Clustering estrogen receptor-positive breast cancer tumors based on hormonal response type

Project 2 - CS-433 Machine Learning, EPFL
19 December 2019

Loftsson, Valentin Oliver (IN) valentin.loftsson@epfl.ch
Dratva, Lisa (SV) lisa.dratva@epfl.ch
Pleskowicz, Michal (IN) michal.pleskowicz@epfl.ch

***Abstract*—We employ unsupervised machine learning techniques to cluster subtypes of estrogen receptor-positive breast cancer, which is the most common variant worldwide. Clustering is done according to hormone responses obtained from *in vivo* models of patient-derived xenografts. Our results facilitate more targeted treatment of patients, responding to the urgent need for personalized medicine to treat breast cancer.**

Introduction

Breast cancer is the second-most common type of cancer in women worldwide [1]. Breast cancers can be divided into subtypes by discriminating the expression of markers inside the tumor cells, such as hormone receptors, where industry research focuses on the estrogen receptor. Tumors expressing hormone receptors are hormone-sensitive, as their cancer cells need the naturally occurring hormones estrogen or progesterone to grow, and can be targeted with endocrine therapies [2]. Estrogen receptor-positive (ER^+) breast cancers make up over 75% of all breast cancers worldwide [3]. Despite the classification of ER^+ breast cancers, which has allowed to subselect the cohort of patients on which endocrine therapies should be effective, its effectiveness is hampered by the great inter-patient heterogeneity that is observed in terms of treatment response. This highlights the need to identify specific ER^+ subtypes to deliver more targeted treatments. Our work integrates into the field of personalized or precision medicine, where patients receive the custom treatment that is most effective for them [4].

ER^+ tumors may be differentiated based on the hormonal response they present. The tumor can respond to either of the three hormones: estradiol (E2), progesterone (P4) or dihydrotestosterone (DHT).

PDX model

Research for breast cancer can be carried out in different models: Cancer cell lines, genetically engineered mice, and patient-derived xenograft (PDX) mouse models. In the PDX model, mice are injected with tumor cells from a cancer patient. The injected mice sometimes grow a tumor, and in such instances, the tumor is termed a PDX. PDXs model the patient's tumor *in vivo* and allow researchers to investigate new treatment options in a controlled environment [5]. The results obtained from studies in mice can then be used as indicators for outcome in human patients and pave the way for clinical trials.

To induce the response observed in some tumors, engrafted mice can be treated with the above-mentioned hormones. Administration of the hormones gives rise to a tumor that responds to the hormone. The tumors with induced response can then be compared to tumors from control mice that haven't received any hormone stimulation.

Data

The comparison is based on the readouts from an RNA sequencing (RNAseq) machine, which returns the relative expression levels of genes from the cancer cells. Three RNAseq data sets are provided:

- PDX data containing 41 mice samples with 15'525 measured genes (9 E2, 11 P4, 3 DHT, 10 control and 8 samples treated with two hormones)
- Patient data containing 617 tumor patient samples with 20'505 measured genes
- The Cancer Genome Atlas (TCGA) online database [6] of breast cancer data containing 1'212 patient samples

The PDX data may be considered labeled because the response was induced by hormonal stimulation and the experimental parameters are known. The two patient data sets, however, contain no labels. A list of 108 genes differentially expressed with regards to the hormonal response from the PDX experiments is also provided.

Collaboration

The project is proposed and supervised by the UPBRI laboratory at EPFL, which investigates the interplay between hormones and breast cancer. Our lead collaborator is Ph.D. student Fabio De Martino, who also performed the PDX mouse experiments, similar to experiments in [7].

Task

The challenge is the following: Given that ER^+ tumors can show response to either E2, P4, DHT or none, optimize an appropriate algorithm to cluster unknown tumors by hormone response; then, demonstrate the findings on the patient data sets.

The idea is to evaluate the algorithms on the PDX data set, which is labeled. The algorithm scoring best on the PDX set is then used to cluster the patient data, as it is expected to discriminate the hormonal response clusters correctly. That is, in humans, we try to find similar expression patterns to determine if the tumor is driven by a certain hormone, in which case we can group such tumors for more targeted and better treatment.

Code repository

The code for this project is entirely hosted on GitHub. Instructions to reproduce our findings and comments detailing every step are provided. Interactive 3D plots are hosted by chart studio.

Exploratory data analysis

The RNAseq data comes pre-normalized by samples to avoid misinterpretation given by technical biases. The PDX and patient sets are not normalized in the same way. However, the expression ratios are conserved. Experimental data for mice treated with two hormones are excluded from the current analysis since their labels are ambiguous, leaving 33 data samples in the PDX set.

Merging patient data sets

We attempt to merge the two patient data sets, as they exhibit similar mean values and feature distributions. It turns out that the smaller patient data set is entirely contained within the bigger patient data set from the online database, reducing the total number of samples we can investigate to the 1'212 samples from the TCGA set.

Feature selection

If some genes' expression levels change after hormonal stimulation, we can infer that the concerned genes are related to the tumor's hormonal response. We can thus limit our analysis of RNAseq data to the genes correlated with hormonal response, reducing the number of genes from over 20'000 to 108. Those genes are either up- or down-regulated after hormonal treatment. Eight genes vary with two treatments instead of one. Of the 108 genes, only 91 appear in both PDX and patient data sets and are thus of interest to us. The terms "genes" and "features" shall be used interchangeably.

Feature correlation analysis

Pearson correlation analysis is performed between genes in the patient data set [8]. We find that 18 out of 21 gene-pairs having Pearson correlation above 60% correspond to genes that are up- or downregulated for the same hormone (see five highest correlations in Table I). These findings support the hypothesis that knowledge gained from the PDX experiments can be transferred to tumor patients.

	correlation	hormone
(KLK14, KLK12)	0.98	e2
(MYBPC1, ATP1A2)	0.86	dht
(SYNPO2, MYBPC1)	0.82	p4
(PDE2A, GIMAP6)	0.78	p4
(UGT2B28, ALOX15B)	0.77	dht

TABLE I: Top 5 correlated pairs of genes in the patients data differentially expressed upon the same hormonal treatment (view full list).

Feature processing

The distributions of TCGA features are heavy-tailed as can be seen in Figure 1. It appears that the distributions are log-normal and so we log-transform the features. Thereafter, we standardize them to facilitate the cluster analysis (view resulting distribution).

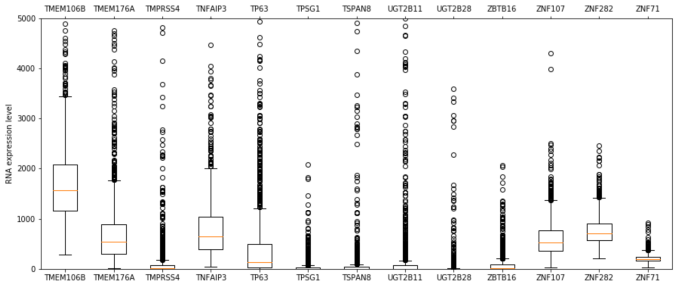

Figure 1: Distribution of 13 sample genes from the patient set, before processing.

Principal component analysis (PCA)

To better understand the labeled PDX data, we analyze it with PCA only considering the 91 genes of interest. Most of the variation (63%) is explained by the first two components, but Figure 2 shows a strong separation by tumor instead of treatment, meaning the variation relates to the original tumor sample (view 3D visualization).

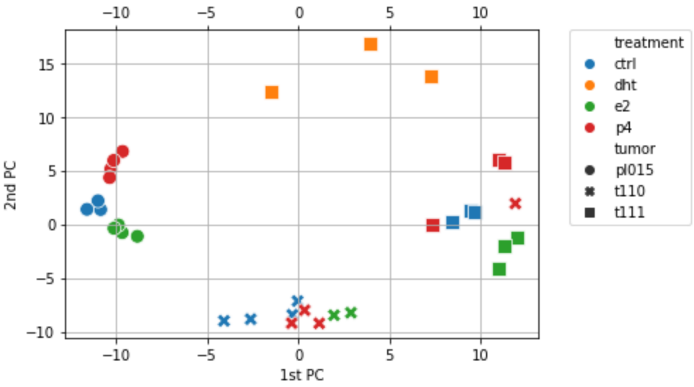

Figure 2: PCA of raw PDX samples. Colors represent the treatments, shapes the original tumor sample. Ctrl denotes mice without any treatment.

Unsupervised learning is likely to return clusters based on the initial tumor tissue, which is not what we want. To address this issue, we standardize the samples coming from the same tumor tissue individually. Figure 3 shows that the bias due to tumor tissue is mostly removed, and the data now clearly clusters by hormonal treatment (i.e. colors cluster together instead of shapes, view 3D visualization).

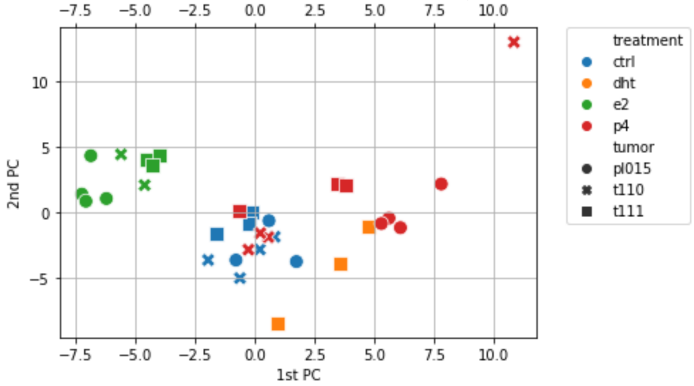

Figure 3: PCA of tumor-standardized PDX samples.

We also perform PCA on the log-transformed and standardized TCGA data set, again on the 91 genes of interest. The resulting interactive 3D plot shows that most data points

gather within a broad range of the first three components. When we examine the explained variance, they explain 33.3% of the variance. We then compare if the composition of genes contributing to the first three principal components of the two data sets is similar, and find that it is very dissimilar.

Clustering approaches

We employ unsupervised learning techniques on both PDX and TCGA data. For the PDX data, the correctness of clusters can be verified using the labels as ground truth. For the TCGA patient set, we can only look at metrics that quantify how well each sample is classified based on cluster cohesion and separation from other clusters, and on physiological expectations.

Adjusted Rand index (ARI)

ARI compares the predicted labels to the actual labels, returning a score between $[-1, 1]$ [9]. We can use it on the PDX data to measure the accuracy of the clusters. It is corrected for chance, meaning that random labeling scores around 0, while perfect labeling scores 1.

Silhouette coefficient

The silhouette coefficient evaluates how similar points are to other points in their assigned cluster, and how dissimilar they are to points outside the cluster. The coefficient lies between $-1 \leq s_i \leq 1$, where 1 denotes a very well-clustered data point and -1 implies a wrongly assigned one [10].

Davies-Bouldin (DB) score

Another metric to evaluate cluster partition is the DB score. It quantifies the similarity of different clusters, such that a lower score means better cluster separation and is thus desirable [11]. The DB score is defined within $\mathbb{R}_+$.

Clustering methods

We calculate the ARI scores for different initial conditions to find the optimal configuration for clustering the PDX data. Using the scores, we then evaluate which clustering algorithm performs best on the PDX data. We compare the following clustering algorithms:

- Agglomerative clustering
- K-means
- Spectral clustering

Our choice of algorithms is explained by the fact that all of them work well on data sets of small size such as PDX. Furthermore, since we physiologically expect only a few clusters (between 3 and 5), we choose algorithms that take as an input the cluster size. With agglomerative clustering, we also include a hierarchical clustering method for comparison.

Results

Method evaluation

The following tables show the ARI score achieved on the PDX data. k denotes the number of clusters, provided as an argument to the functions.

Table II shows that no algorithm performs better than chance when it comes to raw PDX data. This is expected as the algorithms cluster based on the initial tumor (shape from Figure 2), which won't give good results on hormone labeling.

k	ari agglomerative	kmeans	spectral
2	0.034	0.011	-0.045
3	0.020	-0.057	0.106
4	-0.031	0.033	0.059
5	0.019	0.113	0.191
6	0.012	0.180	0.090

TABLE II: ARI scores for raw PDX data

k	ari agglomerative	kmeans	spectral
2	-0.008	0.397	0.034
3	0.026	0.560	0.333
4	0.067	0.484	0.447
5	0.061	0.647	0.130
6	0.110	0.377	0.290

TABLE III: ARI scores for standardized PDX data

After standardization by the initial tumor sample, the ARI scores are significantly better from random cluster assignments, as reported in Table III.

It seems that the control samples pose the biggest problem for clustering. This is explained by Figure 3, where the control samples are not clearly separated from the other samples, but instead show up in between (also true in 3D). If we remove the control samples and run the clustering algorithms again, Table IV shows that a perfect ARI score of 1 can be achieved by K-means (view 3D here). This was possible after trying out different random initial states and then picking the one that resulted in the best ARI score.

k	ari agglomerative	kmeans	spectral
2	-0.024	0.739	0.051
3	0.054	1.000	0.812
4	0.146	0.813	0.397
5	0.804	0.603	0.278
6	0.730	0.602	0.248

TABLE IV: ARI scores for standardized PDX data without control samples and using optimal initial centroids for K-means.

Clustering patient data

From Table IV we identify K-means with $k = 3$ as the algorithm performing best on our data set since the control samples overlap with the treatment samples. We can test two approaches to transfer this knowledge to the patient data set:

- take the cluster centroids resulting in optimal clustering of the PDX samples and apply them
- run the same algorithm

Extracting the cluster centroids is straightforward, as is predicting new data labels based on existing cluster centroids. The result of this approach is visualized in Figure 4, where we plot the results for the first two principal components (view 3D here). Cluster 0 indicates DHT, 1 corresponds to E2 and 2 to P4. A single data point is labeled as DHT at position [5 3.5] while most points cluster as P4 or E2.

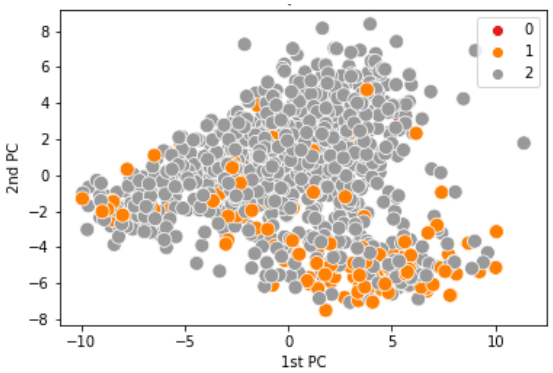

Figure 4: Labeled patients according to centroids from PDX data. 0: DHT, 1: E2, 2: P4.

Now we run the K-means algorithm on the patient data set, with $k = 3$ number of clusters. Results are visualized in Figure 5 (view 3D here).

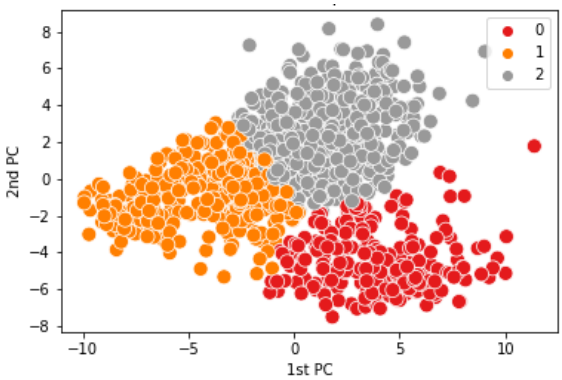

Figure 5: Labeled patients from K-means algorithm using $k = 3$ clusters.

Next, we run the different clustering algorithms on the patient set and compare the silhouette and DB scores, as shown in Table V. Agglomerative clustering shows the best

k	silhouette aggl.	db aggl.	silhouette kmeans	db kmeans	silhouette spectral	db spectral
2	0.316	0.539	0.091	2.643	0.059	6.013
3	0.298	0.537	0.110	2.442	0.023	5.926
4	0.209	0.592	0.086	2.656	0.035	4.864
5	0.192	0.600	0.080	2.842	0.034	4.694
6	0.134	0.635	0.079	2.551	0.007	4.086

TABLE V: Silhouette (higher better) and DB (lower better) scores for patient data.

silhouette and DB scores for all numbers of clusters in the patient data, meaning its clusters are better partitioned than the clusters found with K-means or the spectral method.

Discussion

It is important to note that it is currently not known whether tumor patient data exhibit separation by hormonal response. While it is suspected, other unidentified factors could potentially explain more variance within the tumor population. Thus an interpretation of clusters found through unsupervised learning depends on further advances in cancer research to be validated or refuted.

To analyze Figure 4, it must be understood that variation within the patient data set is much greater than within the PDX data set, such that the first two principal components explain less than 30% of total patient variance. Thus we would not expect the same clear clustering as for the PDX data, where the first two explain double that amount. Figure 5 shows clear cluster boundaries, but the clusters are very close to each other. For K-means ($k = 3$) on the patient data as shown in Figure 5 we report a silhouette score of 0.110 and DB score of 2.442 in Table V. The scores indicate that the clusters overlap and that many samples are likely to be assigned to a wrong cluster. Interestingly, according to Table V agglomerative clustering performs better on the patient data than K-means, with a silhouette score of 0.298 and DB of 0.537 for $k = 3$.

Limitations

Our analysis is limited by the small PDX sample size ($n = 33$) and the apparent differences in data obtained from mouse experiments versus actual human patients. The PDX data shows strong biases not present in the human data, such as bias due to the initial tumor used to grow the mouse tumor. Mouse and human metabolism also differ significantly, which directly impacts gene expression. It is not clear if it makes sense to apply the best performing method found on the PDX samples directly to the patients, even though samples are standardized to remove the tumor bias. The naturally occurring heterogeneity within the patient set further complicates the clustering process. We can hardly draw statistically significant conclusions from clustering such a small data set, especially considering that the DHT hormonal treatment only has three samples available. More samples are needed to draw more accurate conclusions.

In the future, it might be possible to analyze the hormonal composition of the patient's blood and tumor tissue, which means that labels could become available to significantly improve the prediction process [12].

Conclusion

We present a pipeline for processing and analyzing the genetic expression of cancer patients and experimental PDX samples, visualizing their components and discovering latent clusters. We attempted to cluster 1'212 tumor patient samples using 91 relevant features. Three clustering algorithms were applied and evaluated. K-means with $k = 3$ performed best on the labeled PDX data considering the ARI score. This model was then applied to the patient data. Two approaches were followed: (1) directly applying optimal centroids and computing cluster assignments, and (2) running the same algorithm again. While we cannot know if the clustering of patients is correct, we evaluated our results with industry-standard metrics and compared them to physiologically expected results.

REFERENCES

[1] Waks, A. G., Winer, E. P. (2019). Breast cancer treatment: A Review. Jama, 321(3), 288-300.

[2] Hormone Therapy for Breast Cancer Fact Sheet. Retrieved December 1, 2019, from https://www.cancer.gov/types/breast/breast-hormone-therapy-fact-sheet.

[3] Nasrazadani, A., Thomas, R. A., Oesterreich, S., Lee, A. V. (2018). Precision medicine in hormone receptor-positive breast cancer. Frontiers in oncology, 8, 144.

[4] Ginsburg, G. S., Phillips, K. A. (2018). Precision medicine: from science to value. Health Affairs, 37(5), 694-701.

[5] Whittle, J. R., Lewis, M. T., Lindeman, G. J., Visvader, J. E. (2015). Patient-derived xenograft models of breast cancer and their predictive power. Breast cancer research, 17(1), 17.

[6] Broad GDAC Firehose. Retrieved December 2, 2019, from https://gdac.broadinstitute.org/. Browse Breast Invasive Carcinoma data.

[7] Sflomos, G., Dormoy, V., Metsalu, T., Jeitziner, R., Battista, L., Scabia, V., ... Vilo, J. (2016). A preclinical model for ERα-positive breast cancer points to the epithelial microenvironment as determinant of luminal phenotype and hormone response. Cancer cell, 29(3), 407-422.

[8] Benesty, J., Chen, J., Huang, Y., Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1-4). Springer, Berlin, Heidelberg.

[9] Hubert, L., Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1), 193-218.

[10] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[11] Davies, D. L., Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.

[12] Carroll, J. S., Hickey, T. E., Tarulli, G. A., Williams, M., Tilley, W. D. (2017). Deciphering the divergent roles of progestogens in breast cancer. Nature Reviews Cancer, 17(1), 54.