Cancer Dataset Analysis

Sanjeevan Sritharan, Santiago Duque, Suyash Dubey

June 2020

1 Introduction

In this project, we are exploring causes and effects of different types of cancers. We decided to use the datasets from The Cancer Genome Atlas (TCGA). These datasets describe tumour tissues and matched normal tissues, and are widely used for research purposes. From the available cancer cohorts, we selected 5 different types of cancers to analyse: bladder, colon, rectal, prostate and testicular cancer.
These cancers were chosen since they had the same datasets, which is helpful for creating the decision trees and the linear SVMs. The datasets used for creating the decision trees and SVMs are the following:

- From the copy number (gene-level), the gistic2 dataset.
- From DNA methylation, the Methylation450k dataset.
- From phenotype, the Curated survival data dataset.
- From gene / exon expression RNAseq, the IlluminaHiSeq dataset.
- Additionally, for task 3 (finding common proteins in different types of cancer), we used the RPPA protein concentration dataset.

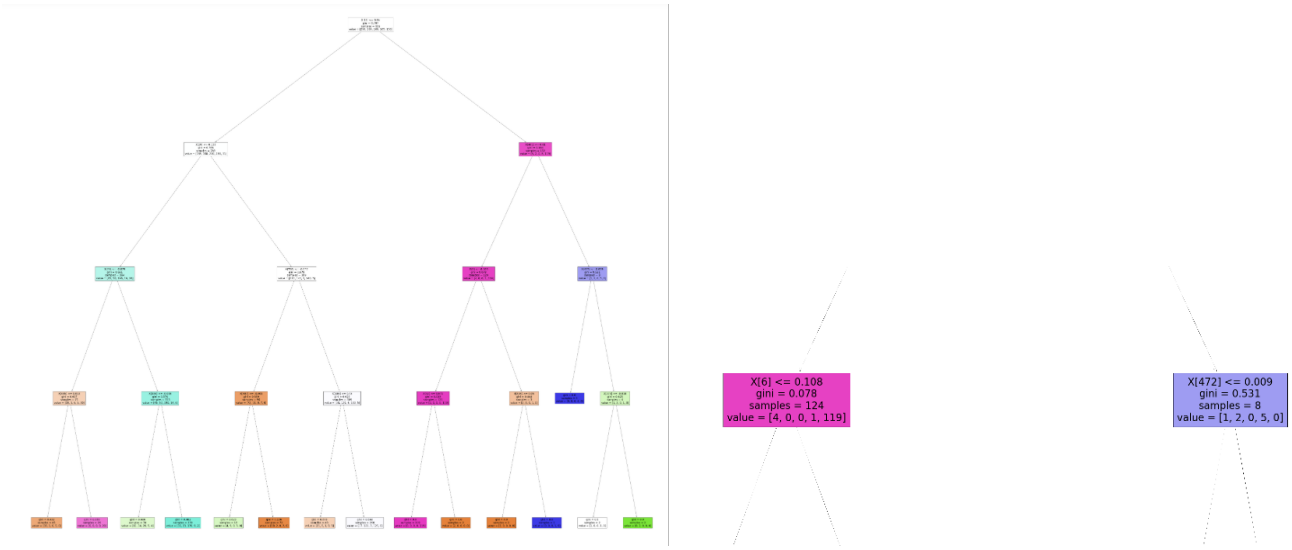
We performed these tasks using Python, and the xenaPython API to handle the different cohorts and datasets.
To create the decision tree to predict cancer type, we used the DecisionTreeClassifier available in the sklearn.tree machine learning library. A code snippet showing the creation of the decision tree for the gistic2 dataset is given below,

For the linear SVM, we used the SVC available in the sklearn.svm machine learning library. A code snippet for the SVM created for the IlluminaHiSeq dataset from gene expression RNAseq is given below,

```
In [9]: # classify on gistic2 data
        classifier = tree.DecisionTreeClassifier(max_depth=4)
        classifier.fit(X,Y)

Out[9]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=4,
                               max_features=None, max_leaf_nodes=None,
                               min_impurity_decrease=0.0, min_impurity_split=None,
                               min_samples_leaf=1, min_samples_split=2,
                               min_weight_fraction_leaf=0.0, presort=False,
                               random_state=None, splitter='best')
```

Creation of the decision tree



A whole decision tree and a detail for two of its nodes, respectively.

2 Findings

Here we will detail out most relevant findings, that is, the conclusion of our research for each of the variables found at the top of the decision trees and SVM classifiers for each dataset.

2.1 Gistic2

- A2MP1 gene (alpha-2-macroglobulin pseudogene 1)
  - This is the most defining variable, which sets the testis cancer apart from the rest.The alpha-2-Macroglobulin is a plasma protein found in blood. This protein is found in high levels in cases of nephrotic syndrome (a disease affecting the kidneys), as well as in female Stress Urinary Incontinence (Wen et al. 2007) . The difference in the concentration amount of alpha-2 M might indicate an imbalance in the excretory systems surrounding the kidneys and bladder. Other than that, we found no evidence tying this gene to testicular cancer.
- AAR2, a protein-coding splicing gene.
  - In our model, it selects for prostate cancer. According to *www.proteinatlas.org/ENSG00000131043-AAR2/pathology*, it has a low-specificity in cancer, and as such isn't a great predictor for cancer types. It is indeed found in large concentrations in prostate cancer, but also in other unrelated variants like skin cancer. We can therefore conclude that it can somewhat predict for prostate cancer, but not definitively.
- AHCY is an enzyme that converts S-adenosylhomocysteine to homocysteine and adenosine.
  - We examined the research found in *www.proteinatlas.org/ENSG00000101444-AHCY/pathology/testis+cancer*, and according to it, AHCY is not prognostic in testis cancer. However, we do see quite a large concentration of it in this particular type. (Belužić et al. 2018) suggests screening for AHCY as a potential preventive measure of cancer, so it could be possible that it is a good predictor for testicular cancer.
- ACSS2 (found with SVM)

- Our model finds the largest separation in this variable between colon and rectal cancer. Indeed, research (*www.proteinatlas.org/ENSG00000131069-ACSS2/pathology*) suggests that we can find this enzyme in large concentrations in colorectal cancer, and in very small concentrations in prostate cancer.

2.2 Illumina:

- ACPP is the prostatic acid phosphatase, and it is the most defining variable in this dataset.
  - As the name suggests, it selects for prostate cancer. It is synthesized under androgen regulation and is secreted by the epithelial cells of the prostate gland. Indeed, all research suggests that it is a very good predictor for this particular type of cancer, and it is used as a prognostic for it.
- APB1 gene (AP complex subunit beta)
  - The corresponding protein catalyzes the degradation of compounds such as putrescine, histamine, spermine, and spermidine, substances involved in allergic and immune responses, cell proliferation, tissue differentiation, tumour formation, and possibly apoptosis. In our model, it selects for both colon and rectal cancer. It seems to not be a strong factor for the prognosis of either of those types of cancer, or for any cancer in general.
    (*www.phosphosite.org/proteinAction?id=2350394showAllSites=true*)
- ADORA1 is the Adenosine A1 receptor.
  - This gene encodes a protein that is an adenosine receptor that belongs to the G-protein coupled receptor family. Inhibition of ADORA1 promotes tumour immune evasion, failure of the immune system to detect the tumour. The deletion of ADORA1 suppresses cell growth in human melanoma cell lines and tumour development, but it also compromises anti-tumour immunity and reduces anti-tumour efficiency (Liu et al. 2020). In our model, it predicts prostate cancer.
- ABCB11 (found through SVM)
  - Our model finds the largest separation in this variable between colon and rectal cancer. The ABCB11 gene is the ATP binding part cassette, subfamily B member 11. It provides instructions for synthesising a protein called the bile export pump (BSEP), which is found in the liver. Mutations in the ABCB11 gene can cause intrahepatic cholestasis. These mutations in the ABCB11 gene also increase the risk of hepatocellular carcinoma, which is the most common type of liver cancer in adults. However, it is also a minor indicator of colon and rectal cancer (*www.proteinatlas.org/ENSG00000073734-ABCB11/pathology*). In our model, it indeed predicts colorectal cancer.

2.3 Methylation450k:

- cg00000714
  - Relevant in studies regarding the prevalence of Lupus across different human populations (Teruel & Sawalha 2017). In our dataset, it predicts for testis cancer. Indeed, there have been some findings regarding the relation between some autoimmune diseases and testicular cancer (Mandel-Brehm et al. 2019), as well as between Lupus and testicular manifestations (testicular tumour) (Kuehn et al. 1989)
- cg00001510
  - No valuable data was available for this particular gene. For the dataset, the decision tree predicts either colon cancer, prostate cancer and rectal cancer.

- cg00009196
  - No valuable data was available for this particular gene. However, from the decision tree, the prediction is for testicular cancer.

2.4 Survival data:

- OS (Overall Survival) time
  - This is the time from the diagnosis until death. In our model, it selects for prostate and testis cancer. This fits well with the known available data, since prostate and testicular cancer have a 5-year survival rate of 100 and 95%, respectively (that is, the number of patients that will live at least 5 years after the diagnosis). In contrast, the 5-year survival rate for bladder, rectal and colon cancer is 77, 67 and 63% - considerably lower than that of testicular and bladder cancer (*https://www.cancer.net/*).
- DSS (Disease-Specific Survival) time.
  - DSS differs from the Overall Survival in that the latter takes into account death due to any cause, while the former only counts if the cause of death is the particular disease we are studying..
- DFI (Disease-Free Interval) time.
  - This is defined as the interval from the completion of chemotherapy to the diagnosis of recurrence. That is, the time it takes for the disease to "come back". In our model, it selects for Bladder cancer. Our research suggests that bladder cancer patients do indeed have a longer DFI, or recurrence-free interval, than other similar times of cancer (Akagashi et al. 2006). The results are not conclusive, however.

The most useful datasets were for the SVM and the decision trees were the gistic2 and IlluminaHiSeq. The Methylation450k was largely inconclusive, since it is difficult to find information about specific genes and encoded chromosome strands.

2.5 Common proteins

We also compared the concentration of different proteins in some types of cancers among those we studied. Unfortunately, only three of them had available datasets regarding the protein expression - bladder, colon and rectal. The protein expression RPPA datasets for testicular cancer and prostate cancer were empty so these cancers were not taken into account during the classification. Some of the common proteins found in large concentrations in both cancers:

- CLAUDIN7
  - This gene encodes a member of the claudin family. Claudins are a family of membrane proteins which are the most important components of the tight junction strands (*https://www.ncbi.nlm.nih.gov/gene/1366*). This protein is very significant in colorectal cancer, according to research (Wang et al. 2018), so it is logical that we would find in in large concentrations in both types.
- Hsp70
  - Hsp70 proteins are central components of the cellular network of molecular chaperones and folding catalysts. They assist a large variety of protein folding processes and help the cells to cope under stressful conditions (Mayer & Bukau 2005). This protein is found to correlate with poor prognosis in bladder cancers (Ciocca & Calderwood 2005).

- MAPKpT202Y204
  - This protein is relevant in squamous cell carcinomas (Joshua D. Campbell et al. 2018), a type of skin cancer, among which bladder cancer is found.
- IGFBP-2
  - The Insulin-like Growth Factor Binding Proteins (IGFBPs) regulate the activity of the Insulin-like Growth Factors (IGFs) ligands. The IGFBP-2 protein is involved in various functions, from embryonic growth to cell differentiation to homeostasis(Khan 2019). This protein is generally thought to be oncogenic, and is found in large amounts in some types of cancer, among which is colorectal cancer (Pickard1 & McCance 2016).

3 Conclusion

Through Machine Learning methods such as Decision Trees and SVMs, many insightful things can be discovered about different types of cancer, their common traits and their differences. Some particular information was difficult or impossible to find, such as certain genes or chromosome parts, but other was widely available, with extensive resources, mostly the proteins.

Indeed, we obtained several interesting common factors in the cancers we studied, as well as some unexpected results.

References

Akagashi, K., Tanda, H., Kato, S., Ohnishi, S., Nakajima, H., Nanbu, A., Nitta, T., Koroku, M., Sato, Y. & Hanzawa, T. (2006), 'Recurrence pattern for superficial bladder cancer'.
URL: *https://pubmed.ncbi.nlm.nih.gov/16834643/*

Belužić, L., Grbeša, I., Belužić, R., Park, J. H., Kong, H. K., Kopjar, N., Espadas, G., Sabidó, E., Lepur, A., Rokić, F., Jerić, I., Brkljačić, L. & Vugrek, O. (2018), 'Knock-down of ahcy and depletion of adenosine induces dna damage and cell cycle arrest'.
URL: *https://www.nature.com/articles/s41598-018-32356-8*

Ciocca, D. R. & Calderwood, S. K. (2005), 'Heat shock proteins in cancer: diagnostic, prognostic, predictive, and treatment implications'.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1176476/*

Joshua D. Campbell, C. Y., Bowlby, R., Pickering, C. R., Chen, Z. & Waes, C. V. (2018), 'Genomic, pathway network, and immunologic features distinguishing squamous carcinomas'.
URL: *https://www.cell.com/cell-reports/pdf/S2211-1247(18)30424-8.pdf*

Khan, S. (2019), 'Igfbp-2 signaling in the brain: From brain development to higher order brain functions'.
URL: *https://www.frontiersin.org/articles/10.3389/fendo.2019.00822/full*

Kuehn, M. W., Oellinger, R., Kustin, G. & Merkel, K. H. (1989), 'Primary testicular manifestation of systemic lupus erythematosus'.
URL: *https://pubmed.ncbi.nlm.nih.gov/2714324/*

Liu, H., Kuang, X., Zhang, Y., Xu, X., Hung, M.-C. & Chen, X. (2020), 'Adora1 inhibition promotes tumor immune evasion by regulating the atf3-pd-l1 axis'.
URL: *https://www.cell.com/cancer-cell/pdfExtended/S1535-6108(20)30095-7*

Mandel-Brehm, C., Dubey, D., Kryzer, T. J. & O'Donovan, B. D. (2019), 'Kelch-like protein 11 antibodies in seminoma-associated paraneoplastic encephalitis'.
URL: *https://www.nejm.org/doi/full/10.1056/NEJMoa1816721*

Mayer, M. P. & Bukau, B. (2005), 'Hsp70 chaperones: Cellular functions and molecular mechanism'.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2773841/*

Pickard1, A. & McCance, D. J. (2016), 'Igf-binding protein 2 – oncogene or tumor suppressor?'.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4343188/*

Teruel, M. & Sawalha, A. (2017), 'Aepigenetic variability in systemic lupus erythematosus: What we learned from genome-wide dna methylation studies'.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5819620/*

Wang, K., Xu, C., Li, W. & Ding, L. (2018), 'Emerging clinical significance of claudin-7 in colorectal cancer: a review'.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6159786/*

Wen, Y., Man, W. C., Sokol, E. R., Polan, M. L. & Chen, B. H. (2007), 'Is alpha2-macroglobulin important in female stress urinary incontinence?', *PubMed* .
URL: *https://pubmed.ncbi.nlm.nih.gov/18077315/*