1. **Analyse survival rates of the patients having that data.**
2. **Analyse trend of the cell fractions of each cancer subtype v/s the analysis in the single-cell paper.**
3. **Box plot of the cell fractions of each cell-type for all the cancer stage categories -> include the non-cancer patients data sent by the prof. Convert RPKM to TPM.**
4. **Diagrams: matrices.**
   1. **Ref link: https://www.nature.com/articles/nbt.4096**

**5. Paper reading:** [**http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006106**](http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006106)

[**http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004132#sec007**](http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004132#sec007)

**Normal Data: https://www.gtexportal.org/home/tissueSummaryPage**

**Tasks performed since last meeting:**

1. Extracted marker genes (given as a supplementary from the paper).
2. Extracted cell-types (merged the 564 columns into 12~13) - Average
3. Extracted data for 30 patients and successfully ran CIBERSORT. Got the deconvolution results.
4. Extracted clinical data for 15 patients.
5. Implemented a linear regression model taking as features the results of CIBERSORT(cell-fraction of the cell types) and predicted output as “day\_to\_live” for the 15 patients.

**Problems Faced:**

1. We were able to extract around 550 marker genes. However, on taking a join with the bulk data (on ensembl IDs without dots removed) we got only around 81 genes.
   1. Our thought: We need to increase the bulk samples. However, the bulk samples already have around 60k genes. So, the strange thing is most of them don’t overlap with the 550 marker genes that are obtained. What can be the reason for this?

2. We need a **script to extract bulk data (gene expression of a patient) and clinical data together**

3. Clinical data is in the form of an XML file. So, **we need to convert them into a CSV.**

4. Furthermore, there is inconsistency among the clinical data. (not the same columns for all patients. This will make training the model difficult).

5. What columns should we choose in the clinical data, as the things to be predicted by the model?

6. How many samples to apply CIBERSORT on? (we applied for 30)

**To Do:**

1. Implement a generic code to extract the bulk data and the clinical data.
2. Find way to convert XML to CSV format. And write a script to merge different clinical data CSVs into one CSV. (this can be done once all CSVs have the same columns).
3. Focus HER2 + immune biomarker genes.
4. Try to run for 164 patients. (IHC\_HER2 +ve)

TUMOR\_STAGE

IHC\_HER2

Filter by case ID having IHC\_HER2 positive

1. Cancer Stages - Terrible accuracy due to lack of data (HER2+ - 163 only)
2. How cell type changes across stages
3. KM plot (survival plot)
4. Should we decrease cell types?
5. Cluster patients according to cell fraction results
6. Then do survival analysis (code sent in mail)

Report:

1. Flowchart; Schematic; Figures
2. Result figures

Predict (based on cell-type values from CIBERSORT):

1. **Cluster patients according to cell fraction results.**
2. **Heat Map plot**
3. **Predicting stages of cancer**
4. **Cancer stages: how to convert to numeric values?**
5. **Plot KM Plot**
6. Survival Rate (KM Plot): Days\_to\_Death

Abstract - 115 words

Introduction - ⅗ para 500-600 words

Flow chart

Dataset

Method & Materials(Data) - include all preprocessing - which marker genes etc

Cibersort - cell fraction and which cell types

Results - 1. Pipeline

2. TCGA BRCA - all patients - include all machine learning methods tried

3. HER2+ patients - ml to predict stages -

4. Survival prediction

Conclusion/Discussion - 200 words and future work

Cancer Stony Brook Retreat - sometime in July - poster presentation

**Abstract:**

The goal is to try to predict the cancer stages of patients using cell fractions of cell-types obtained from CIBERSORT.

**Introduction:**

**Dataset:**

The single-cell gene expression dataset was obtained from the NCBI Gene Expression Omnibus database under the accession code GSE75688.

**Ref link:** <https://www.nature.com/articles/ncomms15081>

The bulk dataset, which consisted of gene expression data for 1064 breast cancer patients was obtained from ….the Prof.

The clinical dataset for the 1064 patients was obtained from ...the Prof.

**Method:**

**Get Marker Genes**

**Goal:** Get a list of the marker genes that will be used to get the signature gene matrix to be used for deconvolution.

1. The reference research article mentioned the marker genes for ER+, HER2+, TNBC tumor marker genes and immune gene sets --- “Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer”.

**Get Signature Genes**

**Goal:** Extract only those rows that contain marker genes from the entire single-cell dataset.

1. The GSE75688 dataset contains data for about 58K genes, deconvolution is performed only on the set of marker genes to obtain the relative cell fractions according to cell-type. Hence, only the set of marker genes were extracted from the GSE75688 dataset in order to perform single-cell deconvolution.
2. Extracted the markers genes mentioned above from the GSE75688 dataset of single cell rna-sequenced data by mapping the gene\_name of the two files.

**Single Cell Matrix Generator**

**Goal:** Generate the signature gene matrix of the form: (marker genes X cell-types)

1. The GSE75688 dataset contained data in the form of genes vs cells(approximately 550), in order to perform deconvolution, the single cells were categorized into cell-types by averaging their TPM values and taking log2(TPM + 1).
2. This is the Signature Gene Matrix (genes vs cell-types)

**Bulk Tissue File Generator**

**Goal:** Preprocess the TCGA\_BRCA dataset to obtain a file of the form: (genes X samples).

TCGA\_BRCA dataset contains the samples vs genes where the relative cell fractions were in FPKM.

1. Converted the FPKM values to TPM values using the conversion formula

TPM = FPKM / (sum of FPKM over all genes/transcripts) \* 10^6

1. Preprocessed the gene\_name in order to get them in the same format as our Signature Gene Matrix
2. This is the genes vs samples bulk tissue file

**Bulk Single Joiner**

**Goal:** To get both the bulk tissue file and the signature gene matrix file to contain the same marker genes in their rows.

1. Performing deconvolution in CIBERSORT requires the marker genes to be the exact same in both the Signature Gene Matrix and the Bulk Mixture Matrix
2. Joined the Signature Gene Matrix and the bulk tissue file on gene\_name
3. Extracting particular columns gives the final Signature Gene Matrix and Bulk Mixture Matrix

**CIBERSORT**

**Goal:** Apply single-cell deconvolution via CIBERSORT to obtain the relative cell-type fractions for each sample.

1. Fed the Signature Gene Matrix and Mixture Matrix to generate the Output file that contains the relative cell-type fractions of each sample

**Clinical Data Pre-processing**

**Goal:** To get the clinical data for the bulk data samples (patients).

1. Load the bulk data and clinical data files.
2. Join them on CASE ID column to obtain the common sets of samples from both files.
3. Save these clinical data and bulk data files.

**Cluster Patients**

**Goal:** Cluster samples(patients) according to their similarity of the cell fractions.

1. Used decision trees, dendograms, hierarchical and k-means clustering (4) methods to check if there exists some patterns among the cancer patients.

**Predict Stages**

**Goal:** Come up with a good machine learning model that can predict the stage of breast cancer based on the cell-type fractions for each patient.

1. Categorized the 12 sub-stages of cancer into three types (Stage I, IA and IB as 1, Stage II, IIA, IIB as 2 and Stage III, IIIA, IIIB, IIIC, IV and X as 3), so that the problem can be modelled as a multi-class classification problem.
2. As the classes were imbalanced, applied two different sampling methods to try and balance the classes for better accuracy. The two methods used were as follows:
   1. SMOTE Sampling - Synthetic Minority Over sampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors, joins them and creates the synthetic samples in the space. The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors. The difference is multiplied by random number between (0, 1) and it is added back to feature. (ref:https://www.datasciencecentral.com/profiles/blogs/handling-imbalanced-data-sets-in-supervised-learning-using-family
   2. ADASYN Sampling - ADAptive SYNthetic (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor. The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data. The algorithm uses Euclidean distance for KNN Algorithm.
3. Applied 8 types of machine learning models (of the Scikit-learn library) to check which model gives the best accuracy in terms of cancer stage prediction:
   1. Multi-Class Logistic Regression.
   2. Random Forest Classification.
   3. Linear SVC.
   4. Linear SVM.
   5. kNN Classifier.
   6. Naive Bayes Classifier.
   7. Convolutional Neural Network (CNN).
   8. Decision Tree.
4. Applied 5-fold cross validation to prevent overfitting and be confident in the accuracy of the constructed model.

Given below are the accuracies of the model shown in a tabular format:

Invalid

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | Model | Accuracy (with 5-fold cross validation on ADASYN sampling) | Accuracy (with 5-fold cross validation on SMOTE sampling) |
| 1 | Multi-Class Logistic Regression | 42.98%%  (multinomial =42.27%) | 46.69%  (multinomial = 46.42%) |
| 2 | Random Forest | 63.62% | 66.40% |
| 3 | Linear SVC (parameter tweaking?) | 42.32% | 45.83% |
| 4 | Linear SVM (C=10, gamma =10) | 69.85% | 69.76% |
| 5 | kNN Classifier (n\_neighbors = 7) | 63.08% | 64.74% |
| 6 | Naive Bayes Classifier | 41.18% | 45.67% |
| 7 | CNN | 51.34% | 55.55% |
| 8 | Decision Tree | 50.41% | 55.45% |

**Results:**

1. 8 types of classifiers were trained based on the BRCA dataset to obtain a model that could predict the stage of breast cancer for a patient.
2. Initial training on the raw dataset yielded low prediction accuracy scores due to the dataset being unbalanced (biased towards the stage II cancer stage).
3. Hence, the dataset was balanced through oversampling using two methods: SMOTE and ADASYN. This drastically improved the cancer stage prediction accuracy scores.
4. Overall, the top 3 models were: a) SVM (C=10, gamma=10, SMOTE sampling) with an accuracy of ~70% b) Random Forest (SMOTE sampling ) with an accuracy of ~66% c) kNN classifier (SMOTE sampling) with an accuracy of 65%.

**Conclusion and Discussion:**

Using single-cell deconvolution technique CIBERSORT on the TCGA-BRCA patients dataset we were able to successfully obtain the relative cell-fraction of the required cell-types for each patient. Based on the results obtained from CIBERSORT, we were also able to successfully implement an SVM model that could predict the breast cancer stage with around 70% accuracy. We also tried to cluster the cancer patients using hierarchical and k-means clustering to observe if patients were clustered according to the cancer stage type or not.

Future Work:

1. Experimenting with other single-cell deconvolution techniques (besides CIBERSORT) to obtain cell fraction of samples and then use these fractions as features to create new models and compare them to the models based on CIBERSORT generated cell fractions to see if it leads to better prediction accuracies.
2. Improving the accuracy of the cancer stage prediction model through:
   1. Hyper-parameter tuning of the best models.
   2. Obtaining a larger training dataset on cancer patients.

Method:

1. Classify DAYS\_TO\_DEATH and CANCER\_STAGES into categories so that we can apply classifying models like: Naive Bayes Classifier.

CANCER STAGES divided into:

1. Early Stages
2. Mid Stage
3. Late Stage

DAYS\_TO\_DEATH divided into:

1. <2 years
2. 2 to 5 years
3. >5 years
4. Try different sampling methods: SMOTE Sampling
5. Try to get a probabilistic distribution (empirical PDF) of days\_to\_death to fill in the missing values.

Can try taking log values of the cell fractions

Models we’ll try to predict “Survival Rate”:

**Dataset Processing Information:**

1. SMOTE sampling applied to balance the dataset increase the samples of each category to 624.
2. Train set size: 80% Test set size: 20%

Experiments:

1. Hyper Parameter tweaking
2. Different sampling methods
3. One-hot encoding -> all methods
4. K-fold cross validation
5. Top Features?

Invalid

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | Model  (SMOTE sampling) | Accuracy (Test set) | Accuracy (with 5-fold cross validation) |
| 1 | Multi-Class Logistic Regression | 44.26%  (multinomial = 44.8%) | 46.69%  (multinomial = 46.42%) |
| 2 | Random Forest | 69.06% | 66.40% |
| 3 | Linear SVC (parameter tweaking?) | 44% | 45.83% |
| 4 | Linear SVM (C=10, gamma =10) | 70.4% | 69.76% |
| 5 | kNN Classifier (n\_neighbors = 7) | 64.8% | 64.74% |
| 6 | Naive Bayes Classifier | 46.4% | 45.67% |
| 7 | CNN | 55.1% | 55.55% |
| 8 | Decision Tree | 58.67% | 55.45% |

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | Model | Accuracy (with 5-fold cross validation on ADASYN sampling) | Accuracy (with 5-fold cross validation on SMOTE sampling) |
| 1 | Multi-Class Logistic Regression | 39.26%  (multinomial =39.91%) | 40.44%  (multinomial = 40.44%) |
| 2 | Random Forest | 63.18% | 64.91% |
| 3 | Linear SVC (parameter tweaking?) | 39.97% | 40.65% |
| 4 | Linear SVM (C=10, gamma =10) | 61.82% | 60.63% |
| 5 | kNN Classifier (n\_neighbors = 4) | 60.90% | 64.10% |
| 6 | Naive Bayes Classifier | 39.97% | 42.58% |
| 7 | CNN | 44.52% | 44.66% |
| 8 | Decision Tree | 47.89% | 54.76% |

Challegenes: (1) computational (2) biological estimate cell fractions

Summary: our analysis

Results: (1) flowchart, we built a pipeline for cancer single cell decov. (2) estimate cell fractions breast cancer (3) associate cell fractions with clinical outcomes, survival rates (4) comparison across breast cancer subtypes (her2, …) vs single cell paper

Materials and Methods: (1) datasets (bulk tissue, single cell) (2) data processing (FPKM-TPM) (3) classifiers (4) accuracy

Discussion/Conclusion

200-300 words