Diary

**2019.04.05**

In the previous study, only CAGE and Microarray can catch the downregulation type transcription factors. (e.g. GATA6 vs CAV1)

* Gene name should be converted and uniform the method name.
* To do:

Understand the code from Sara.

Expand the list.

Normalization: ZScore, Log and Log+ZScore.

* Figure with times catch in the list from 3 dataset and the deviation.

**2019.04.23**

The problem in the code:

The function "*fromsymboltoentrez*" is able to get the gene Entrez ID from the NCBI website. MATLAB connects to the NCBI server by function *webread.*

* Problem: The function *webread*got the message error 502 due to disconnection with NCBI server and then code stopped every time when error occurred.
* Solution: Aldo taught me to use function *try/catch* to solve the stop problem. And then tried to reconnect to the server three times if the server return error. If it still cannot connect to the server more than three times, then we set the Entrez ID as NaN

**2019.04.29**

The Gene ID was in the second column of *cage\_peaks* in the *FANTOM5\_CAGE* dataset. There is no need to convert chromosome locus to gene symbol. Based on Sara’s example (DKK1 and GATA6), I created the function for CAGE dataset.

The summary of function:

First it compares the TF gene and regulated gene of the list with genes in the *FANTOM5\_CAGE* dataset row-by-row in the list. If both TF gene and regulated gene are in the *FANTOM\_CAGE*, the function compute the Pearson correlation between TF gene and regulated gene. If there are more than one peaks associated to the gene, the function gets the mean across the peaks and then calculates the correlation between TF gene and regulated gene.

The result showed:

54.39 % caught by microarray in downregulation type.

46.15% caught by RNA-seq in downregulation type.

44.00 % caught by CAGE in downregulation type.

64.13 % caught by microarray in upregulation type.

42.17% caught by RNA-seq in upregulation type.

51.39 % caught by CAGE in downregulation type

Carlo:

* C.C. Carlo when sending the mail.
* For the normalization part, I should discuss with Claudio first.

1. Create the function.
2. Set the normalization option in the function.
3. Storage the result.

**2019.05.03**

Sara got the result of true positive rate for CCLE- microarray and Atlas- RNA-seq from LOG then ZSCORE.

True positive rate of CCLE-Microarray

Upregulation: 62.0%

Downregulation: 52.6%

True positive rate of Atlas RNA-seq

Upregulation: 56.6

Downregulation: 34.6%

I got the result of CAGE with and without normalization.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| True positive rate of different normalizations. | | | | |
| Normalization | Type of cell | Dataset | Upregulation | Downregulation |
| without | Cancer | Atlas RNA-seq | 0.4211 | 0.4884 |
| CCLE-Microarray | 0.6353 | 0.5833 |
| CAGE | 0.5358 | 0.4634 |
| ZSCORE | Cancer | Atlas RNA-seq | 0.5132 | 0.5814 |
| CCLE-Microarray | 0.6235 | 0.5833 |
| CAGE | 0.6000 | 0.4146 |
| LOG | Cancer | Atlas RNA-seq | 0.6184 | 0.3023 |
| CCLE-Microarray | 0.6471 | 0.6042 |
| CAGE | 0.5385 | 0.4634 |
| LOG ->  ZSCORE | Cancer | Atlas RNA-seq | 0.5658 | 0.3488 |
| CCLE-Microarray | 0.6118 | 0.5625 |
| CAGE | 0.5321 | 0.4390 |

Carlo:

* Need to calculate: sensitivity, specificity, F1 score, AUC, AUPR, precision-up and precision-down.
* Calculate the average across all peaks and save it as matrix.
* What is the type of three datasets (cancer?)
* Find another dataset from human, will do the animal datasets as well.
* In this analysis we check the sign of correlation. Check correlation values later.
* Check the overlap of true positive transcription factor genes across different type of datasets (normal vs cancer).
* Title: Transcription factor gene regulatory interaction confirmed by multi-technologies genome wide expression analysis (or investigation).

**2019.05.07**

Q: There are different results between:

Normalize then mean.

Mean then normalize.

In CAGE dataset.

A: Do the both analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| True positive rate of different normalizations. | | | | |
| Dataset | Type of cell | Normalization | Upregulation | Downregulation |
| Atlas RNA-seq | Cancer | Without | 0.4211 | 0.4884 |
| Zscore | 0.5132 | 0.5814 |
| Log | 0.6184 | 0.3023 |
| Log -> Zscore | 0.5658 | 0.3488 |
| CCLE-Mircroarray | Cancer | Without | 0.6353 | 0.5833 |
| Zscore | 0.6235 | 0.5833 |
| Log | 0.6471 | 0.6042 |
| Log -> Zscore | 0.6118 | 0.5625 |
| CAGE  (Normalize first) | Cancer | Without | 0.4762 | 0.4634 |
| Zscore | 0.6190 | 0.4146 |
| Log | 0.4921 | 0.4390 |
| Log -> Zscore | 0.4762 | 0.4390 |
| CAGE  (Mean first) | Cancer | Zscore | 0.6190 | 0.3171 |
| Log | 0.4762 | 0.4390 |
| Log -> Zscore | 0.4921 | 0.3659 |

Sensitivity: true positive rate of upregulation/downregulation.

Specificity: true positive rate of downregulation/upregulation.

**2019.05.09**

For the AUC-ROC, Claudio taught me to use the true positive rate (TPR) and false positive rate (FPR=1-specificity) for AUC calculating.

In MATLAB:

y = [0; TPR; 1];

x = [0; FPR; 1];

trapz(y,x);

**2019.05.10**

I finished to calculate the statistic value including sensitivity, F1 score, AUC, precision-positive and precision negative (please see the attached file). For the AUC, Claudio taught me using the true positive rate and false positive rate to plot and integrate the area. The RNA seq and CAGE did not have good separability in up- and downregulation. In this result the positive and negative are just based on sign of correlation, maybe we should try to filter by the value of correlation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | F1 score | AUC | Precision-  positive | Precision-negative |
| CCLE  microarray | Without | Down | 0.5833 | 0.5234 | 0.6093 | 0.4746 | 0.7297 |
| Up | 0.6353 | 0.6792 | 0.7297 | 0.4746 |
| Zscore | Down | 0.5833 | 0.5158 | 0.6034 | 0.4667 | 0.7260 |
| Up | 0.6235 | 0.6709 | 0.7260 | 0.4667 |
| Log | Down | 0.6042 | 0.5421 | 0.6256 | 0.4915 | 0.7432 |
| Up | 0.6471 | 0.6918 | 0.7432 | 0.4915 |
| Log + Zscore | Down | 0.5625 | 0.5000 | 0.5871 | 0.4500 | 0.7123 |
| Up | 0.6118 | 0.6582 | 0.7123 | 0.4500 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | F1 score | AUC | Precision-positive | Precision-negative |
| Atlas  RNAseq | Without | Down | 0.4884 | 0.3889 | 0.5453 | 0.3231 | 0.5926 |
| Up | 0.4211 | 0.4923 | 0.5926 | 0.3231 |
| Zscore | Down | 0.5814 | 0.4762 | 0.5473 | 0.4032 | 0.6842 |
| Up | 0.5132 | 0.5865 | 0.6842 | 0.4032 |
| Log | Down | 0.3023 | 0.3059 | 0.5396 | 0.3095 | 0.6104 |
| Up | 0.6184 | 0.6144 | 0.6104 | 0.3095 |
| Log + Zscore | Down | 0.3488 | 0.3297 | 0.5427 | 0.3125 | 0.6056 |
| Up | 0.5658 | 0.5850 | 0.6056 | 0.3125 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | F1 score | AUC | Precision-positive | Precision-negative |
| FANTOM5CAGE | Without | Down | 0.4634 | 0.4086 | 0.5302 | 0.3654 | 0.5769 |
| Up | 0.4762 | 0.5217 | 0.5769 | 0.3654 |
| Zscore | Down | 0.3171 | 0.3333 | 0.5319 | 0.3514 | 0.5821 |
| Up | 0.6190 | 0.6000 | 0.5821 | 0.3514 |
| Log | Down | 0.4390 | 0.3913 | 0.5424 | 0.3529 | 0.5660 |
| Up | 0.4762 | 0.5172 | 0.5660 | 0.3529 |
| Log + Zscore | Down | 0.3659 | 0.3409 | 0.5710 | 0.3191 | 0.5439 |
| Up | 0.4921 | 0.5167 | 0.5439 | 0.3191 |

**2019.05.13**

Carlo:

* Compare the sign of Pearson correlation, Spearman correlation and distance correlation with the list.
* Compare the values of Pearson correlation, Spearman correlation and distance correlation which are higher than 0.5 or lower than -0.5 with the list.

**2019.05.14**

Dear Carlo,

We discussed on Monday:

* Compare sign from Pearson correlation, Spearman correlation and distance correlation with the regulation type in the list.
* Compare the values of Pearson correlation, Spearman correlation and distance correlation which are higher than 0.5 or lower than -0.5 with the regulation type in the list.
* Redo the Table.

I did the results for sign of Pearson correlation and the value of Pearson correlation (attached file). The threshold of the value is higher than 0.5 or lower than -0.5. The positive class for sensitivity, specificity and AUC is upregulation. Notably in threshold cases, there are just few passing the threshold (6 in upregulation of microarray and 1 in downregulation of microarray). In the downregulation, only one passes the threshold and its value is positive, so the specificity is zero. In the upregulation, few are higher than threshold but the correlation value are all positive. The sensitivity is 1. Because the specificity and false negative are zero, some values in F1 score and precision are no value.

**Sign (Pearson correlation)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| Microarray | Without | Upregulation (+) | 0.6353 | 0.5833 | 0.6792 | 0.6093 | 0.7297 |
| Downregulation (-) | 0.5234 | 0.4746 |
| Zscore | Upregulation (+) | 0.6235 | 0.5833 | 0.6709 | 0.6034 | 0.7260 |
| Downregulation (-) | 0.5185 | 0.4667 |
| Log | Upregulation (+) | 0.6471 | 0.6042 | 0.6918 | 0.6256 | 0.7432 |
| Downregulation (-) | 0.5421 | 0.4915 |
| Log+Zscore | Upregulation (+) | 0.6118 | 0.5625 | 0.6582 | 0.5871 | 0.7123 |
| Downregulation (-) | 0.5000 | 0.4500 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| RNAseq | Without | Upregulation (+) | 0.4211 | 0.4884 | 0.4923 | 0.5453 | 0.5926 |
| Downregulation (-) | 0.3889 | 0.3231 |
| Zscore | Upregulation (+) | 0.5132 | 0.5814 | 0.5865 | 0.5473 | 0.6842 |
| Downregulation (-) | 0.4762 | 0.4032 |
| Log | Upregulation (+) | 0.6184 | 0.3023 | 0.6144 | 0.5396 | 0.6104 |
| Downregulation (-) | 0.3059 | 0.3095 |
| Log+Zscore | Upregulation (+) | 0.5658 | 0.3488 | 0.5850 | 0.5427 | 0.6056 |
| Downregulation (-) | 0.3297 | 0.3125 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| FANTOM5  CAGE | Without | Upregulation (+) | 0.4762 | 0.4634 | 0.5217 | 0.5302 | 0.5769 |
| Downregulation (-) | 0.4068 | 0.3654 |
| Zscore | Upregulation (+) | 0.6190 | 0.3717 | 0.6000 | 0.5319 | 0.5821 |
| Downregulation (-) | 0.3333 | 0.3514 |
| Log | Upregulation (+) | 0.4762 | 0.4390 | 0.5172 | 0.5424 | 0.5660 |
| Downregulation (-) | 0.3913 | 0.3529 |
| Log+Zscore | Upregulation (+) | 0.4921 | 0.3659 | 0.5167 | 0.5710 | 0.5439 |
| Downregulation (-) | 0.3409 | 0.3191 |

**Correlation value (Pearson correlation)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| Microarray | Without | Upregulation (+) | 1 | 0 | 0.9231 | 0.5000 | 0.8571 |
| Downregulation (-) | NaN | NaN |
| Zscore | Upregulation (+) | 1 | 0 | 0.9231 | 0.5000 | 0.8571 |
| Downregulation (-) | NaN | NaN |
| Log | Upregulation (+) | 1 | 0 | 0.9231 | 0.5000 | 0.8571 |
| Downregulation (-) | NaN | NaN |
| Log+Zscore | Upregulation (+) | 1 | 0 | 0.9231 | 0.5000 | 0.8571 |
| Downregulation (-) | NaN | NaN |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| RNAseq | Without | Upregulation (+) | 1 | 0 | 0.8571 | 0.5000 | 0.7500 |
| Downregulation (-) | NaN | NaN |
| Zscore | Upregulation (+) | 1 | 0 | 0.9091 | 0.5000 | 0.8333 |
| Downregulation (-) | NaN | NaN |
| Log | Upregulation (+) | 1 | 0 | 0.9091 | 0.5000 | 0.8333 |
| Downregulation (-) | NaN | NaN |
| Log+Zscore | Upregulation (+) | 1 | 0 | 0.8889 | 0.5000 | 0.8000 |
| Downregulation (-) | NaN | NaN |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| FANTOM5  CAGE | Without | Upregulation (+) | 1 | 0 | 0.8000 | 0.5000 | 0.6667 |
| Downregulation (-) | NaN | NaN |
| Zscore | Upregulation (+) | 1 | 0 | 0.6909 | 0.5000 | 0.5278 |
| Downregulation (-) | NaN | NaN |
| Log | Upregulation (+) | NaN | 1 | NaN | NaN | NaN |
| Downregulation (-) | 1.000 | 1 |
| Log+Zscore | Upregulation (+) | 1 | 1 | 1 | 1 | 1 |
| Downregulation (-) | 1 | 1 |

**Sign (Spearman correlation)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| Microarray | Without | Upregulation (+) | 0.6353 | 0.5417 | 0.6708 | 0.5885 | 0.7105 |
| Downregulation (-) | 0.4952 | 0.4561 |
| Zscore | Upregulation (+) | 0.6235 | 0.5717 | 0.6625 | 0.5826 | 0.7067 |
| Downregulation (-) | 0.4906 | 0.4483 |
| Log | Upregulation (+) | 0.6353 | 0.5717 | 0.6708 | 0.5885 | 0.7105 |
| Downregulation (-) | 0.4952 | 0.4561 |
| Log+Zscore | Upregulation (+) | 0.6353 | 0.5417 | 0.6708 | 0.5885 | 0.7105 |
| Downregulation (-) | 0.4952 | 0.4561 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| RNAseq | Without | Upregulation (+) | 0.6184 | 0.3023 | 0.6144 | 0.5396 | 0.6104 |
| Downregulation (-) | 0.3059 | 0.3095 |
| Zscore | Upregulation (+) | 0.6316 | 0.6774 | 0.6957 | 0.6530 | 0.7742 |
| Downregulation (-) | 0.5800 | 0.5088 |
| Log | Upregulation (+) | 0.6184 | 0.3023 | 0.6144 | 0.5396 | 0.6104 |
| Downregulation (-) | 0.3059 | 0.3095 |
| Log+Zscore | Upregulation (+) | 0.5526 | 0.3721 | 0.5793 | 0.5376 | 0.6087 |
| Downregulation (-) | 0.3441 | 0.3200 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| FANTOM5  CAGE | Without | Upregulation (+) | 0.5079 | 0.4146 | 0.5378 | 0.5387 | 0.5714 |
| Downregulation (-) | 0.3820 | 0.3542 |
| Zscore | Upregulation (+) | 0.6984 | 0.3415 | 0.6567 | 0.5199 | 0.6197 |
| Downregulation (-) | 0.3478 | 0.4242 |
| Log | Upregulation (+) | 0.5079 | 0.4146 | 0.5378 | 0.5387 | 0.5714 |
| Downregulation (-) | 0.3820 | 0.3542 |
| Log+Zscore | Upregulation (+) | 0.4603 | 0.3415 | 0.4874 | 0.5991 | 0.5179 |
| Downregulation (-) | 0.3146 | 0.2917 |

**Correlation value (Spearman correlation)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| Microarray | Without | Upregulation (+) | NaN | NaN | NaN | NaN | NaN |
| Downregulation (-) | NaN | NaN |
| Zscore | Upregulation (+) | NaN | NaN | NaN | NaN | NaN |
| Downregulation (-) | NaN | NaN |
| Log | Upregulation (+) | NaN | NaN | NaN | NaN | NaN |
| Downregulation (-) | NaN | NaN |
| Log+Zscore | Upregulation (+) | NaN | NaN | NaN | NaN | NaN |
| Downregulation (-) | NaN | NaN |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| RNAseq | Without | Upregulation (+) | 1 | 0 | 0.6667 | 0.5000 | 0.5000 |
| Downregulation (-) | NaN | NaN |
| Zscore | Upregulation (+) | 1 | 0 | 0.9091 | 0.5000 | 0.8333 |
| Downregulation (-) | NaN | NaN |
| Log | Upregulation (+) | 1 | 0 | 0.6667 | 0.5000 | 0.5000 |
| Downregulation (-) | NaN | NaN |
| Log+Zscore | Upregulation (+) | 1 | NaN | 1 | NaN | 1 |
| Downregulation (-) | NaN | NaN |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Normalization | Regulation type | Sensitivity | Specificity | F1 score | AUC | Precision |
| FANTOM5  CAGE | Without | Upregulation (+) | NaN | 1 | NaN | NaN | NaN |
| Downregulation (-) | 1 | 1 |
| Zscore | Upregulation (+) | 0.9600 | 0 | 0.7059 | 0.5200 | 0.5581 |
| Downregulation (-) | NaN | 0 |
| Log | Upregulation (+) | NaN | 1 | NaN | NaN | NaN |
| Downregulation (-) | 1 | 1 |
| Log+Zscore | Upregulation (+) | 1 | NaN | 1 | NaN | 1 |
| Downregulation (-) | NaN | NaN |

**2019.05.24**

* Q: The distance correlation value is between 0 and 1. How could I know the negative correlation and positive correlation from the value.

REPLY: I did not remember this. Please check by yourself with the help of Claudio and let me know the solution. I am curious too to know the reason and how to solve it.

Dear Carlo,

Master Project

* I found that the RNA-seq dataset has same gene name with different expression value when I run the distance correlation. I made the average of repetitive gene as I did in CAGE dataset.
* I discussed the implement of permutation test in Pearson and Spearman correlation with Claudio: Pearson and Spearman correlations have positive and negative value, the permutation test code should be modified from the distance correlation.
* I finished to run the Pearson and Spearman correlation with the permutation test threshold (p value < 0.05). The distance correlation problem need to be solved.

Carlo:

* Determine the sign from same order subtraction and do the average of the sign.
* Check the code how the value is always positive.
* Check the approach to construct the regulatory network from expression data.
  + Dream challenge
  + Andrea Califano (ARACNe)

2019.05.28

REPLY:

For instance, did you notice that the 0.05 cut-off approach significantly improve the prediction of up-regulation but reduce the prediction of down-regulation?

Please, check that there is not any mistake in what you perform to build the null model.

For instance, you need to create two different null models, one for the negative and another for the positive interactions and each of them should be used independently to define the 0.05 cut-off.