Gayatri Ratakonda

Yorktown High School

Research Plan– Analyzing CNA Patterns to Determine the Efficacy of Breast Cancer Treatment

Rationale –

Breast Cancer is an incredibly deadly disease affecting approximately 1 in 8 women at some point during their lifespan (BC.org, 2017). Due to the advancement of modern technology, the 5-year and 10-year survival rate for breast cancer patients is approximately 90% and 83% respectively after treatment has been recovered from, making it highly treatable. The problem for most cancer types, breast included, is in catching the cancer early before it is able to metastasize beyond its starting point. If the patient has metastatic breast cancer at the time of diagnosis, the 5-year survival rate drops dramatically to 27% due to it having left its original tissue and spreading throughout the body (CDC.gov, 2017). In order to effectively diagnose a patient, mammography tests need to be conducted that require invasive biopsy and blood work in order to search for possible genomic indicators of the cancer. Breast cancer specifically has many of these indicators, BRCA1, BRCA2, HERC2, and ERBB2 being amongst the most commonly appearing (Osborne, 2004). These tests are both inefficient and inconvenient for the patients due to them taking approximately 2 hours to conduct, and up to a week for analysis. Proposed solutions to biopsies are direct DNA samples but breast cancer oncogenes are only present in the affected tissues, so their presence won't appear in samples from other places in the body. In order to analyze patients for breast cancer genes, a method other than a direct test must be used. Copy Number Aberrations (CNAs) are changes in the DNA of an individual and are found everywhere in the body. These changes are specifically in the number of times a specific gene is expressed. The presence of a CNA can mean the deletion or amplification of a gene. Through analysis of the patterns of CNAs, it has been found previously that CNAs can be indicative of the

physical type of breast cancer (lobular vs. ductal carcinomas) (Chi, 2018). By analyzing CNA patterns of 2,433 breast cancer patients from a publicly viewable Redcap dataset, patterns indicative of those oncogenes can also be found. An established correlation between the oncogenes and CNA patterns could lead to the elimination of invasive testing. On top of this, the treatment that is most likely to be successful can also be deduced from the copy number aberration (CNA) pattern observed. Through analyzing the CNA pattern of these patients, both the processes of diagnosis and treatment can be simplified for future patients.

Research Questions –

- Can the analysis of CNA patterns of breast cancer patients with different activated tumor oncogenes lead to the observation of a few key patterns indicative of a certain type of breast cancer?
- Can the indicative copy number aberration (CNA) patterns allow for better prognostics due to being linked with one successful type of treatment?

Hypotheses –

- There will emerge a few key copy number aberration (CNA) patterns common in almost all breast cancer patients of the same type, allowing for a cheap and noninvasive diagnostic procedure (Alonso, 2017).
- The CNA patterns determined in the first phase can be used to determine the most effective treatment for that tumor specifically.

Procedures –

Role of Mentor – The mentor will instruct the student on which programming software (Pandas, Matplotlib, and SciKit Learn) and what program will best suit the type of analysis being done.

Role of Student – The student will carry out all programming procedures, format the data so it is usable in the Jupyter system, and utilize the program under the guidance of the mentor to analyze the CNA data. All data that will be utilized in this study will be from a publicly available source and I will not have any contact with any participants. All data will be de-identified and anonymous.

Jupyter Notebook

Formatting

· The 2,433 pieces of anonymous and de-identified data will be sorted according to their MB number assigned by the people who built the database (Pereira, 2016). No contact with live participants will occur. They will compare 47 different parameters from age to tumor stage to individual survival rate. The clinical data along with the patient data will then be merged by the student through the use of the Jupyter notebook import system in order to condense the possible parameters for the study.

Matplotlib

· Matplotlib is a piece of Jupyter software manufactured by John D. Hunter in 2003 that works mainly to plot different parameters in relation to each other. By inputting values

for two parameters, a logistic regression is plotted with the level of correlation being shown in the slope of the plotted line.

· 16 of the 173 analyzed genes for CNA patterns provided from the dataset will be selected for further CNA analysis by looking at their correlations through regression. Genes with higher correlation with the "DIED WITHIN FIVE YEARS" parameter will be chosen for the study.

SciKit Learn

· SciKit Learn is a Python learning library. It will mark the correlation between the observed copy number aberration (CNA) pattern for the patient and their observed positive gene traits. This is important in order to determine if the CNA pattern can effectively determine the presence of this gene.

· The CNA patterns for each of the patients will be recorded, and the correlation factors to genes will be shown. The SciKit Learn software will issue an accuracy prediction of how many times given 100 samples it will guess the gene positives correctly given the CNA pattern.

Tensor Flow

· Tensor Flow is a neural network that is used for machine learning. A neural network is an adaptive program that takes in data from an outward source and builds a program that can detect changes from the normal set by the given test data. Once it takes into account a change, it can describe how irregular it is and how it can be recognized as irregular.

· The established copy number aberration (CNA) correlation will give the machine an input of values from the CNA dataset. Tensor flow will establish these to be the "norms"

for each gene. When test data is inputted, the formula for detecting the presence of genes will adapt to the extra parameters, such as age and stage of tumor at detection.

· The model can then take into account the two parameters, type of cancer treatment and survival rate in order to determine the most effective breast cancer treatment given the CNA data available for each sample.

References –

1. Alonso, M. H., Aussó, S., Lopez-Doriga, A., Cordero, D., Guinó, E., Solé, X., . . . Moreno, V. (2017, July 06). Comprehensive analysis of copy number aberrations in microsatellite stable colon cancer in view of stromal component. Retrieved from https://www.nature.com/articles/bjc2017208
2. Chi, C., Murphy, L. C., & Hu, P. (2018, January 29). Recurrent copy number alterations in young women with breast cancer. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5837756/
3. Hegselmann, S., Gruelich, L., Varghese, J., & Dugas, M. (2018, November 29). Reproducible Survival Prediction with SEER Cancer Data. Retrieved from http://proceedings.mlr.press/v85/hegselmann18a.html
4. Hieronymus, H., Murali, R., Tin, A., Yadav, K., Abida, W., Moller, H., . . . Sawyers, C. L. (2018, September 04). Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6145837/
5. Pereira, B., Chin, S., Rueda, O. M., Vollan, H. M., Provenzano, E., Bardwell, H. A., . . . Caldas, C. (2016, May 10). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Retrieved from https://www.nature.com/articles/ncomms11479
6. STAFF, C. (2019, February 28). Breast Cancer - Statistics. Retrieved from https://www.cancer.net/cancer-types/breast-cancer/statistics