Computational Biology: Understanding Cancers

Student ID: 29845998

***Abstract*—In this report we are going to investigate 3 Kidney cancers along with normal samples from the TCGA TARGET GTEx cohort. The cancer types are Chromophobe, Clear Cell and Papillary. We investigate the structure of the data using dimensionality reduction and K-means clustering, followed by linear models to find similarities between the cancer types. We then use Random Forest Classifiers and Decision Trees to find the most important features which differentiate and predict these cancers.**

I. The Data

The data used in this report is gene expression RNAseq - RSEM expectedcount (DESeq2 standardized) from the TCGA TARGET GTEx cohort. This data set is combination of TCGA and GTEx data sets, with samples that have been reprocessed through the same data pipeline so they can be used together. The features represent Ensembl genes each with their own Ensembl ID, beginning with ENSG. The data set is huge and so to get the samples we want we used the visualisation tools available from the host website (xenabrowser.net) to get a list of samples names. Then, using the xenaPython package created by Xena we are able to extract only the samples we need (1013 out of the total 19039) along with their corresponding 60499 features. The data is then stored in arrays using the Numpy package and then put together into data frames using the Pandas package. The data consists of 129 normal samples, 66 Chromophobe, 530 Clear Cell and 288 Papillary. The data is imbalanced and so this is going to influence our results, especially for the Chromophobe samples. From quickly inspecting the data there a few columns which are all zeroes, these have been removed to help reduce the size of the data set (Now 1013 samples x 53798 features).

II. Visualization

To get some idea of how the data looks we ran some dimensionality reduction. To do this we used LDA to represent the data in 2D, allowing us to plot it. In Figure 1 we can see the results. As expected, the normal samples are separate from the cancer samples. However, it is interesting to note that the Chromophobe samples are very similar to Clear Cell (at least in this transformation). This could be because they appear similar in 2D but if we were to use a higher dimension representation and test the distance between the two clusters, we would find that they are not so similar. It could also be that these cancer types are structurally quite similar, this will become more clear once more investigation has been done. Due to the lack of Chromophobe data it might seem like it is similar to Clear Cell, and if we had more data it would 'bridge' the gap between Papillary and Clear Cell. To further investigate these results we ran K-means on this reduced data and find that a cluster of 3 is very clearly the best choice. This can intuitively be seen from Figure 1 but to support this we plot the elbow graph Figure 2.

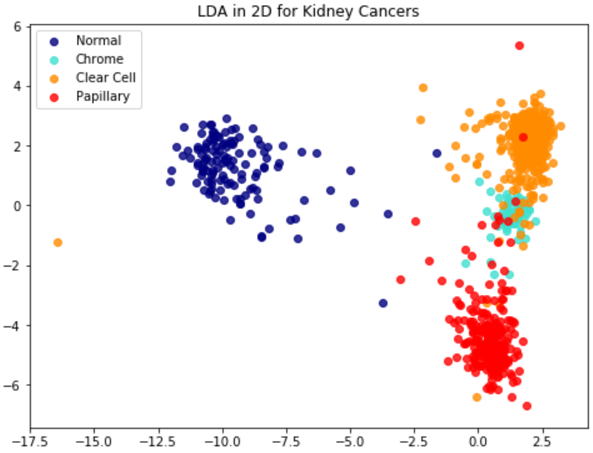

Fig. 1: Reduced data to 2D and plotted the four classes

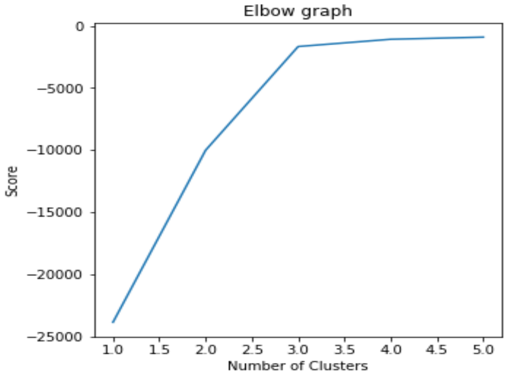

Fig. 2: Plotting the different scores for K-means on varying cluster size for reduced data

To investigate further we run K-means on the whole data to see if the 3 clusters is because of the choice of dimensionality reduction or if the data does indeed follow a 3 grouping structure. The elbow graph in Figure 3 suggests that a good choice of clusters is either 2 or 3 clusters, and possibly 4 (at a push). Although it is hard to tell. After choosing a cluster size of 4 we find that it groups them with an accuracy of 0.943, mainly mislabelling Chromophobe samples as different classes, not just Clear Cell.

III. Machine Learning - SGD

To investigate this similarity further we ran sklearns SGD classifier on a training set of the data (80% of each class, leaving the remaining for the test set). When running the

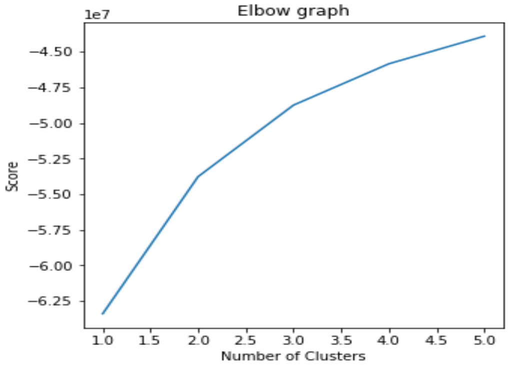

Fig. 3: Plotting the scores for varying cluster numbers for K-means on all the data

model on the test set we have an accuracy of 93.5%. To get a better idea of where this is incorrectly classifying we ran test data only comprised on one group to see how the model does. It manages 100% accuracy on Normal samples, 84.6% for Chromophobe, 90.6% for clear cell and 98% papillary. The lower accuracy on Chromophobe and Clear Cell does support the idea that they are quite similar and so it is hard to differentiate between them. Meaning that perhaps Figure 1 is a good representation of the data. However, it is important to note that as there is very little data for Chromophobe the model will find it difficult to predict it. The high accuracy for the classes (excluding Chromophobe) would suggest that the data is linearly separable, meaning that there should be features which will allow us to differentiate between the classes.

A. Machine Learning - Decision trees and Random Forest

To get down to the features that are causing these differences we shall use Sklearns DecisionTreeClassifier and RandomForestClassifier to analyse each cancer type separately and determine which features differentiate them from normal samples, then find which features are important for differentiating the cancers from each other.

After running a few preliminary tests with decision trees on all pairings, we find that cancers can be differentiated from normal samples with 100% accuracy for many features. This makes sense as our linear separator was able to distinguish normal samples with 100% accuracy. However, as there are likely to be many we ran a random forest with 200 decision trees using and looked at the most important features. These can be seen in Figure 4-6. Where the sum of all the feature importance's is 1. The plots show that out of $\approx 54,000$ features only about 40 are noticeably important. Examining these three forests further we can see the top 3 feautures for each group:

Group1

ENSG00000125967.16 is the gene for NECAB3.

ENSG00000223728.3 none found.

ENSG00000224417.2 is a lncRNA which is a subgroup of Long non-coding RNA.

Group2

ENSG00000139144.9 is gene PIK3C2G

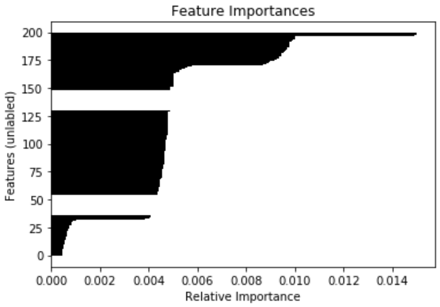

Fig. 4: Chromophobe

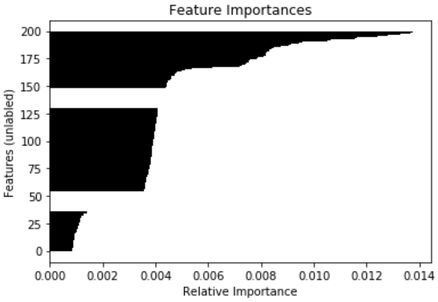

Fig. 5: Clear Cell

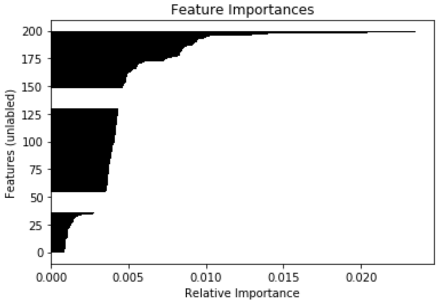

Fig. 6: Papillary

ENSG00000254789.1 is gene RP11-531H8.2 (Clone-based (Vega) gene) lincRNA 465bp

ENSG00000147257.13 is for glypican 3 GCP3 (LOOK INTO MORE)

Group3

ENSG00000281769.1 is a lincRNA which is a subgroup of Long non-coding RNA. This genomic region is 465bp.

ENSG00000249601.2 LINC01187 (HGNC Symbol) two transcripts of lincRNA

ENSG00000158955.10 is gene WNT9B

The top 40 features for each pairing are stored and then compared against each other to see if any of the same features are found. We would expect that Chromophobe and Clear to have some in common as they are similar but this is not the case. In fact they have none in common. This could just be down to the randomness of the trees but it is unlikely to not have any in common if this was due to randomness, especially as they are so similar in the cluster plot. We find that Chromophobe and Papillary have two of their top 40 in common. ENSG00000066230.10 - SLC9A3 (HGNC Symbol) gene and ENSG00000167580.7 - AQP2 gene.

Clear cell and Papillary have two. ENSG00000128045.6 is gene for RASL11B and ENSG00000254789.1 has been seen above.

Next we ran a decision tree on all of the data to see where it splits the data. Unfortunately the tree is too big to visualise but we can visualise the top few and analyze the most important features, i.e the nodes with the highest entropy in Figure 7.

At the top of the tree we have ENSG00000129521.13 splitting values for $\leq$ 13.185, with a gini of 0.625. This is able to almost entirely isolate Clear Cell samples. This gene codes for protein EGLN3 (HGNC Symbol). The follow branches form false are very low entropy as they are just fine tuning but do manage to maintain the same amount of samples for Clear Cell (488).

Following on from the top layer down the True branch we have ENSG00000004939.13 $\leq$ 9.898 with a gini of 0.622. We are able to completely factor out all the normal samples but this also takes most of the Chromophobe data with it. If we follow the true branch we are left with only a small amount of Clear cell and most Papillary samples. This gene codes for protein SLC4A1.

Following the false branch we have a node for ENSG00000225329.3 $\leq$ 7.397 with a gini of 0.525. is able to isolate all the remaining normal samples if false. If true it leaves most of the Chromophobe samples with very few clear and papillary. If you follow this branch down it eventually isolates the Chromophobe samples.

The other features do not have produce much more entropy and are more fine tuning. Interestingly, there are nodes which have most of the classes isolated with a few samples missing. This further suggests that it is possible to isolate the classes based on a few features despite them being overall quite similar. This similarity comes from the fact that the clusters at the beginning of the analysis suggest there are 3 clusters, not 4.

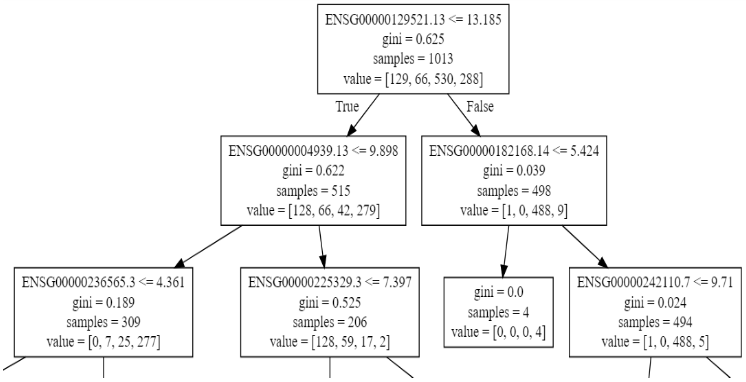

Fig. 7: Top 3 layers of the decision tree showing some of the features with highest purity

IV. Conclusions

Interestingly, our results from K-means and linear model suggest that Chromophobe and Clear Cell are very similar, but in our decision tree on all of the data we found that Chromophobe and the normal samples are often grouped together by ENSG00000129521.13 and ENSG00000004939.13. Unfortunately these genes did not result in any particularly interesting results and so there was not much to discuss. However, due to the lack of Chromophobe samples, it is hard to draw good conclusions.

V. Future Work

As decision trees and random forest have random elements, a more robust approach would be to run multiple forests and inspect the most consistent features. Also, there are more data sets which we could use to investigate and see if we obtain similar results. Originally we had planned to also compare normal and cancerous samples of testis and ovaries, as they share a common origin and this might mean they have similar important features despite being different organs. However, the time it took to gather the data meant that we had to abandon this. Given more time this would definitely be of interest.