Gene Signature Discovery

Timothy Fisher (Author), Huiwen Ju

CSC 8630 Project Report

April 30, 2019

1 Goal

Based on microRNA, mRNA quantification data and other clinical data of the patients with a breast cancer subtype, we want to build a model by machine learning to predict 5-year survival rate and tumor stage. If those labels cannot be predicted accurately, we want to do a data mining by clustering to find out if the data can be classified into other subclasses and the genes that distinguish the subclasses. Furthermore, we want to find out the biological pathways those genes are involved in and make hypotheses about microRNA-mRNA pair signature.

2 Data Preprocessing

After the data is downloaded from TCGA database [1], the gene data and the label data are in different places (Fig. 1). By building a dictionary to link a patient's case ID, race, 5-year survival, tumor stage, microRNA filename and mRNA filename, we extract the useful data and put them into a .csv file (Fig. 2). Also, we use 5-year survival label determined from the clinical data instead of the survival label already existing in the data to exclude two types of cases: 1. dead when the data is collected but survived more than 5 years after diagnosis, 2. alive when the data is collected but has not survived more than 5 years.

The label distribution within the data (Fig. 3) is biased, i.e., more alive than dead, more tumor stage iii than other stages and more white than other races.

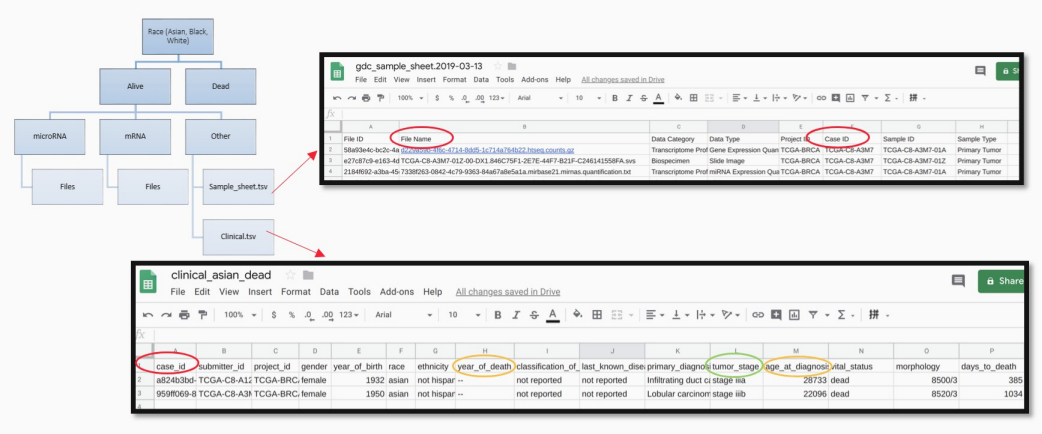


Fig. 1: Raw data.

	A	B	C	D	E	F	G	H
1		race	5-yr survival	tumor stage	microRNA1	...	mRNA1	...
2	sample 1							
3	sample 2							
4	...							

Fig. 2: Preprocessed data.

```
tot_stage: 335
tot_surv: 335
count_alive: 298, 0.89
count_dead: 37, 0.11
count_stage i: 63, 0.19
count_stage ii: 178, 0.53
count_stage iii: 81, 0.24
count_stage iv: 13, 0.04
count_asian: 10, 0.03
coun_black: 50, 0.15
count_white: 275, 0.82
```

Fig. 3: Label distribution (after comma: ratio).

3 Results

3.1 Predict 5-year survival and tumor stage

We try various types of machine learning models, including logistic regression, SVM, random forest, neural network and naive Bayes, to predict 5-year

survival (Fig. 4) and tumor stage (Fig. 5). 75% and 25% data are used as training and testing data, respectively. We use normalized confusion matrix to evaluate the prediction quality, and the quality is bad.

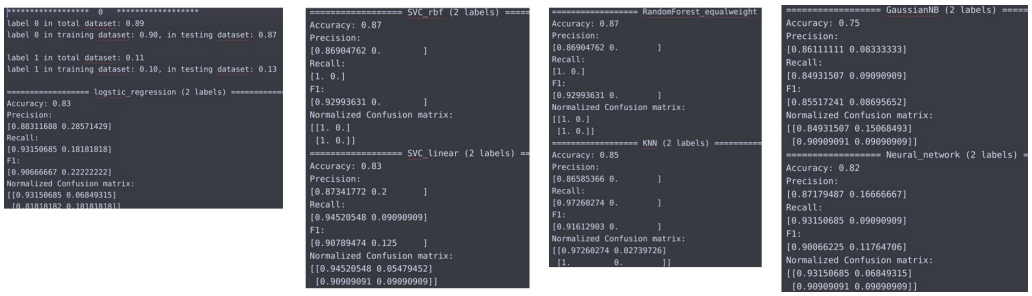


Fig. 4: Machine learning results for 5-year survival prediction.

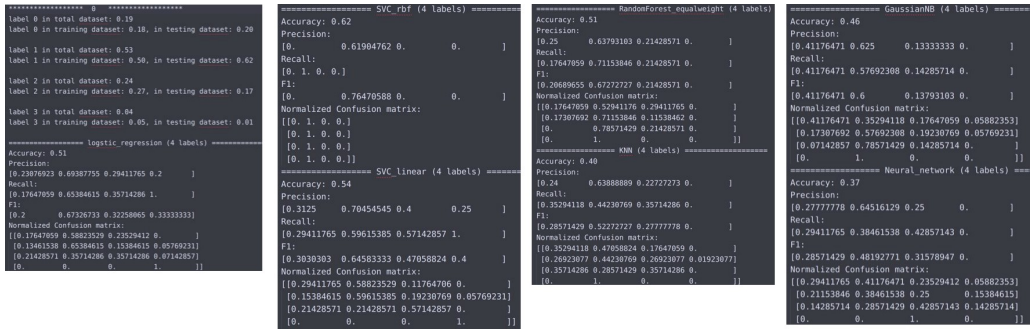


Fig. 5: Machine learning results for tumor stage prediction.

3.2 PCA

Then we apply PCA to the gene quantification data and get the variance each principle component accounts for (Fig. 6). We find that the top 2 or 3 principle components cannot separate the data according to the three types of label (Fig. 7).

We also run machine learning trainings on the principle components accounting for 50%, 90% and 98% total variance, respectively, to predict the three types of labels and still get poor results (Fig. 8). Results with normalized variance for each gene counting are poor, too.

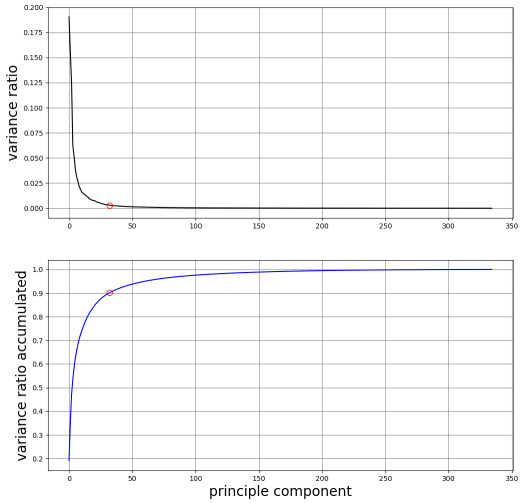


Fig. 6: Variance each principle component contributes to (up). Accumulated variance the principle components contribute to (down).

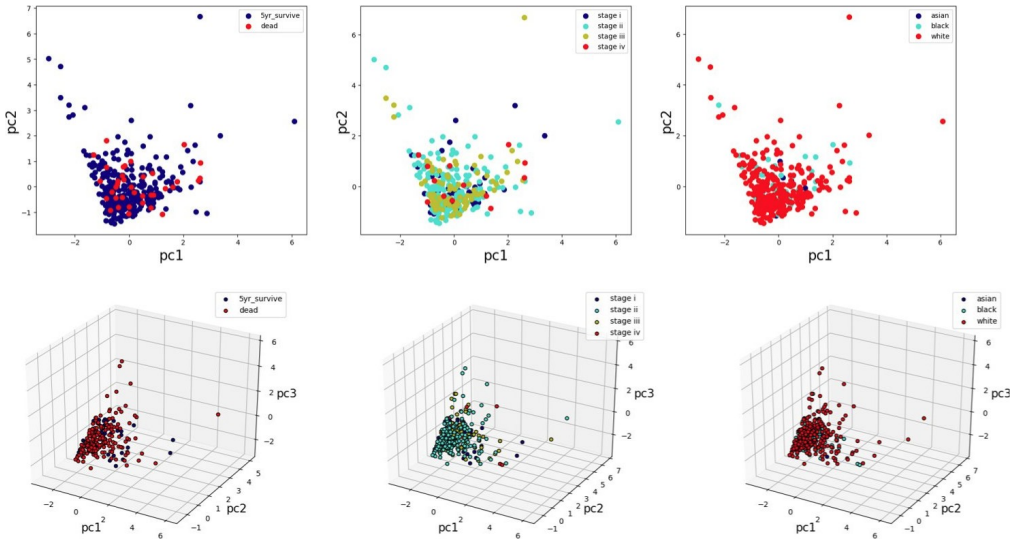


Fig. 7: The first 2 (up) or 3 (down) principle components cannot separate the data according to the label: 5-year survival (left), tumor stage (middle) and race (right).

3.3 Clustering

We then run k-means clustering on the whole data using the number of clusters 2-8. After that, machine learning methods are applied on the dataset

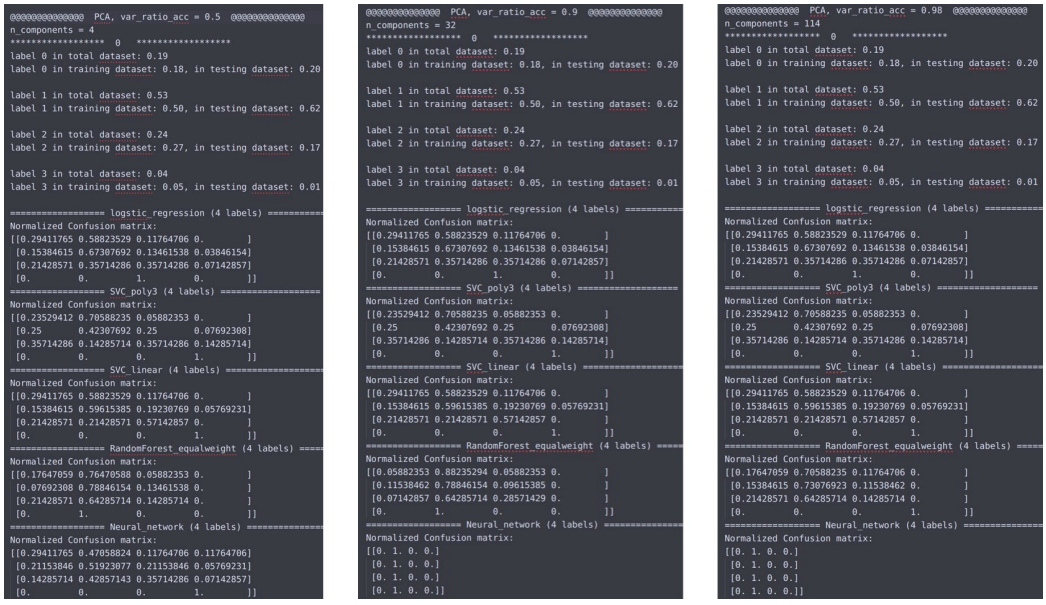


Fig. 8: Machine learning based on principle components accounting for 50%, 90% and 98% total variance (from left to right) to predict tumor stage.

with clustering-generated labels. SVM with linear kernel always gives accurate prediction for 2 to 8 clusters (Fig. 9).

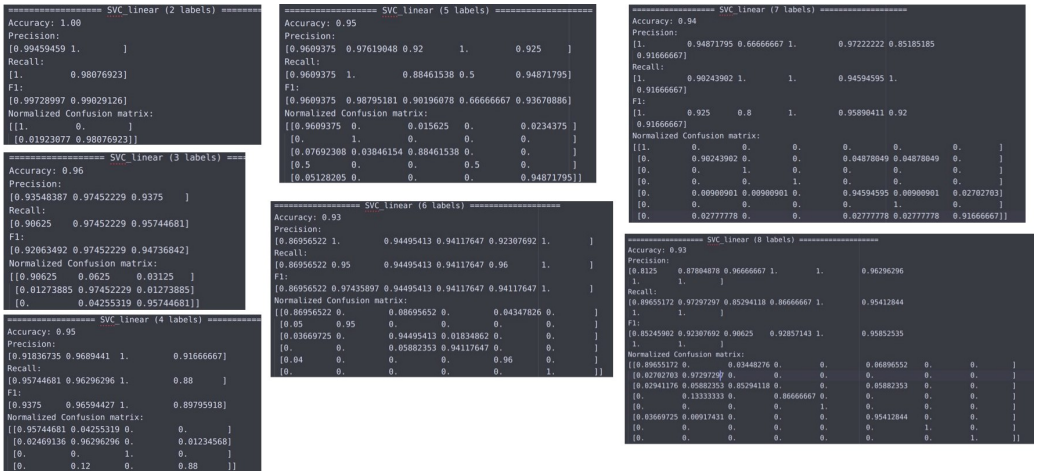


Fig. 9: SVM with linear kernel trained on labels generated by clustering with 2-8 clusters shows accurate prediction.

Next, we try to find out the genes that are likely to play a role in distinguishing each cluster pair. Clusters having less than 5 samples are excluded from this analysis since they have too little statistical power. First, we distill the genes that are differentially expressed between each pair of clusters. To quantify the difference of expression level, we use two methods. One is the ratio of the distance between the two means and the sum of the two standard deviation:

$$r = (mean1 - mean2)/(std1 + std2). \quad (1)$$

The other is as calculating the t-score, using the pooled standard deviation instead to reduce the effect of sample size:

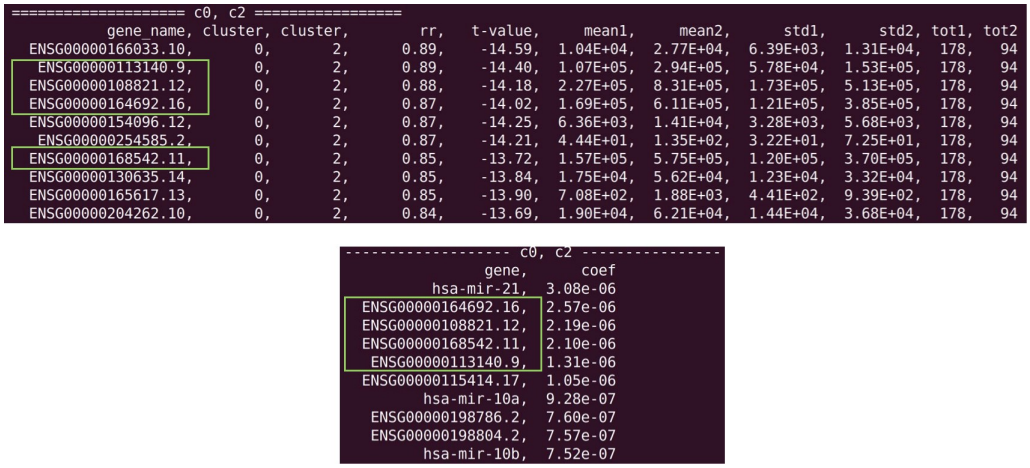
$$t = (mean1 - mean2)/std_{pooled}. \quad (2)$$

We find that the r and t are highly positively correlated. The genes are sorted by r in decreasing direction for each pair of clusters and the top 10 are extracted (Fig. 10 upper panel).

Second, we extract the 10 genes whose coefficients in the SVM linear model are the largest, indicating that those genes play important roles in classifying the cluster pair (Fig. 10 lower panel). For each cluster pair, comparing the genes extracted by the ratio and by the SVM coefficient, we find that some of the genes overlap (Fig. 10 green boxes) and we save the overlapping protein-encoding genes for further analysis (Fig. 11), assuming those might be the most critical genes in separating the cluster pair.

Then the overlapping gene set for each cluster pair is analyzed by Gene Ontology (GO) term enrichment technique. This technique can find the most significant pathways the input gene set is involved in. For each clustering, there is always a cluster pair (for clustering with 8 clusters, there are 2 pairs) whose overlapping gene set contains some specific genes (ENSG00000210082, ENSG00000198886, ENSG00000198888, ENSG00000198899, ENSG0000019-8804, ENSG00000198763) and it is always significantly involved in several pathways (negative regulation of signaling receptor activity, NADH dehydrogenase activity, response to hyperoxia , proton transmembrane transporter activity, reactive oxygen species metabolic process). Very likely, a cluster pair is discovered in the clustering (2-8 clusters) repeatedly. However, the overlapping gene sets for other cluster pairs are not shown to be significantly involved in any pathways.

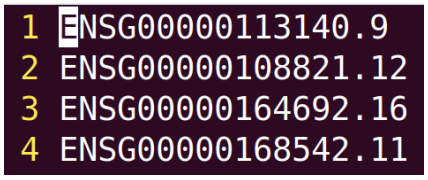
To see whether those pathways might relate to the labels, we look at the label distribution (i.e. race, tumor stage and 5-year survival) in that



```
================== c0, c2 ================
          gene_name, cluster, cluster,        rr,   t-value,      mean1,      mean2,       std1,       std2, tot1, tot2
 ENSG00000166033.10,       0,       2,      0.89,    -14.59,  1.04E+04,  2.77E+04,  6.39E+03,  1.31E+04,  178,   94
  ENSG00000113140.9,       0,       2,      0.89,    -14.40,  1.07E+05,  2.94E+05,  5.78E+04,  1.53E+05,  178,   94
 ENSG00000108821.12,       0,       2,      0.88,    -14.18,  2.27E+05,  8.31E+05,  1.73E+05,  5.13E+05,  178,   94
 ENSG00000164692.16,       0,       2,      0.87,    -14.02,  1.69E+05,  6.11E+05,  1.21E+05,  3.85E+05,  178,   94
 ENSG00000154096.12,       0,       2,      0.87,    -14.25,  6.36E+03,  1.41E+04,  3.28E+03,  5.68E+03,  178,   94
  ENSG00000254585.2,       0,       2,      0.87,    -14.21,  4.44E+01,  1.35E+02,  3.22E+01,  7.25E+01,  178,   94
 ENSG00000168542.11,       0,       2,      0.85,    -13.72,  1.57E+05,  5.75E+05,  1.20E+05,  3.70E+05,  178,   94
 ENSG00000130635.14,       0,       2,      0.85,    -13.84,  1.75E+04,  5.62E+04,  1.23E+04,  3.32E+04,  178,   94
 ENSG00000165617.13,       0,       2,      0.85,    -13.90,  7.08E+02,  1.88E+03,  4.41E+02,  9.39E+02,  178,   94
 ENSG00000204262.10,       0,       2,      0.84,    -13.69,  1.90E+04,  6.21E+04,  1.44E+04,  3.68E+04,  178,   94

------------------- c0, c2 ----------------
                gene,      coef
         hsa-mir-21,  3.08e-06
 ENSG00000164692.16,  2.57e-06
 ENSG00000108821.12,  2.19e-06
 ENSG00000168542.11,  2.10e-06
  ENSG00000113140.9,  1.31e-06
 ENSG00000115414.17,  1.05e-06
        hsa-mir-10a,  9.28e-07
  ENSG00000198786.2,  7.60e-07
  ENSG00000198804.2,  7.57e-07
        hsa-mir-10b,  7.52e-07
```

Fig. 10: Critical gene extraction for cluster 0 and cluster 2 (clustering with 4 clusters) by ratio (up) and SVM linear coefficient (down). The genes within the green boxes appear in both files.



```
1 ENSG00000113140.9
2 ENSG00000108821.12
3 ENSG00000164692.16
4 ENSG00000168542.11
```

Fig. 11: Overlapping protein-encoding gene set for cluster 0 and cluster 1 in clustering with 4 clusters.

reoccurring cluster pair (Fig. 12 orange lines). In all of those pairs, one cluster always has higher-than-background asian and black ratio and lower-than-background white ratio, and the other one always has higher-than-background white ratio and lower-than-background asian and black ratio. The pathways might be able to explain the pairwise race ratio difference.

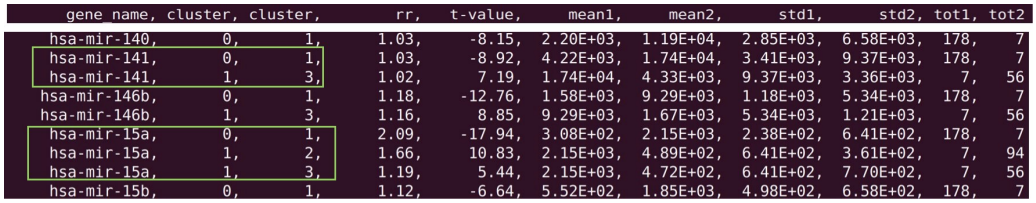
Also, for each clustering, we find that some genes are in the top gene set for several cluster pairs (Fig. 13).

num_cluster=2													
		total: 335											
		alive (298)	dead (37)		stage i (63)	ii (178)	iii (81)	iv (13)		asian (10)	black (50)	white (275)	
background	# samples	0.89	0.11		0.19	0.53	0.24	0.04		0.03	0.15	0.82	
cluster 0	97	0.85	0.15		0.18	0.48	0.28	0.06		0.03	0.13	0.84	
cluster 1	238	0.91	0.09		0.19	0.55	0.23	0.03		0.03	0.16	0.82	
num_cluster=3													
		total: 335											
		alive (298)	dead (37)		stage i (63)	ii (178)	iii (81)	iv (13)		asian (10)	black (50)	white (275)	
background	# samples	0.89	0.11		0.19	0.53	0.24	0.04		0.03	0.15	0.82	
cluster 0	54	0.83	0.17		0.17	0.48	0.28	0.07		0.02	0.15	0.83	
cluster 1	223	0.91	0.09	less dead	0.2	0.54	0.23	0.03	more i, ii, less iii, iv	0.04	0.15	0.81	more asian, less white
cluster 2	58	0.88	0.12	more dead	0.16	0.55	0.24	0.05	less i, more ii, more iv	0.02	0.14	0.84	less asian, less black, more white
num_cluster=4													
		total: 335											
		alive (298)	dead (37)		stage i (63)	ii (178)	iii (81)	iv (13)		asian (10)	black (50)	white (275)	
background	# samples	0.89	0.11		0.19	0.53	0.24	0.04		0.03	0.15	0.82	
cluster 0	178	0.92	0.08	less dead	0.18	0.57	0.22	0.02	less i, more ii, less iii, iv	0.03	0.18	0.79	more black, less white
cluster 1	7	0.71	0.29		0.29	0.29	0.14	0.29		0	0.29	0.71	
cluster 2	94	0.86	0.14		0.22	0.45	0.29	0.04		0.04	0.11	0.85	
cluster 3	56	0.88	0.12	more dead	0.14	0.57	0.23	0.05	less i, more ii, less iii, more iv	0.02	0.11	0.88	less asian, less black, more white
num_cluster=5													
		total: 335											
		alive (298)	dead (37)		stage i (63)	ii (178)	iii (81)	iv (13)		asian (10)	black (50)	white (275)	
background	# samples	0.89	0.11		0.19	0.53	0.24	0.04		0.03	0.15	0.82	
cluster 0	207	0.9	0.1	less dead	0.16	0.57	0.23	0.03	less i, more ii, less iii, iv	0.03	0.16	0.81	more black, less white
cluster 1	7	1	0	less dead	0	0.57	0.43	0	less i, more iii, less iv	0	0.14	0.86	less asian, black, more white
cluster 2	51	0.84	0.16		0.24	0.49	0.24	0.04		0.04	0.2	0.76	
cluster 3	6	0.83	0.17		0.33	0.67	0	0		0	0.17	0.83	
cluster 4	64	0.88	0.12		0.23	0.42	0.28	0.06		0.03	0.06	0.91	
num_cluster=6													
		total: 335											
		alive (298)	dead (37)		stage i (63)	ii (178)	iii (81)	iv (13)		asian (10)	black (50)	white (275)	
background	# samples	0.89	0.11		0.19	0.53	0.24	0.04		0.03	0.15	0.82	
cluster 0	7	0.85	0.14		0.29	0.71	0	0		0	0.14	0.86	
cluster 1	33	0.85	0.15		0.18	0.48	0.27	0.06		0.03	0.21	0.76	
cluster 2	162	0.9	0.1	less dead	0.17	0.55	0.26	0.02	less i, more ii, iii, less iv	0.04	0.17	0.79	more asian, black, less white
cluster 3	24	0.96	0.04		0.17	0.62	0.12	0.08		0	0.29	0.71	
cluster 4	45	0.91	0.09	less dead	0.16	0.62	0.18	0.04	less i, more ii, less iii	0	0.09	0.91	less asian, less black, more white
cluster 5	64	0.86	0.14		0.25	0.39	0.3	0.06		0.03	0.06	0.91	
num_cluster=7													
		total: 335											
		alive (298)	dead (37)		stage i (63)	ii (178)	iii (81)	iv (13)		asian (10)	black (50)	white (275)	
background	# samples	0.89	0.11		0.19	0.53	0.24	0.04		0.03	0.15	0.82	
cluster 0	24	0.96	0.04		0.17	0.62	0.12	0.08		0	0.29	0.71	
cluster 1	5	0.8	0.2		0.4	0.6	0	0		0	0.2	0.8	
cluster 2	68	0.87	0.13		0.24	0.41	0.29	0.06		0.04	0.06	0.9	
cluster 3	158	0.89	0.11		0.17	0.56	0.25	0.02	less i, more ii, iii, less iv	0.04	0.17	0.79	more asian, black, less white
cluster 4	34	0.85	0.15		0.32	0.47	0.26	0.06		0.03	0.21	0.76	
cluster 5	45	0.91	0.09	less dead	0.16	0.62	0.18	0.04	less i, more ii, less iii	0	0.09	0.91	less asian, black, more white
cluster 6	1	1	0		0	0	1	0		0	0	1	
num_cluster=8													
		total: 335											
		alive (298)	dead (37)		stage i (63)	ii (178)	iii (81)	iv (13)		asian (10)	black (50)	white (275)	
background	# samples	0.89	0.11		0.19	0.53	0.24	0.04		0.03	0.15	0.82	
cluster 0	5	0.8	0.2		0.4	0.6	0	0		0	0.2	0.8	
cluster 1	49	0.9	0.1		0.18	0.61	0.2	0		0.08	0.2	0.71	
cluster 2	1	1	0		0	0	1	0		0	0	1	
cluster 3	7	1	0	less dead	0	0.57	0.43	0	less i, more ii, iii, less iv	0	0.14	0.86	less asian, black, more white
cluster 4	136	0.92	0.08	less dead	0.18	0.55	0.24	0.03	less i, more ii, less iv	0.03	0.16	0.81	more black, less white
cluster 5	12	0.75	0.25		0.17	0.33	0.33	0.17		0	0.17	0.83	
cluster 6	63	0.87	0.13	more dead	0.24	0.4	0.3	0.06	more i, less ii, more iii, iv	0.03	0.06	0.9	less black, more white
cluster 7	62	0.85	0.15	more dead	0.16	0.6	0.19	0.05	less i, more ii, less iii, more iv	0	0.16	0.84	less asian, more black, more white

Fig. 12: Label distribution within each cluster for all clustering. Each black block is for a clustering with a certain number of clusters. Green lines show the background ratio, i.e., ratio of a label in the whole data set. Orange lines show the situation for the reoccurring cluster pair.

4 Conclusion

Machine learning models based on the microRNA and mRNA quantification data for predicting 5-year survival of a breast cancer subtype, tumor stage and race have poor performance. Doing PCA on the microRNA and mRNA quantification data cannot improve the performance. However, SVM with



gene_name,	cluster,	cluster,	rr,	t-value,	mean1,	mean2,	std1,	std2,	tot1,	tot2
hsa-mir-140,	0,	1,	1.03,	-8.15,	2.20E+03,	1.19E+04,	2.85E+03,	6.58E+03,	178,	7
hsa-mir-141,	0,	1,	1.03,	-8.92,	4.22E+03,	1.74E+04,	3.41E+03,	9.37E+03,	178,	7
hsa-mir-141,	1,	3,	1.02,	7.19,	1.74E+04,	4.33E+03,	9.37E+03,	3.36E+03,	7,	56
hsa-mir-146b,	0,	1,	1.18,	-12.76,	1.58E+03,	9.29E+03,	1.18E+03,	5.34E+03,	178,	7
hsa-mir-146b,	1,	3,	1.16,	8.85,	9.29E+03,	1.67E+03,	5.34E+03,	1.21E+03,	7,	56
hsa-mir-15a,	0,	1,	2.09,	-17.94,	3.08E+02,	2.15E+03,	2.38E+02,	6.41E+02,	178,	7
hsa-mir-15a,	1,	2,	1.66,	10.83,	2.15E+03,	4.89E+02,	6.41E+02,	3.61E+02,	7,	94
hsa-mir-15a,	1,	3,	1.19,	5.44,	2.15E+03,	4.72E+02,	6.41E+02,	7.70E+02,	7,	56
hsa-mir-15b,	0,	1,	1.12,	-6.64,	5.52E+02,	1.85E+03,	4.98E+02,	6.58E+02,	178,	7

Fig. 13: The genes in the green boxes are in the top gene set of several cluster pairs in clustering with 4 clusters.

linear kernel shows high accuracy for predicting the labels generated by k-means clustering (2-8 clusters). Using the top gene set suggested by the ratio (1) and SVM coefficient for each cluster pair as input, the Gene Ontology (GO) term enrichment technique suggests that the input gene set of one reoccurring cluster pair significantly takes part in several pathways. Those pathways might be some signatures which can explain the gene expression difference and the classification of the patients in the cluster pair.

In the future, more data can be used for clustering. In this project, a lot of samples are discarded in the process of generating the 5-year survival label. Also, the overlapping gene sets that are not shown to be significantly involved in pathways according to GO still can be looked at in more depth.

5 My contribution

The aim of my individual contribution was to preprocess data downloaded by and prepare a .csv file for using sklearn machine learning package, run machine learning to predict labels, apply PCA, run clustering, extract the genes that may account for the classification of the clusters (the files containing the genes for each cluster pair, a label distribution within each cluster and to interpret the results after using getting the Kpathways results.

How I worked: write python scripts to fulfill the aim above. A general description for each script can be found in the run.sh file in my zipped codes. I created all figures in this report.

References

[1] Genomic data commons data portal. [Online]. Available: https://portal.gdc.cancer.gov/