Classification and Interactive visualization of tumor types in Neurofibromatosis based on RNA-seq and Drug Screening data

Introduction to the Motivating Problem

Neurofibromatosis (NF) is a rare genetic disorder of the nervous system affecting the development of nerve cell tissues. The three entities of neurofibromatosis are Types 1 (NF1), Types 2 (NF2), and schwannomatosis (SWN), with NF1 being the most common and SWN the rarest [1]. NF usually causes benign tumors with a nearly 10% chance of malignancy. The condition does not currently have a cure; thus, the most important elements of management are early diagnosis and treatment of the effects of the disease [1]. The objective of this proposal is to identify molecules associated with different tumor types of NF1 and NF2 and possible therapeutic targets using machine learning algorithms to classify the RNA sequence data and drug screening data.

Identifying transcriptome (RNA) sequence signatures that are unique to specific tumor types has been an area of interest and applied to other types of cancers. These analyzed signatures help identify expression profiles that help in disease prognosis and treatment [2]. Previous genomic profiling studies for NF demonstrated that NF tumor types have limited genomic variants with great phenotypic heterogeneity. Few studies published previously characterized the tumors associated with NF1 type using Latent Variable analysis and supervised machine learning [3]. Related RNA seq work has been done to leverage deep neural networks for other types of cancers [4]. A study found the predictive performance of two feed-forward neural networks and a linear model to be comparable [5].

Previous work has also been done on drug-screening data, specifically the relationship of drugs with their target proteins. Ma et al. analyzed large topological modules extracted from networks like protein-protein interaction (PPI) using a couple of different cluster detection algorithms [8]. Evaluating the performance of these algorithms using metrics like shortest-path and diffusion correlation, the results of this study indicated that the Walktrap clustering algorithm achieved the best performance overall.

By utilizing the proposed workflow below, we will attempt to successfully analyze the data and clearly visualize the results. The payoff of these efforts will help identify important gene targets for the development of a treatment or therapy as well as for the classification of NF types. This project will also benefit NF2 patients as there is no previous classification done for the tumors normally associated with this type of NF. The drug study will help identify potential drug targets based on RNAseq results. It will be highly useful for researchers when trying to develop clinical trials, and consequently, it will be beneficial for the patients who would take therapeutics in the future if a drug is developed. Further studies can be done on larger datasets and untested patients as well, to prove the validity of the project.

Since we could not find any previous studies on the classification of NF2 tumors based on RNA seq data, in this project, we plan to identify the transcriptomic signatures of tumors associated with two types of NF - NF1 and NF2 - as well as possible drug targets. The traditional approach to drug therapy development has been to test new or repurposed drugs in various model systems based on hypotheses of tumor pathogenesis [9]. Several large scale projects characterizing drug sensitivities have been done, but many ignored drugs commonly used to treat tumors common to NF1 and NF2 [10]. For our approach, we will be examining drug screening data and looking specifically at those thought to be effective against NF tumors.

Proposed Methods

The first step of our project was to clean and normalize the datasets [14]. Later, dimensionality reduction would be done using PCA for downstream classification [11]. In case PCA did not produce desirable results, we proposed an alternative dimensionality reduction technique. Our backup plan was to exploit the autoencoder framework by Geddes [7] and experiment with methods in the random projection step, such as spectral projections. Using these reduced dimensions, supervised algorithms like random forest were used for training and identifying molecular signatures that differ in each of the tumor types. Random forest classifier was considered because it is commonly used for the classification of cancer types and the study on NF1 classification also included this classifier with Latent variables [3][13]. Accuracy, Precision, Recall, F1-score, and ROC curve were the proposed metrics to be used to validate performance. The most highly expressed genes for each tumor type would be used to find drug associations. For integrative visualization, we planned to use D3 to generate a heatmap to visualize the distribution between the tumor types and their genes. We also planned to generate a second heatmap to display the correlation between PCA latent variables and tumor types, and explore bioinformatics tools like iDEP to generate volcano and scatter plots.

Performed Methods and Experiments

Datasets

The datasets used were obtained from NF researchers who granted access via the NF Hackathon hosted by the Children's Tumor Foundation. Since the data is highly sensitive due to patient privacy concerns, access to the datasets had to be requested through synapse.org.

The RNAseq meta dataset includes a total of 6,130,214 records which include differential expression data for tumor samples from 255 patients. These samples are primarily associated with two different diagnoses, Neurofibromatosis 1 and 2, but also had non-NF tumors for control purposes. The tumor types included in the datasets are Cutaneous Neurofibroma (cNF), Plexiform Neurofibroma (pNF), Low-grade Glioma, High-grade Glioma, Meningioma, and Schwannoma, with the latter two being tumor types of NF2. While it may seem like 255 is a small sample size to work on, it should be noted that NF is a rare disease so large sample sizes are uncommon, and several studies [3] have used similar sized datasets (or even smaller) for their studies. Furthermore, each data point in this dataset has about 50k features corresponding to the differential expression of each gene, which further drives the importance of dimensionality reduction, but also justifies the dataset size since studies have found that for random forests, purity or near purity tends to be more effective when the feature space is large and the sample size is small [20].

The genomic datasets included information about tumors associated with patients, the tumor type, diagnosis, and the RNA seq total counts obtained for each gene in the tumor sample. The drug dataset included genes and information on their drug association, i.e. possible targets and dose responses.

Filtering/Cleaning

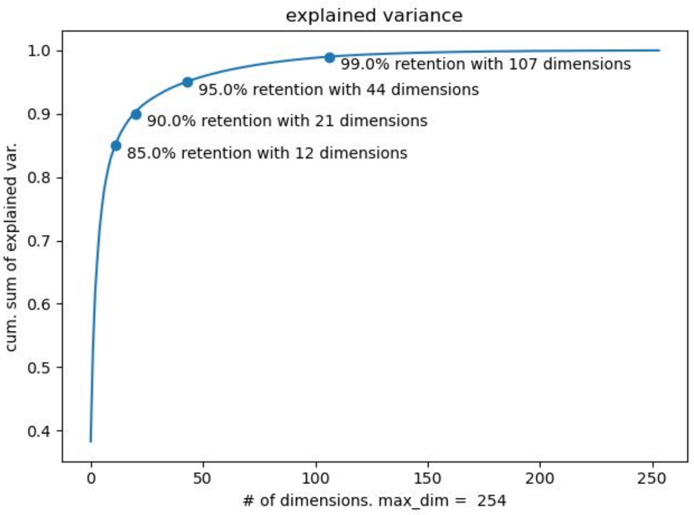
RNA seq has large dimensions as there are a large number of genes in the count data. Some genes in the dataset were very sparse across all samples, showing little evidence of differential expression [14]. They may also add a burden when estimating false discovery rates and detecting differentially expressed genes. Some of the tumor types in the given datasets are marked as NaN, so we tried to obtain the missing information from the modelOf column and dropped the rows for which we couldn't find a value.

Normalization

An important consideration for differential expression tests is variance. We shrink our estimates to have statistical meaning by normalizing the data. We used Variance stabilizing transformation (vst) of Deseq2 package of R to use for dimensionality reduction. We tried the rlog() method and it took several hours to run, so we chose not to use it. Minmaxscaler and log transformation were also tested. We used vst in the final analysis as it takes into account the size factor (depth of normalization) and the mean dispersion across the samples and is often preferred before machine learning analysis. For the drug analysis, the RNAseq data was normalized by subtracting the mean z-score of the normal tumors from the other tumor types.

Dimensionality Reduction

Since each data point contains data for about 50k features, there is a need for dimensionality reduction in order to avoid the curse of dimensionality. In our case, feature selection is an important step to optimize the dimensions which would reduce computational cost as well as improve accuracy for classification of gene expression data (phenotypes). The extracted number of features influences the classification results by over/under-fitting the data. PCA emphasizes variation and brings out strong patterns in the dataset. We apply PCA to our dataset via singular value decomposition. To determine the number of PCA components, we make the plot on the left and choose the number that retains 99% of the data variance. Currently, the latent space generated from PCA produces results that are good enough to not warrant experimentation with the alternative autoencoder technique.



Classification

Using these reduced dimensions, the Random forest algorithm is used for training and identifying molecular signatures that differ in each of the tumor types [3][13]. The crucial steps of classification can be considered as model fitting and validation. Machine learning on gene expression data could be a valuable new tool to understand differences between and within entities. For this project, we used a supervised learning - Random forest algorithm - as most papers for cancer tumor sub-classification used this method. As previously mentioned, data goes through dimension reduction before it is used for model fitting. Various hyper parameter tests are performed after splitting 107 dimensions obtained by PCA into 20 percent test and 80 percent training sets. Grid_Search with a 5-fold cross validation is used to test different metrics, like the number of trees in the forest ‘n_estimators': [4,6,16,100], maximum depth of the tree 'max_depth': [2,6,8,16,20] and random states 614 and 12345. Some of the parameters tested for training and testing accuracy are mentioned in the table below.

For the final classification we used a random state of 614, n_estimator =16 , max_depth =100. The ROC curve is generated by generating the y_score using .predict_proba. However, since the data at hand consists of multiclasses, the roc_curve is generated for each class. Label_binarize is used to generate y_scores for the y_test values. We observed false positive rates for the ROC curve, and this could be due to very few samples for each subtype of NF due to the rarity of the disease. Moreover, the distinction of features between the subtypes of some diseases is difficult as gene counts vary for each person, depending on various factors. Obtaining more samples for each subtype will improve the prediction of the classification method.

Table 1: Random Forest Classification - Different Hyper parameters tested for 107 PCA dimensions

Hyperparameter max_depth	Hyperparameter n_estimators	Random_State	Train Accuracy	Test Accuracy
5	80	12345	0.99	0.74
8	100	12345	1.0	0.76
5	100	12345	0.97	0.78
16	**100**	**614**	**1.0**	**0.86**
6	100	614	0.99	0.80
8	100	12345	1.0	0.76

Evaluation

The classification performance is measured using the metrics module. Different metrics are generated. From one perspective, these metrics are functions of four classification scenarios, true/false positive (TP/FP) and true/false negative (TN/FN). Accuracy incorporates all four scenarios, whereas precision ignores FN and recall ignores FP. The F1-score is the harmonic mean of precision and recall. Finally, the receiver operating characteristic curve (ROC curve) only considers TP and FP.

In our dataset, a TN occurs when the random forest classifier labels a healthy patient as “normal”. Conversely, a FN is when a sick patient is classified as healthy. TP and FP are defined similarly. A TP classification is when the classifier labels a sick patient with the correct disease and vice versa for FP. In general, a good classifier will have a large number of TP and TN and a small number of FP and FN. However, for medical diagnoses, avoiding FN is more important than reducing FP. The metrics that weigh FN more heavily should be considered first. Thus, metrics in order of importance are: recall, accuracy, f1-score, precision, and ROC curve.

Table 2: Evaluation of Random Forest Classification metrics and Comparison with the Study for NF1 subclassification

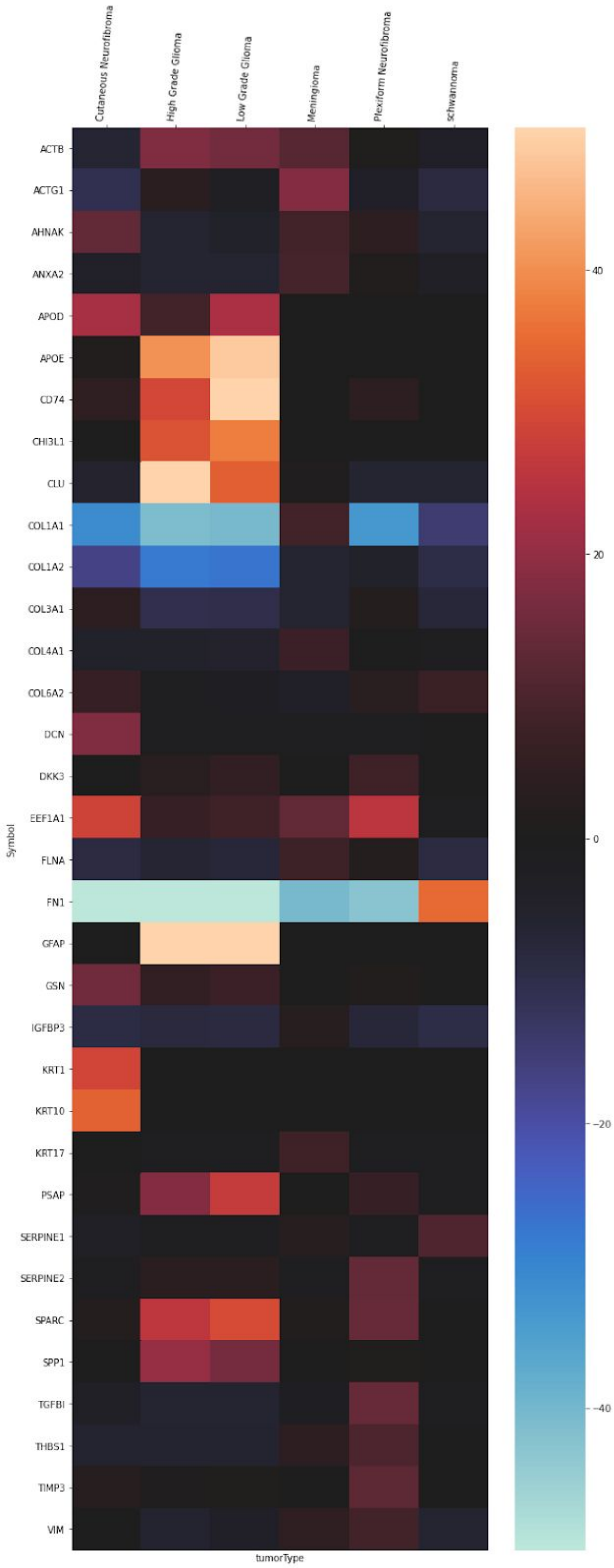
Testing	Metrics - Accuracy	Metrics - Precision	Metrics - Recall	Metrics - F1-Score
Experiment 1*	0.883116883116883	1.0	0.883116883116883	0.934376229830775
Team20	0.862745098039215	0.846855622590916	0.862745098039215	0.846984551396315
Experiment 2**	0.753246753246753	1.0	0.753246753246753	0.845461271077034

The Roc_Auc score is 0.9505794200797201.

Conclusion / Discussion

For the RNA seq classification analysis, the NF1 subclassification study [3] focuses on 3 tumor types of NF1 and one undefined type and considers two sets of latent variables, Experiment 2** included all latent variables and Set 1 included top 40 latent variables. The metrics show that the accuracy of Experiment 1* is better. The current project focuses on 8 types belonging to both NF1 and NF2 and a non tumor type. Our problem is twice as hard, given that we have twice as many prediction labels. Even with a harder problem, we were able to achieve an accuracy of 86 percent for classifying the tumor types. The current dataset size is smaller for tumor classification usually performed, but the task at hand is to sub classify the tumors which is an advancing field now. Considering the smaller size of the datasets due to the rarity of the disease, we tested the classification method using different PCA dimensions and test scores and tuned hyperparameters to design the classification to produce ideal results. Testing this method on more sample sizes for each subtype will also enable us to improve on the accuracy.

For the drug association analysis, we were able to obtain the top 10 highly expressed genes for each tumor type. Since there was overlap in the most disturbed genes amongst the tumor types, we ended up with 34 total genes (see figure below). Among these, FN1 and COL1A1 were both highly expressed for all 6 of the tumor types. COL1A2 was also highly expressed for most tumor types, but was not in the top 10 of Plexiform Neurofibroma and Meningioma tumor types. The gene FN1, or fibronectin 1, encodes fibronectin which is involved in cell adhesion and migration processes such as metastasis [15]. COL1A1 and COL1A2 (collagen type I alpha 1 chain and alpha 2 chain, respectively) encode the alpha chains of fibril-forming collagen found in most connective tissue. All three of these genes have been found to be candidate prognostic factors for different types of cancers [16, 17].



Following the finding of the most highly expressed genes, we found drug targets associated with these genes. In particular, zinc and copper seemed to be a common association for these genes according to the drug target explorer data. Zinc is vital in host defenses against the initiation and promotion of malignancies, and decreased levels have been seen in cancerous cells [18]. Copper, an essential element for biological processes, has been found to have elevated levels in tumor tissue and seems to play an important role in inflammation and tumor growth [19]. These two elements are the most common among the many potential drug targets we found in this study. With knowledge of a person's differentially expressed genes and the tumor type, it might help researchers to know associations they can specifically target in order to work on potential therapeutics. This, in turn, helps pave the way for personalized medicine.

In addition to this static visualization, we have generated interactive versions to display the specific z-score standardized values and target drugs of the top 34 gene symbols (heatmap.html). Furthermore, we have also visualized the workings of PCA in terms of a standardized correlation heatmap between the PCA components and tumor types. In that heatmap, it can be observed that as the number of dimensions increase, the correlation values decrease, signifying less important dimensions.

Distribution of Work

All team members contributed a similar amount of effort.

References:

1. Farschtschi, S., Mautner, V. F., McLean, A., Schulz, A., Friedrich, R. E., & Rosahl, S. K. (2020). The Neurofibromatoses. *Deutsches Arzteblatt international*, *117*(20), 354–360. https://doi.org/10.3238/arztebl.2020.0354
2. Zishuang Z., Zhi-Ping L. (2019). Identifying Cancer Biomarkers from High-Throughput RNA Sequencing Data by Machine Learning. Intelligent Computing Theories and Application, th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part II
3. Jineta B., Robert J. A., Jaclyn N.T., Aaron B., Xiaochun Z.,Chang I. M.,Christine A. P.,Jaishri O. B.,Justin G.,Angela H.,Casey S. G., & Sara JC.G.(2020). Integrative Analysis Identifies Candidate Tumor Microenvironment and Intracellular Signaling Pathways that Define Tumor Heterogeneity in NF1. *Genes 2020, 11(2), 226.* https://doi.org/10.3390/genes11020226
4. Way, G. P., Sanchez-Vega, F., La, K., Armenia, J., Chatila, W. K., Luna, A., Sander, C., Cherniack, A. D., Mina, M., Ciriello, G., Schultz, N., Cancer Genome Atlas Research Network, Sanchez, Y., & Greene, C. S. (2018). Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell reports, 23(1), 172–180.e3. https://doi.org/10.1016/j.celrep.2018.03.046
5. Urda D., Montes-Torres J., Moreno F., Franco L., Jerez J.M. (2017) Deep Learning to Analyze RNA-Seq Gene Expression Data. In: Rojas I., Joya G., Catala A. (eds) Advances in Computational Intelligence. IWANN 2017. Lecture Notes in Computer Science, vol 10306. Springer, Cham. https://doi.org/10.1007/978-3-319-59147-6_5
6. Vidman, L., Källberg, D., & Rydén, P. (2019). Cluster analysis on high dimensional RNA-seq data with applications to cancer research - An evaluation study. *PloS one*, *14*(12), e0219102. https://doi.org/10.1371/journal.pone.0219102
7. Geddes, T. A., Kim, T., Nan, L., Burchfield, J. G., Yang, J., Tao, D., & Yang, P. (2019). Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis. *BMC bioinformatics*, *20*(Suppl 19), 660. https://doi.org/10.1186/s12859-019-3179-5
8. Ma, J., Wang, J., Ghoraie, L. S., Men, X., Haibe-Kains, B., & Dai, P. (2019). A Comparative Study of Cluster Detection Algorithms in Protein-Protein Interaction for Drug Target Discovery and Drug Repurposing. Frontiers in pharmacology, 10, 109. https://doi.org/10.3389/fphar.2019.00109
9. Synodos for NF2 Consortium, Allaway, R., Angus, S. P., Beauchamp, R. L., Blakeley, J. O., Bott, M., ... & Clapp, D. W. (2018). Traditional and systems biology based drug discovery for the rare tumor syndrome neurofibromatosis type 2. *PloS one*, *13*(6), e0197350.
10. Guo, J., Grovola, M. R., Xie, H., Coggins, G. E., Duggan, P., Hasan, R., Huang, J., Lin, D. W., Song, C., Witek, G. M., Berritt, S., Schultz, D. C., & Field, J. (2017). Comprehensive pharmacological profiling of neurofibromatosis cell lines. *American journal of cancer research*, *7*(4), 923–934.
11. Gregory P. W., Robert J. A., Stephanie J. B., Camilo E. F., Yolanda S. & Casey S. G.(2017).A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma. *BMC Genomics, 18, 127.* https://doi.org/10.1186/s12864-017-3519-7
12. Kiselev, V.Y., Andrews, T.S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 20, 273–282 (2019). https://doi.org/10.1038/s41576-018-0088-9

13. David, G. P. I., Karoly S., Inge H. B., Marie k., Marieke L. K., Judith V. M. G. B.(2019). Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. Plos Computational biology. https://doi.org/10.1371/journal.pcbi.1006826
14. Gokmen Z., Dincer G., Selcuk K., Vahap E., Gozde E. Z., Izzet P.D., Ahmet O.(2017). A comprehensive simulation study on classification of RNA-Seq data. Plos one. 10.1371/journal.pone.0182507
15. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-45 PubMed
16. Li B, Shen W, Peng H, et al. Fibronectin 1 promotes melanoma proliferation and metastasis by inhibiting apoptosis and regulating EMT. Onco Targets Ther. 2019;12:3207-3221. Published 2019 May 1. doi:10.2147/OTT.S195703
17. Li J, Ding Y, Li A. Identification of COL1A1 and COL1A2 as candidate prognostic factors in gastric cancer. World J Surg Oncol. 2016;14(1):297. Published 2016 Nov 29. doi:10.1186/s12957-016-1056-5
18. Dhawan DK, Chadha VD. Zinc: a promising agent in dietary chemoprevention of cancer. Indian J Med Res. 2010;132(6):676-682.
19. Wang F, Jiao P, Qi M, Frezza M, Dou QP, Yan B. Turning tumor-promoting copper into an anti-cancer weapon via high-throughput chemistry. Curr Med Chem. 2010;17(25):2685-2698. doi:10.2174/092986710791859315
20. Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. Genomics, 99(6), 323-329.