Gene Selection and Clustering of Breast Cancer Data

Farzana Ahamed Bhuiyan,[1] MD Bulbul Sharif,[1] Paul Joshua Tinker,[1] William Eberle,[1]
Douglas A. Talbert,[1] Sheikh Khaled Ghafoor,[1] Lewis Frey,[2]
{fbhuiyan42, msharif42, pjtinker42}@students.tntech.edu, {weberle, dtalbert, sghafoor}@tntech.edu, frey@musc.edu
[1]Department of Computer Science, Tennessee Technological University, Cookeville, TN 38501
[2] Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC 29425

Abstract

In this work, we first attempt to reproduce an earlier study on gene selection and clustering, and then we extend this work by applying a different type of hierarchical clustering to discover interesting subsets of genes from breast cancer data. Reproduction of such studies is a known challenge and an active area of research in bioinformatics. The work presented in this paper is three-fold. First, we reproduce a study conducted at the University of North Carolina to generate an initial set of genes. Second, we apply an approach called Distance Weighted Discrimination to fuse multiple, disparate breast cancer datasets into a single validation set. Third, we perform hierarchical clustering and k-means clustering on this validation set to discover natural groupings and compare the clusters generated by both methods. While applying the hierarchical clustering is part of the reproduction step, we extend the research by trying two different forms of hierarchical clustering. We also apply k-means clustering for the same purpose and compare all three methods using Kaplan-Meier estimation and Cox proportional hazards regression. We discover that among the three methods, k-means clustering gives us the best results.

Introduction

For women in the U.S., the death rate due to breast cancer is second only to lung cancer. Moreover, breast cancer is a highly heterogeneous disease with different molecular subtypes. Precision medicine (Collins and Varmus 2015; Frey, Bernstam, and Denny 2016) is an approach that takes into account individual variability in genes, environment, and lifestyle to better predict which treatment and prevention strategies for a particular disease will work in which groups of people. Considering the severe complication, it is imperative that researchers engage precision medicine in breast cancer diagnosis and treatment, and for that, a clearer idea about the genomes of breast tumors needs to be realized.

Unfortunately, computational studies, such as those that identify connections between genetic patterns and cancer characterization, have often proven to be difficult to reproduce (Ioannidis 2009), and work is ongoing to improve the reproducibility (Ioannidis 2009; Sandve et al. 2013; Nekrutenko and Taylor 2012). Thus, our first objective in this work is to reproduce existing research (Hu 2006). Using the same data and procedures, we obtain similar results from the conduct of an independent study where our techniques are as closely matched to the original experiments as possible.

Currently, mankind has the capacity to capture and analyze biological information at the genetic level. However, due to the expense involved in the collection and processing of biological samples, most datasets are comprised of very few samples. This has led some researchers to explore methods for merging data sets from various studies into a single set. As with the original work, to generate a validation set we have merged four publicly available breast cancer expression datasets using Distance Weighted Discrimination (DWD) (Benito 2004a).

Finally, we perform both hierarchical clustering and k-means clustering to seek interesting sub-populations of genes. Much of the previous analyses of gene expression data to classify breast tumors uses hierarchical clustering. However, we will extend the work by trying different forms of hierarchical clustering and also k-means clustering to discover distinctive molecular portraits of each tumor. We will show that we can successfully achieve high reproducibility in identifying most of the subgroups previously identified in other studies. Using the Kaplan-Meier survival analysis and Cox proportional-hazards model, we will compare the differences in outcomes and associations with other clinical parameters between each of the groups. While the hierarchical clustering technique will give us some distinct subgroups of genes, we show that k-means clustering perform better to identify the distinct subgroups which are consistently predictive of a patient's clinical outcomes as evidenced by the prediction of Relapse-free survival (RFS) and Overall survival (OS) of each identified group.

Related Work

Human breast tumors vary in their natural history and treatment responsiveness. It is proposed that the phenotypic diversity of breast tumors might be accompanied by a corresponding diversity in gene expression patterns (Perou 2000). They show that the tumors could be classified into subtypes that differ widely in their patterns of gene expression. We used a reference paper (Hu 2006) for our study where we reproduce some of their work and also used the paper to com-

pare our results. Hu et. al sought to determine a new breast tumor intrinsic gene list from a training set. Then this gene set was used to predict the survival rate. They created a single data set from three different datasets using DWD (Benito 2004a) to validate their derived gene list for the prediction of breast cancer type. Finally, to identify hierarchical clusters, they used the data set to obtain a distinctive molecular portrait of each tumor type.

In a similar study (Sørlie 2001) the authors found that tumor classification based on gene expression patterns can be used as a prognostic marker in a subset of patients receiving uniform therapy with respect to OS and RFS. In another study (Sørlie 2003), they tried to redefine the previously defined breast tumor subtypes that could be distinguished by their distinct gene expression patterns. Their results support the idea that many of these subtypes of breast tumors are biologically separate disease entities.

Breast cancer is such a complicated disease that even the strongest metastasis predictors can not precisely classify breast tumors according to their clinical behavior. In this paper (Van't Veer 2002), they used supervised classification to identify a short-interval gene expression signature in patients without tumor cells and hierarchical clustering to cluster tumors based on their measured similarities over their significant genes. In their paper (Sotiriou 2003), the authors used microarray technology to examine thousands of genes simultaneously for the molecular classification of human cancers. They carried out unsupervised cluster analysis to find natural groups in the profiles. Although their sample size was not large enough to determine the high reproducibility of smaller subgroups, they could find several subgroupings previously identified in other studies in their dendrogram.

Data

Each of the experiments mentioned previously was conducted using the relative gene expression abundance found in a given tumor sample. Gene expression can be thought of as gene transcription, in which the DNA is copied in a gene to produce an RNA transcript called messenger RNA (mRNA). A simple microarray is comprised of a solid surface covered in thousands of microscopic divots. Each divot contains many synthetically produced, single-stranded DNA, which represent a particular gene. Two groups of sample mRNA are prepared and then converted to complementary DNA (cDNA) and then combined and linked to the microarray slide. When exposed to certain laser lights and temperatures, the divots will emit levels of fluorescence. By measuring the wavelengths of the fluorescence, one can measure the abundance of experimental vs. control that bound to a particular gene's site, thus yielding the relative abundance.

For our experiments, we will be accessing data from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (Sørlie 2001), using the breastCancerNKI package (Schroeder 2011) of Bioconductor (Gentleman 2004), and comparing our results against previous work (Hu 2006; Sotiriou 2003). Storing this data so that it can be easily accessed while not losing fidelity is not a trivial task. Each new batch of samples has the potential of coming from a different group of researchers, being prepared with differing protocols, and being analyzed on ever-changing technology. The designers of GEO have incorporated several methods for encapsulating all of the data relevant to each study. We chose to use the Matrix format (Sørlie 2001). Matrix is preprocessed format of data, so we did not need any normalization, summarization, or filtering. These files can contain up to three main sections of data: phenotype, feature, and expression data. Phenotype data contains information about the origin of the sample genome being analyzed. For our studies, this included age, survival rate, tumor size, and several other pieces of clinical information useful in classifying the tumor. Feature data describes the microarray platform used in the experiment, the subsequent data gathered, and identifiers such as gene name and gene symbol. Expression data is the abundance value assigned to that particular sample by the microarray analysis, represented as a matrix of real numbers.

Feature Extraction

As our first step of reproduction, we derived an objective "intrinsic subtype" classifier that can be used clinically. For this purpose, first, we collected a breast tumor intrinsic list from the reference paper (Hu 2006). For developing this gene set, 9 normal breast samples and 105 breast tumor samples were taken as the training set. These raw data were filtered to include only features with single intensity 30 units over background in both Cy5 and Cy3 channels and for which this signal intensity criteria was met in at least 70% of the samples. After that, the mean of the non-missing expression values was computed separately in each batch on a gene by gene basis (Sørlie 2003). Then for each sample, the batch mean for that gene was subtracted. For each gene, the average "within-pair variance" and "between-subject variance" were computed and the ratio D = (within-pair variance) / (between-subject variance) was computed. The genes with a small value of D were declared to be intrinsic. The choice of a value of D as a cutoff was somewhat arbitrarily set at one standard deviation below the average. Following this, an intrinsic gene set was identified consisting of 1410 microarray elements representing 1300 genes.

Data Preparation

To evaluate the intrinsic gene set we need an independent test dataset. In microbiology, if the size of the validation dataset is too small validation becomes unconvincing. For this reason and following (Hu 2006), we generated a validation set by merging four publicly available breast cancer expression datasets using DWD (Benito 2004a) and used it to show the clinical significance of our intrinsic classifications. The datasets that we used were Stanford datasets (Sørlie 2001; 2003), Rosetta dataset (Van't Veer 2002) and Singapore dataset (Sotiriou 2003). While three of these four datasets were also used in (Hu 2006), they are not the exact same dataset but an extension of those datsets with more samples.

However, getting the datasets was not a trivial task. For

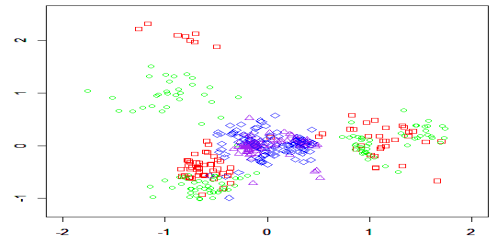


Figure 1: MDS Plot Before Applying DWD Method

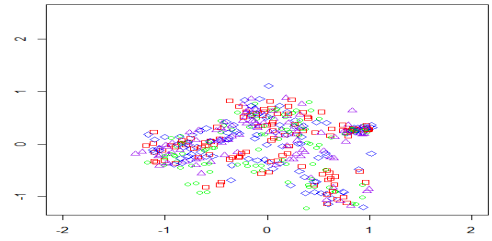
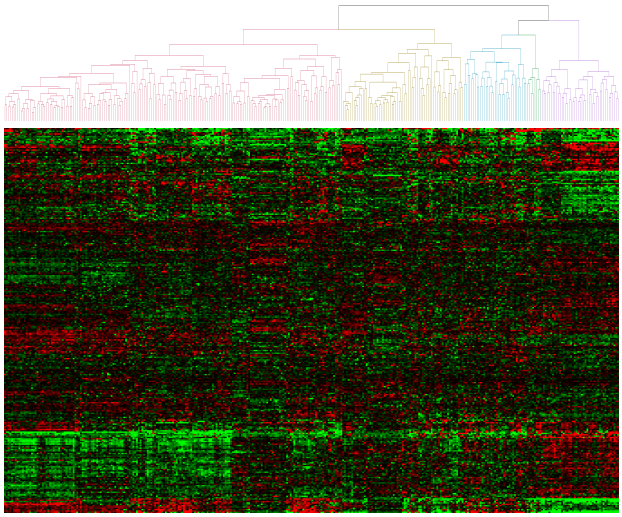


Figure 2: MDS Plot After Applying DWD Method

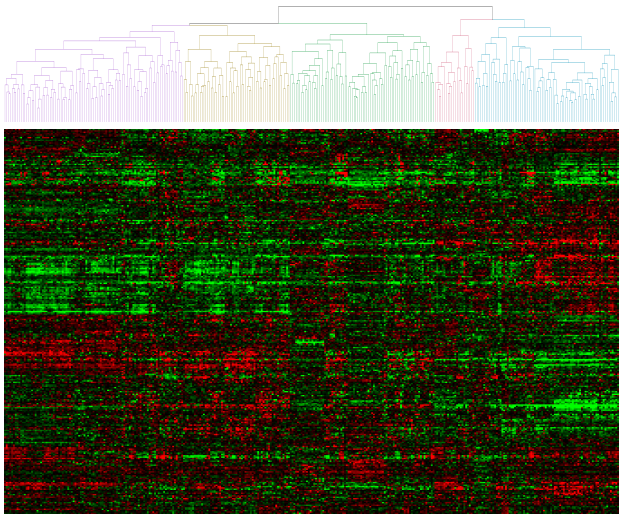
(a) Dendrogram



(b) Heatmap

Figure 3: Hierarchical Clustering Using Euclidean Distance

(a) Dendrogram



(b) Heatmap

Figure 4: Hierarchical Clustering Using Pearson Uncentered Correlation

the Stanford (Sørlie 2001; 2003) datasets, we first downloaded 8 matrix files of the same series but of different platforms and used the Geoquery R package (Davis and Meltzer 2007) to load the data as expression sets and the inSilicoMerging package (Taminau 2012) to merge them into a single expression set. Then we took the expression data and the feature data and formed a table. For the Rosetta (Van't Veer 2002) dataset we loaded the matrix data from the breastCancerNKI package (Schroeder 2011) in an expression set format. We again took the expression data and the feature data to form a table. Finally, for the Singapore (Sotiriou 2003) dataset, we found the data from the supplementary resources of the paper directly in the table format.

In each of the datasets, there were missing values for each tuple. For each dataset, we used the "Gene symbol" as the primary key, and then we deleted the records with a missing primary key. For the other missing attributes, we did random imputation to fill in the missing feature values with appropriate values using the KNN method (Hu 2006). However, a single gene can have multiple gene symbols. To get something unique, we converted the symbols to "Gene ID" using a conversion tool (Kim 2015). For some gene samples, we could not find the corresponding Gene ID. Therefore, after the ID conversion, we deleted all the rows for which Gene ID was not found.

Then, we used DWD to combine the datasets together using the DWD Java tool (Benito 2004b). Systematic differences due to experimental features of microarray experiments are present in most large microarray data sets. Many different experimental features can cause biases including different sources of RNA or different microarray platforms. From Figure 1, one will notice that if we simply merge the data without applying DWD, there would be a very strong dataset bias. Performing DWD, we first combined the two Stanford datasets, and then combined this with the Rosetta dataset and finally the Singapore dataset. As shown in Figure 2, after applying DWD, all the datasets are mixed together, and the biases are removed in the merged dataset. Figures 1 & 2 are generated using the Bioconductor package "limma" (Ritchie 2015). Finally, we found 288 genes that are present both in the combined dataset and the intrinsic gene list, and we will perform our experiments using these 288 genes.

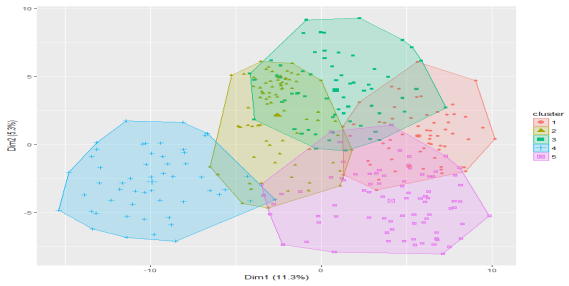


Figure 5: k-means Clustering

Experimental Setup

We performed both hierarchical clustering and k-means clustering on our combined test set of 288 genes. Since we used the combined set, not the 4 sets separately, we believe a more meaningful result will be realized as any interesting findings would need to be present across all four sets. To find the tumor subtypes, first, we performed agglomerative hierarchical clustering which was also done in the reference paper (Hu 2006). For this purpose, we used the software Cluster 3.0 (Eisen and de Hoon 1998), which was developed based on the work in (Eisen 1998). We computed the distance using two different methods. First, we tried Euclidean Distance (Eisen and de Hoon 1998) and second, we tried the uncentered correlation based on Pearson correlation. We drew the dendrograms and heatmaps of Figures 3 and 4 using the library called "dendextend" (Galili 2015) and "heatmap.2" (Warnes 2016) respectively. These figures help us visualize and compare trees of hierarchical clustering. We used the dendrogram branching pattern in Figure 3a and 4a and the knowledge of the reference paper (Hu 2006) to determine the number of biologically relevant tumor subtypes within the cluster.

Then we performed k-means clustering. Since in k-means clustering we need some chosen number of clusters (k), we used the same number of clusters that we found from our hierarchical clustering analysis. As our distance measure, we used Euclidean Distance. We did the ggplot2-based elegant visualization of the k-means method shown in Figure 5 using the CRAN- Package "factoextra" (Kassambara 2017).

After finding the groups, we compared the results and associations with other clinical parameters between each of the groups. We did Kaplan-Meier survival analysis on the combined set, to find out whether the groups were predictive of RFS and OS. Here a curve can be considered an estimate of the survival curve for all people with the same circumstances. As shown in Figures 6 and 7, any point on the curve gives the proportion surviving at a particular time after the start of the experiment (Rich 2010). Each step goes down in the curve when they have a measured relapse of the disease. If the curves are flattened, it suggests that patients have gone into remission. Finally, we compared these Kaplan-Meier survival plots using the Co model.

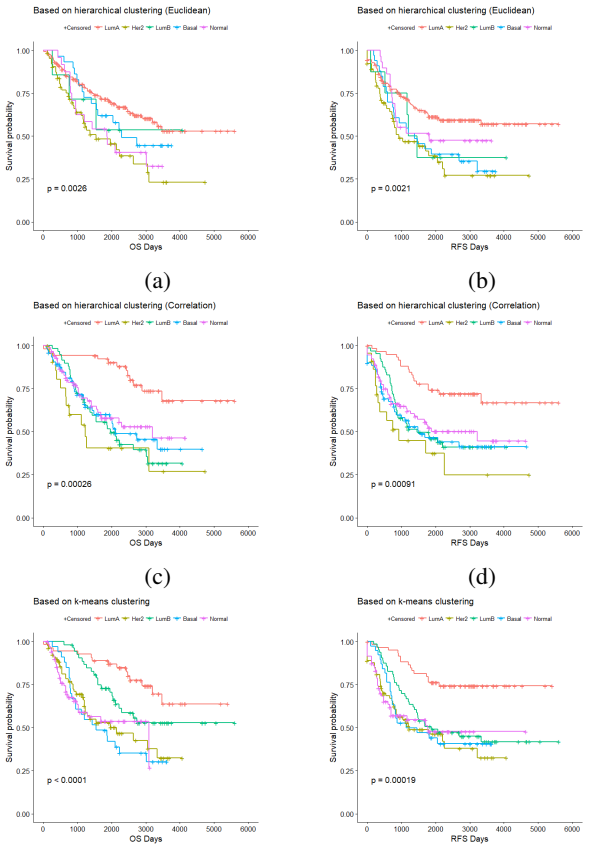


Figure 6: Kaplan-Meier survival curves of breast tumors subtypes. (a) OS curve classified by Hierarchical Clustering using Euclidean Distance (b) RFS curve classified by Hierarchical Clustering using Euclidean Distance (c) OS curve classified by Hierarchical Clustering using Uncentered Correlation (d) RFS curve classified by Hierarchical Clustering using Uncentered Correlation (e) OS curve classified by k-Means Clustering (f) RFS curve classified by k-Means Clustering

Results

Based on the dendrogram branching pattern (Figure 3a and 4a) and our knowledge of the previous classifications made by Zhiyuan et. al (Hu 2006), we identified 5 potential groups within the cluster in Figure 3 and 4. From these five biologically relevant tumor groups (figure 6 and 7), we proceeded to look for differences in outcomes and associations with other clinical parameters

In Kaplan-Meier curve analysis, from the "p" value we can get an overall significance of the model, which in our case defines whether our identified groups have a significant influence on survival time. A p-value that is less than 0.05 is considered to be significant. We can see from Figure 6 that the Kaplan-Meier curves based on our classified subgroups shows highly significant differences in OS and RFS between the subgroups.

In the combined test set, the standard clinical parameters

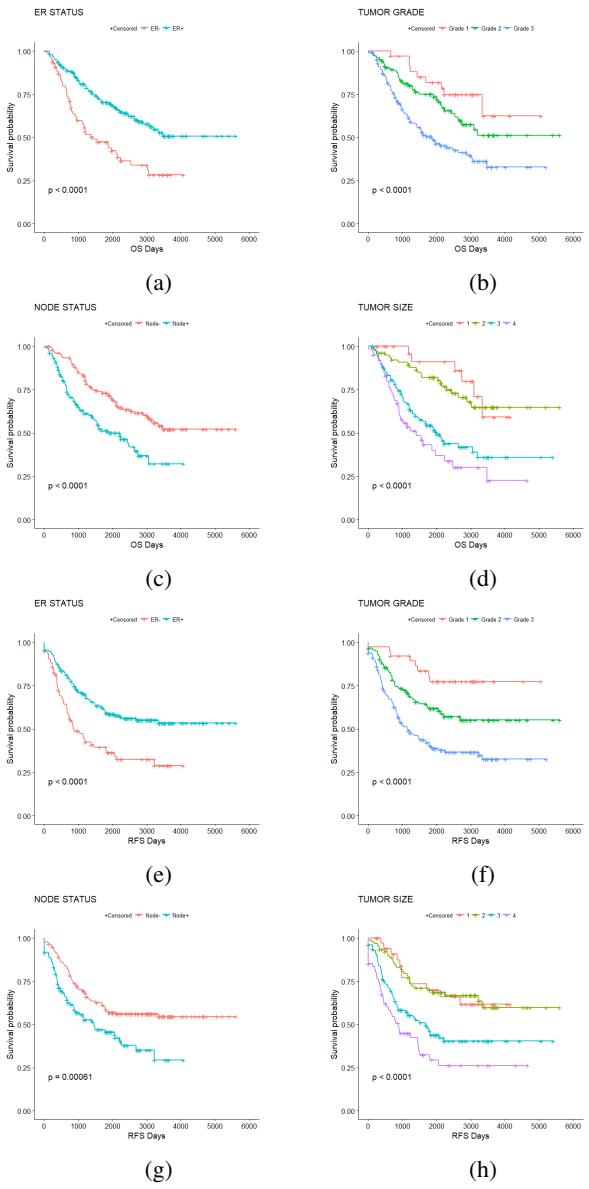


Figure 7: Kaplan-Meier survival curves for common clinical parameters. (a) OS curve for ER status (b) OS curve for Tumor Grade (1= well-differentiated, 2= intermediate, 3= poor) (c) OS curve for Node Status (d) OS curve for Tumor Size (1: diameter $\leq$ 1cm; 2: diameter $\leq$ 2cm; 3: diameter $\leq$ 3cm; 4: diameter $>$ 3cm (e) RFS curve for ER status (f) RFS curve for Tumor Grade (g) RFS curve for Node Status (h) RFS curve for Tumor Size

of ER status, grade, node status, and tumor size were important predictors of both OS and RFS using Kaplan-Meier analysis (Figure 7), demonstrating that the combination of four different datasets did not destroy these standard markers' prognostic abilities. To investigate the ability of each clustering method to give us new information about breast cancer types, we performed a comparison of the approaches and showed differences in the survival models (OS and RFS) among the approaches using a multivariate Cox proportional hazards analysis (see Tables 1 and 2). We also put them in the context of known clinical relevant variables (see Tables 3 and 4).

Though we did not show the survival curve for age, we included age in the tables 3 and 4, which was a continuous variable formatted as decade-years. In the tables, the column marked "z" indicates the Wald statistic value which evaluates whether the beta (β) coefficient of a given variable is statistically significantly different from 0. Also, as evident from Tables 1 and 2, the p-values for all three methods are significant, indicating that the models are significant. From the values in the tables, we can conclude that k-means clustering has highly statistically significant coefficients. Among the two versions of hierarchical clustering, the one with an uncentered correlation has a more significant coefficient than the one with Euclidean Distance.

Table 1: Cox proportional hazards analysis of clustering methods in relation to OS

Clustering Method	coef	exp(coef)	se(coef)	z	p	liklihood ratio test
Hierarchical Clustering Using Euclidean Distance	0.127	1.136	0.060	2.12	0.034	4.24
Hierarchical Clustering Using Uncentered Correlation	0.1583	1.1715	0.0628	2.52	0.012	6.61
k-means Clustering	0.2236	1.2506	0.0644	3.47	0.00052	12

Table 2: Cox proportional hazards analysis of clustering methods in relation to RFS

Clustering Method	coef	exp(coef)	se(coef)	z	p	liklihood ratio test
Hierarchical Clustering Using Euclidean Distance	0.1155	1.1224	0.0544	2.12	0.034	4.28
Hierarchical Clustering Using Uncentered Correlation	0.1449	1.1559	0.0575	2.52	0.012	6.58
k-means Clustering	0.1908	1.2102	0.0587	3.25	0.0012	10.5

Table 3: Cox proportional hazards analysis of standard clinical factors in relation to OS

standard clinical parameter	coef	exp(coef)	se(coef)	z	p	liklihood ratio test
Age	0.02132	1.02155	0.00698	3.06	0.0022	9.13
ER Status	-0.800	0.450	0.184	-4.33	0.000015	17.6
Grade	0.607	1.835	0.146	4.15	0.000033	19.5
Node Status	0.713	2.040	0.185	3.86	0.00011	14.8
Tumor Size	0.611	1.842	0.109	5.61	0.00000021	33.4

Table 4: Cox proportional hazards analysis of standard clinical factors in relation to RFS

standard clinical parameter	coef	exp(coef)	se(coef)	z	p	liklihood ratio test
Age	0.00779	1.00782	0.00686	1.13	0.26	1.27
ER Status	-0.714	0.489	0.171	-4.17	0.00003	16.3
Grade	0.682	1.979	0.139	4.92	0.00000084	28.2
Node Status	0.571	1.770	0.168	3.4	0.00068	11.3
Tumor Size	0.5144	1.6726	0.0978	5.26	0.00000014	28.8

Another feature of high importance is the sign of the regression coefficients (coef) in the Cox model results. A positive sign means that the hazard (risk of death) for subjects with higher values of this variable is higher and therefore the prognosis worse. From Tables 3 and 4, we can see that for ER status we get the best prognosis. The p-value for ER-status is 0.000015, with a hazard ratio HR= 0.450, indicating a strong relationship between the ER-status of the patient and a decreased risk of death. The hazard ratios of covariates can be interpreted as multiplicative effects on the hazard. For instance, keeping the other covariates constant,

being ER+ reduces the risk by a factor of 0.45. So being ER+ is associated with a good prognostic. Similarly, the p-value for node-status is 0.00011, with a hazard ratio HR = 2.040, indicating a strong relationship between node-status and increased risk of death. That means that higher values of node-status is associated with a poor survival, holding the other covariates constant.

Discussion

We were successfully able to reproduce an experiment and build our own experiments on top of that baseline. However, since we do not have enough domain knowledge, we do not classify the tumor samples as the original paper (Hu 2006) did. In our experiments, the merged dataset shows significant predictors of outcome both in Kaplan-Meier survival analysis and Cox proportional hazards analysis. Moreover, in spite of reducing the intrinsic set to only 288 genes, our classification was still able to predict outcomes successfully. Finally, performing multivariate analysis, we found that tumor classification based on gene expression patterns can be used as a prognostic marker for OS and RFS in a subset of uniform therapy patients.

Conclusion

Using this study, we might have a much clearer picture of the genomes of breast cancer and can generate data about the intrinsic characteristics of a tumor, thereby providing useful diagnostic, prognostic, and predictive information. Comparisons of the clustering methods give us different perspectives on the genes involved in breast cancer subtypes. All of this can lead to clinical research that makes truly personalized breast cancer medicine possible.

Future Work

Since the main goal of these paper is reproduction, we do not experiment with the clustering factors. In future we would like to build a model with both clinical and clustering factors and analyze the combined model with respect to the model based only on standard clinical factors. We would also like to create a lifelong machine learning system that will learn continuously from gene expression sets and will adjust to new situations (Liu 2018). We are also interested in Transfer Learning (West, Ventura, and Warnick 2007; Torrey and Shavlik 2009), where we will use the knowledge of this work of breast tumor on a different kind of tumor.

References

Benito, Monica, e. a. 2004a. Adjustment of systematic microarray data biases. *Bioinformatics* 20(1):105–114.

Benito, Monica, e. a. 2004b. Adjustment of systematic microarray data biases. *Bioinformatics* 20(1):105–114.

Collins, F. S., and Varmus, H. 2015. A new initiative on precision medicine. *New England Journal of Medicine* 372(9):793–795.

Davis, S., and Meltzer, P. S. 2007. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics* 23(14):1846–1847.

Eisen, M., and de Hoon, M. 1998. Cluster 3.0 manual. *University of Tokyo: Human Genome Center*.

Eisen, Michael B, e. a. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95(25):14863–14868.

Frey, L. J.; Bernstam, E. V.; and Denny, J. C. 2016. Precision medicine informatics.

Galili, T. 2015. Extending 'dendrogram' functionality in r [r package dendextend version 1.7.0].

Gentleman, Robert C, e. a. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* 5(10):R80.

Hu, Zhiyuan, e. a. 2006. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC genomics* 7(1):96.

Ioannidis, John PA, e. a. 2009. Repeatability of published microarray gene expression analyses. *Nature genetics* 41(2):149.

Kassambara, A. 2017. Extract and visualize the results of multivariate data analyses [r package factoextra version 1.0.5].

Kim, Sunghwan, e. a. 2015. Pubchem substance and compound databases. *Nucleic acids research* 44(D1):D1202–D1213.

Liu, B. 2018. Lifelong learning - learn as "humans do".

Nekrutenko, A., and Taylor, J. 2012. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics* 13(9):667.

Perou, Charles M, e. a. 2000. Molecular portraits of human breast tumours. *Nature* 406(6797):747.

Rich, Jason T, e. a. 2010. A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head and Neck Surgery* 143(3):331–336.

Ritchie, Matthew E, e. a. 2015. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research* 43(7):e47–e47.

Sandve, G. K.; Nekrutenko, A.; Taylor, J.; and Hovig, E. 2013. Ten simple rules for reproducible computational research. *PLoS computational biology* 9(10):e1003285.

Schroeder, Markus, e. a. 2011. breastcancernki: Gene expression dataset. *R package version* 1(6).

Sørlie, Therese, e. a. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* 98(19):10869–10874.

Sørlie, Therese, e. a. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences* 100(14):8418–8423.

Sotiriou, Christos, e. a. 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences* 100(18):10393–10398.

Taminau, Jonatan, e. a. 2012. Unlocking the potential of publicly available microarray data using insilicodb and insilicomerging r/bioconductor packages. *BMC bioinformatics* 13(1):335.

Torrey, L., and Shavlik, J. 2009. Transfer learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* 1:242.

Van't Veer, Laura J, e. a. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *nature* 415(6871):530.

Warnes, G. R. 2016. gplots.

West, J.; Ventura, D.; and Warnick, S. 2007. Spring research presentation: A theoretical foundation for inductive transfer. *Brigham Young University, College of Physical and Mathematical Sciences*.