor is also based on some statistics which is sees as deterministive however, realized with approximation, and estimation on distribution for instance. For my work, I also tried the tsne method to reduce the dimension and select the features with embedding networks. It is computed on the measure of energy and distances. (cmbining K-L distance and fourier transform.) Through setting the metrice as euclidean distance and random state included with 10 jobs for each batch, the tsne training is conducted after the PCA with embedding.

2. CLASSIFICATION AND CLUSTERING

After the PCA initialized with standard statistics, the tsne is conducted with embedding network. As the result, the 60000 training data is classified into 10 classes with supervised learning.(THe biomarkers of genes are also classified with differentiative private GANs.) Then, I also tried some basic method to cluster the data. For clustering, the learning is not based on labels(ground truth). More methods used the

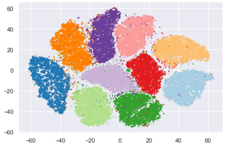


Fig. 2. training with tsne

computation of the distance as well. The first method I tried is the KNN which cluster the samples with the k-fold nearest criteria. Usually, according to the property of the dataset, the cluster can be decided by computation of mdistance with continual data and voting with separate data. However, the most important thing is to decide the cluster number.(THe starting point is not importnat for KNN though.) In my work, I use the WCSS which is also called the elbow method to decide the best cluster number. It is based on the gradient computation. THe minimum point is found through iteration of the gradient. As can be seen in the result, the best cluster umber is 4 and our cluster is quite fast which consumes less than 1second on the laptop. For the next cluster method

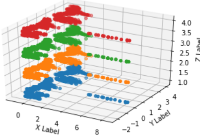


Fig. 3. KNN++

hierarchical clustering, the clustering is again based on the distance between two clusters with maximum clust criteria. And with the transcate of the last clustering, the distribution is shown as the dendrogram with 4 leaf nodes. The

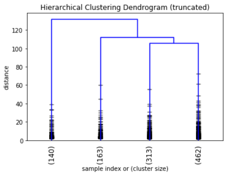


Fig. 4. hierarchical clustering

result show as the scatter plot also manifest the advantage of t-SNE which is a more advanced method since the genes are clustered into clusters not only based on distance but also other information. However, as the differentration and reprogramming of the single cell is usually studied on more fined level, which requires the biomarkers to be detected and thus their transcription factor can computed to decompose the genes into groups which can leads to the evolution of the cells later. THe biomarkers are classified with dp-GANs in my work. The encoding and decoding part of the GANs makes it unique in its competitive system. THe generator generates the data against the calssification of the descriminators while the discriminator tries it best to classify the data according to the ground truth. GANs is usally a supervised method in classification tasks. Unlike dc-GANs and other classifiers based on convolutional network, the convergence of the generator is realized through satisfying privacy budget.T weight computation is literally applied with differentiation method which is different from those stochastics model with iteration or approximation using distribution on random resampled data or information inference based on priors. The criteria is calculated analytically using differentiation knowledge. As a result, the biomarkers are classified into 4 classes.

CLustering Scatter$_{12}.png$

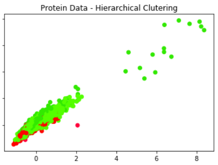
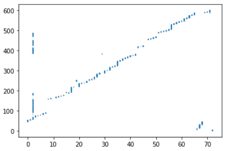


Fig. 5. hierarchical clustering

3. PREDICTION WITH SIGMOID REGRESSION

Last we predict the differentiation of ESC using 9 critical cells with more than 1700 genes involved. As the raw data records 64 cells in all using micro array, we conduct the feature selection with the TF matrix. The unrelated or outliers are exluded from computation first. And through the gradient search, we find the best TF dynamically with the micra-array data on pseudotime analysis. With the GSE and GSM coordinates recorded(Shown as the cell trajectory 2D), we analyse the expression of the cell and genes on pseudo time. As we further use the TF to regress the cell evolution, with the label given with 9 cells, at A1 cell line, we manage to detect the differentiation time and predict the TF of them dynamically with sigmoid function which maximizes the likelihood after logarithm. Through minimizing th loss-value, the gradient leads to minimum(although only local minimum is calculated) point and we successfully predict the 9 cell's expression and thus their TF at A1 cell line.



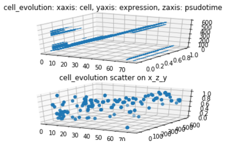


Fig. 7. evolution on pseudotime

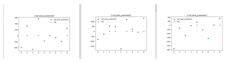
1-3.PNG 

Fig. 8. cell expression prediction

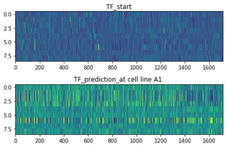


Fig. 9. TF predictivity

Using the TF ahieved with highest accuracy, the evolution of the 9 cells related to A1 cell line is analysized on pseudotime. Notice that, most of the genes are kept normalized distributed and thus we have our MLE matrices following chisquare distribution. ANd the whole processing is based on the gene data shuffled with permutation test as well. As our data achieved in lab is cultured on MEA and the research of interest is the epillepsy. I basically focus more on ESC and which express the MAP, TUBB3 and etc. Those cells data are usually sampled at cortex and hippocampus. ———

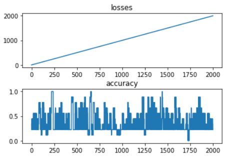
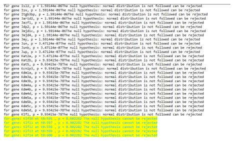


Fig. 10. prediction accuracy



```
for gene Isl2, p = 1.59144e-06The null hypothesis: normal distribution is not followed can be rejected
for gene Isx, p = 1.59144e-06The null hypothesis: normal distribution is not followed can be rejected
for gene Itgb3bp, p = 1.59144e-06The null hypothesis: normal distribution is not followed can be rejected
for gene Jarid2, p = 1.59144e-06The null hypothesis: normal distribution is not followed can be rejected
for gene Jazf1, p = 1.59144e-06The null hypothesis: normal distribution is not followed can be rejected
for gene Jdp2, p = 1.59144e-06The null hypothesis: normal distribution is not followed can be rejected
for gene Jmjd1c, p = 1.59144e-06The null hypothesis: normal distribution is not followed can be rejected
for gene Jmjd4, p = 1.59144e-06The null hypothesis: normal distribution is not followed can be rejected
for gene Jmy, p = 1.59144e-06The null hypothesis: normal distribution is not followed can be rejected
for gene Jun, p = 3.47124e-07The null hypothesis: normal distribution is not followed can be rejected
for gene Junb, p = 3.47124e-07The null hypothesis: normal distribution is not followed can be rejected
for gene Jup, p = 3.47124e-07The null hypothesis: normal distribution is not followed can be rejected
for gene Kat2a, p = 3.47124e-07The null hypothesis: normal distribution is not followed can be rejected
for gene Kat2b, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kat5, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kcnip3, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm1a, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm3a, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm3b, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm4a, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm4b, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm4d, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm5a, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm5b, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm5c, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Kdm5d, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene Klf1, p = 9.93415e-78The null hypothesis: normal distribution is not followed can be rejected
for gene: Klf10 at t0:596 , p = 0.902282 The null hypothesis cannot be rejected
for gene: Klf11 at t0:597 , p = 0.902282 The null hypothesis cannot be rejected
for gene: Klf12 at t0:598 , p = 0.902282 The null hypothesis cannot be rejected
for gene: Klf13 at t0:599 , p = 0.902282 The null hypothesis cannot be rejected
for gene: Klf14 at t0:600 , p = 0.902282 The null hypothesis cannot be rejected
```

Fig. 11. distribution of the genes

-

4. SUMMARY

TO summarize, the analysis on single cell can be conducted combined with machine learning and signal processing techniques well. THrough preprocessing, we get better distributed data with lower dimention and more diterminant features. THus, the differentiation. classification and reprogramming can be conducted better later on either pseudotime analysis or expression prediction. The TF is one useful quantity defining the cell evolution with its relation between time, space and gene encoding information. In the future, I would like to apply more analysis on pseudo time and explore more analysis techniques with single cell RNA sequencing maybe. Of course, machine learning will always be of one useful assistant tool. As the landcape modle describes the encodng of differentiation and reprogramming using only four TFs, I would like to explore more on realted model in the future and try the prediction and detection on 4 TF only. Althouygh the cell of my interest in related to ESC rather than NSC and NPC. ————————

REFERENCES

[1]Quantifying the Wddinton landscape and biologica paths for developing and differentiation

[2] Human embryonic stem cell-derived neural cells from spontaneously active neuronal network in vitro.