Genes clustering selection to survival prediction in breast cancer patients

Seleção de agrupamento de genes para predição de sobrevida em pacientes com câncer de mama

Selección de agrupamiento de genes para la predicción de la supervivencia en pacientes con cáncer de mama

Khennedy Bacule dos Santos[1], Israel Tojal da Silva.[2] Mariana Cúri.[1]

1 Instituto de Ciências Matemáticas e de Computação (ICMC), USP - São Carlos, São Paulo (SP), Brasil.
2 A.C.Camargo Cancer Center, São Paulo (SP), Brasil

Autor correspondente: Khennedy Bacule dos Santos
E-mail: khennedy@usp.br

Abstract

The risk stratification based on molecular data for predicting cancer progression or outcome is an important undertaking for supporting clinical decision making in oncology. In this work, we use Cox model and K-means to define a prognostic gene expression-based signature. Our approach reaches a better C-index (0.8341) and outperforms the Cox model by using clinical data alone (0.6348). Overall, this shows that the genetic signature found is related to the evolution of the patient's clinical condition, detecting molecular features related to prognosis in breast cancer.

Keywords: Machine learning; Breast cancer; Genes expression

Resumo

A estratificação de risco com base em dados moleculares para prever a progressão ou o resultado do câncer é um empreendimento importante para apoiar a tomada de decisão clínica em oncologia. Neste trabalho, usamos o modelo de Cox e K-means para definir uma assinatura baseada na expressão gênica prognóstica. Nossa abordagem atingiu um

índice C-index (0,8341) e supera o modelo de Cox usando apenas dados clínicos (0,6348). No geral, isso mostra que a assinatura genética encontrada está relacionada à evolução do quadro clínico da paciente, detectando características moleculares relacionadas ao prognóstico no câncer de mama.
Descritores: Aprendizado de máquina; Câncer de mama; Expressão gênica

Resumen

La estratificación del riesgo basada en datos moleculares para predecir la progresión o el resultado del cáncer es una tarea importante para respaldar la toma de decisiones clínicas en oncología. En este trabajo, usamos el modelo de Cox y K-means para definir una firma basada en la expresión génica de pronóstico. Nuestro enfoque logró un C-index (0,8341) y supera al modelo de Cox utilizando solo datos clínicos (0,6348). En general, esto demuestra que la firma genética encontrada está relacionada con la evolución del estado clínico de la paciente, detectando características moleculares relacionadas con el pronóstico en cáncer de mama.
Descriptores: Aprendizaje Automático; Cáncer de mama; Expresión génica

Introduction

Cancer is a disease that affects the whole world population, affecting about 19.3 million people in the year 2020 [1]. In Brazil, the National Cancer Institute (INCA) shows that in 2020 more than 600,000 Brazilians suffered from cancer. The breast and prostate cancers are the most common incidents for women and men, respectively. The number of deaths of females with breast cancer in 2019 was 18,068, which represents 16.4% of female deaths. Of note, a research based on data from subjects gathered from the year 2000 to 2015 reported that the number of breast cancer cases tends to increase in women [2]. The study shows that by 2030 there will be an increase from 13.3% to 22.9% in relation to 2015. The projection of the number of deaths for the year 2030 is up to 40.9% than the year 2015.

All cancers arise because of changes that have occurred in the DNA sequence of cell genome [3]. Understanding the relationships between the genomic aspect of cancer and the clinical features of the disease has therefore become a priority in cancer research.

In recent years, advances in state-of-the-art sequencing technologies Next Generation Sequencing (NGS) have allowed us to unveil the main molecular aspects of a wide range of neoplasms [4].

Breast cancer is a complex disease and presents itself in a heterogeneous way in the various histological and molecular subtypes. For example, in the histological aspect, the classification of breast carcinoma comprises the following subgroups: i) *in situ*, characterized by the proliferation of neoplastic cells inside the mammary ducts (DCIS) or inside the lobules (LCIS), and ii) invasive, where the cells tumors have already surpassed the cellular barrier of the basal layer. In the molecular context, tumors are classified into four subtypes: Luminal A (tumors that have positive estrogen and progesterone receptors), Luminal B (have positive estrogen and/or progesterone receptors), HER2 (do not express hormone receptors, but have the expression of the HER2 protein) and Triple Negative (they have neither hormone expression nor the HER2 protein, being negative, therefore, for estrogen, progesterone and HER2 - **TNBC**, *Triple-Negative Breast Cancer*).

Due to the heterogeneity of cancer, the application of statistical methods becomes viable to detect different aspects of the disease. One way of analyzing it is to take advantage of survival methods, such as the Cox model, inferring the risk of death from the data collected as a function of the patient's characteristics. The Cox model is widely used in health, with applications in several types of cancer. Most studies seek to identify the best effect of procedures applied to different groups of patients [5,6,7,8]. Methods such as the Cox model bring a strong interpretation of the data, which is why they are widely used in the health field. However, most of these methods have a limitation regarding the number of characteristics in the analysis, which makes them unfeasible to be applied in some situations. For example, in genetic data, since each person has thousands of genes. Thus, different ways of dealing with this large volume of information, and still having the benefit of interpretation, are quite relevant and remain to be further vetted. The application of a penalty in the Cox model allows the feature selection in situations with a huge number of features [9,10,11,12].

Other approaches are also found, such as the union of methods, capable of dealing with large volumes of information [13] and, in some cases, with the ability to obtain non-

linear relationships [2,14,15]. In general, these approaches bring better results compared to the classical use of the Cox model. But most of them bring a loss of model interpretability [2,14,15].

This research proposes a method to predict survival with high-dimensional data from patients with early-stage breast cancer (stages I and II), using clinical and genetic protein data. The research is based on the fundamentals of survival analysis and on a way to reduce the dimensionality of genetic information, using a clustering method. In this way, we will have the ability to insert the genetic material into the survival model, and still could interpret the model.

Methods

Dataset

The clinical and molecular information was downloaded from the *The Cancer Genome Atlas* (*TCGA*) project and available at https://portal.gdc.cancer.gov [16]. In this project we used two datasets. The first contains several clinical information of patients. The second is the genetic information of these patients. The base used is called **TCGA-BRCA,** which are data from 1219 patients with breast cancer. Clinical data have a wide variety of information, with 154 clinical features.

In the gene expression database, 60483 genes are represented. With this large volume of genes, we can separate them into two large groups by their functions. 19595 genes are associated in the protein coding process; these genes have a fundamental role since proteins are present in various biological processes of the body, such as the creation of antibodies. The remaining genes, those that are not protein-coding, are related to internal processes in the cell.

Cox proportional hazards model

In the survival analysis, several models were proposed to identify factors that influence survival. Some of them, assuming known probability distributions for the failure time. In a more general context, without assuming any specific distribution for the time to failure, there is the Cox model, used in this work to predict the risk of death of patients with breast cancer.

The Cox model is a regression that relates the failure rate between two groups of different individuals as a constant. For example, suppose we have two study groups: one with stage 1 and one with stage 2 cancer. The failure rate function for group 1 is $\lambda_0(t)$ and for group 2, $\lambda_1(t)$. We can define that

$$\frac{\lambda_0(t)}{\lambda_1(t)} = K \quad (1)$$

where *K* is the ratio between the failure rates of group 1 and 2, assumed to be constant over time *t*. Generalizing to more than two groups, defined from *n* covariates, the Cox model defines risk as follows:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp\left(\beta_1 x_1 + ... + \beta_n x_n\right), \quad (2)$$

for each *x* that represents some characteristic given by the study we have a parameter β pegged. The Cox method is called semi-parametric because it has two components: one non-parametric and one parametric. The part dependent on the covariates, characteristic of the individuals, is the parametric part, while the factor $\lambda_0(t)$ is the non-parametric part.

The name proportional hazards given to this model comes from the fact that the failure rate between two individuals is constant. Earlier, we saw that the ratio of risk of death between two groups (in stages 1 and 2, respectively) is *K*. In this more general case of *n* covariates, we have that the individuals *i* and *j* have the following failure rate relationship:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp\{\boldsymbol{x}_i'\beta\}}{\lambda_0(t) \exp\{\boldsymbol{x}_j'\beta\}} = \exp\{\boldsymbol{x}_i'\beta - \boldsymbol{x}_j'\beta\}. \quad (3)$$

If this ratio is greater (or less) than 1, then the individual *i* can be said to have a greater (or lesser) risk of death than the individual *j*, in a constant time. With the Cox model, we can obtain a lot of information about the influence of covariates on survival. The significance of each parameter β in the model tells about the importance of the corresponding covariate in survival (or failure function). In this regard, the estimation of the β's must be done and one could consider using the maximum likelihood method, classically. However, to circumvent the existence of the non-parametric component $\lambda_0(t)$, the partial maximum likelihood method is adopted.

K-Means gene selection

Figure 1 represents the methodology proposed in this research for the prediction of survival, where gene selection is based on the k-means clustering method.

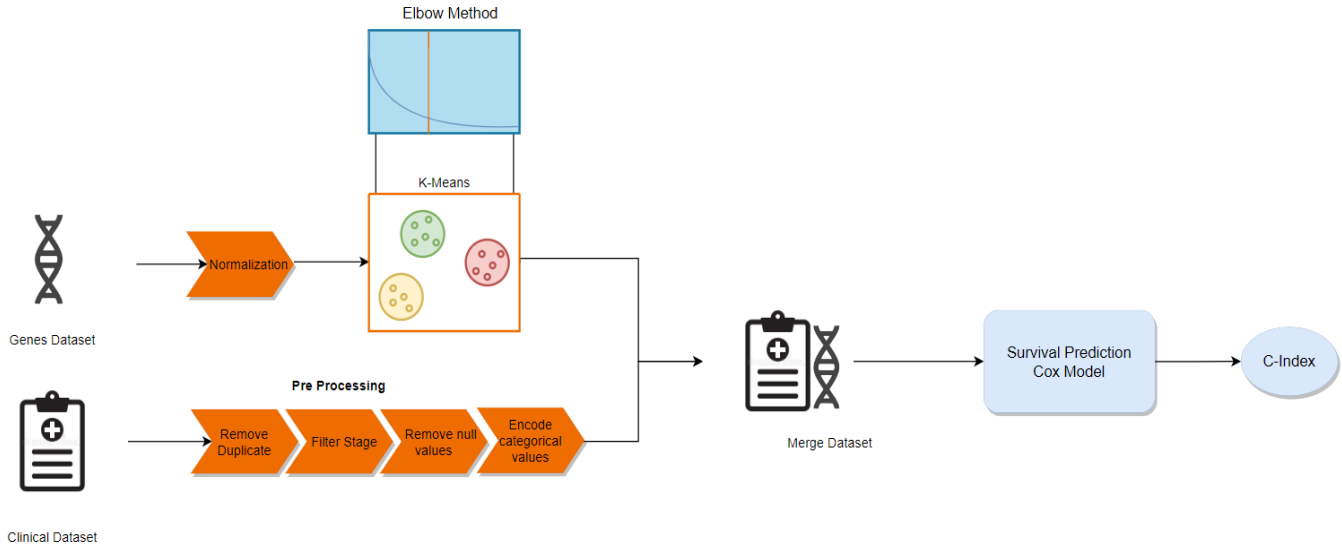

Figure 1 - Approach to survival prediction with clinical and genes data, using clustering genes selection.

The high dimensionality of genetic data makes methods such as the Cox model unfeasible to be applied. As demonstrated, this research performs methods for the selection of genes that will aggregate clinical data, which will be the input for the Cox regression that will enable the prediction of patient survival. The k-means method represents your data in a plane $\mathbb{R}^p$ where *p* is the dimension of the data we are working on, in this case after reducing the number of genes, selecting only those that encode proteins we have *p = 19595*.

Patients contribute with their gene expression to the possible grouping of genes that are related in their expression level. In this way we have $i = (1, 2, ..., 1092)$ patients, where the patient p_i has 19595 protein expression genes. We can consider that the input to the k-means model consists of the matrix X_{gp} where *g* are the genes and *p* the patients. A gene is represented by a point u_g in the plane $\mathbb{R}^p$, formed by the vector $u_g = (p_{11}, p_{21}, ..., p_{i1})$. In this way, the grouping consists of identifying genes that have similarity in their expression. Figure 2 exemplifies the approach, in this case we see three

groups formed by genes, and the contribution to the identification of genes are given by patients.

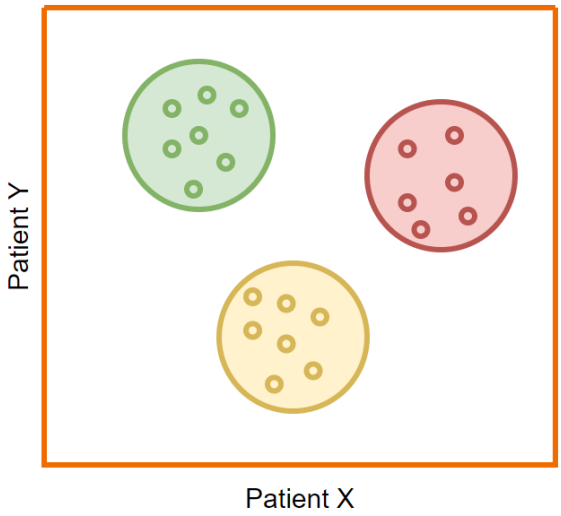

Figure 2 - Genes clustering using K-means

Euclidean distance is used to form the groups in the cluster analysis. One of the tasks within the k-means algorithm is to define the value of *k* that represents the number of groups, the method used to estimate this value is the *Elbow*. The *Elbow* method estimates the value of *k* by varying the value of *k*, for each execution it checks the variance created by the defined groups.

The choice of *k* is then given when the insertion of new groups is still relevant and does not specify the data. By estimating the number of groups, we will obtain a grouping of genes and in this way the gene selection process begins. K-means when starting its execution creates *k* points μ, called centroids. At the end of the execution, the centroid will be at the central point of its respective group, so that the point μ is the midpoint of the group. Thus, we will consider that the gene closest to the centroid will be selected for the prediction of survival. We can see in Figure 3 that the genes selected in groups are the closest to the centroid.

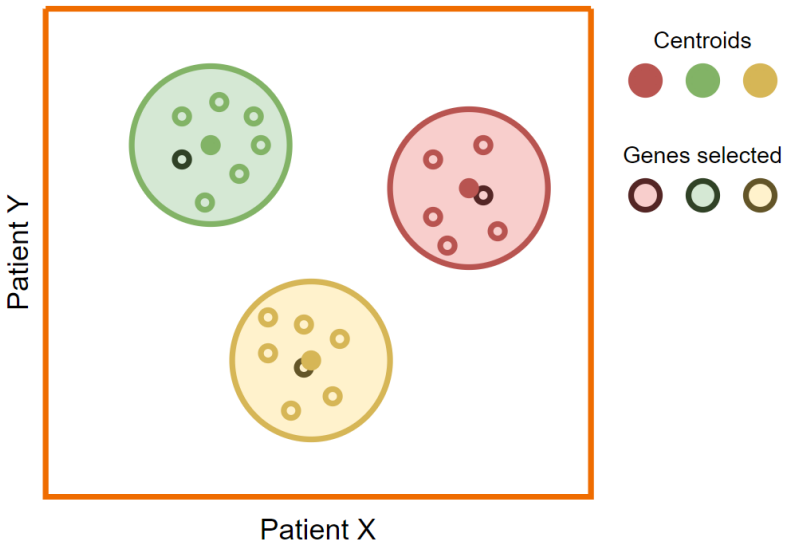

Figure 3 - Genes selection approach using k-means

With this approach, when choosing the gene closest to the centroid, we were able to remove some possible outliers in the solution, since the genes chosen are those that are closest to the center of their group.

The next step is the unification of the selected genes with the clinical characteristics for the prediction of survival. In this case, to compare the proposed methodology with the Cox model using only clinical data, the C-index metric will be used.

Concordance index

The C-index metric [17] implements a logic that if the patient has died at time T, his risk probability must be greater than the case of a censorship if it is censored at a time higher than *T*. The C-index then evaluates the data by pairs and those pairs that satisfy the mentioned logic are computed as a concordant pair and those that do not are non-concordant pairs. In this way we can give the following logic equation:

$$C_{index} = \frac{\sum_{i,j} T_j < T_i \cdot \eta_j > \eta_i \cdot \delta_j}{\sum_{i,j} T_j < T_i \cdot \delta_j} \quad (4)$$

where η represents the risk value and δ denotes the occurrence of the event or censoring, with value 1 for the event and 0 for censoring.

Preprocessing

The number of clinical records is 1219, but there are cases of repetition of patients, only the last records of these patients were selected. Newer records were excluded for patients who had more than one occurrence in the database. After this selection, the number of examples is 1098, guaranteeing the uniqueness of patient registration.

Five characteristics were considered, where the characteristics **days_to_last_follow_up** and **vital_status.clinical** are mandatory information for the Cox model and are not considered covariates.

The covariates considered in the Cox model, the features **Subtype** and **pathologic_stage** are variables that are not in a numerical domain, these two variables must be treated so that their insertion is possible in the Cox model. The form used for this treatment was the one-hot-encoding technique. The feature **age** is not necessary for any treatment.

The gene expression was normalized using min-max scaling, a way to represent the data in the range between 0 and 1. The database is separated into two groups, 70% is used for training the Cox model and the remaining 30% is used for testing. The information with the occurrence of the event was equally divided between the two bases.

Results and discussion

The Cox method considering only the clinical data estimated 6 values of β's to predict the risk of death of cancer patients. The interpretation of the values of the β's can bring relevant information (Figure 4). For example, in the execution of the method with only clinical data it is shown that the **Her2** subtype has a higher risk than the other subtypes, including the subtype **Basal**, the reference in analysis. Thus, patients with this subtype have about 2 times more risk of death than the others. Our results reinforce previous findings which HER2 overexpression is a marker of poor prognosis in breast cancer [18].

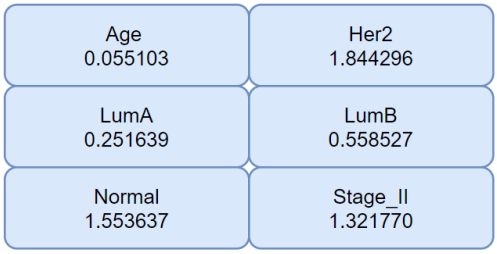

Figure 4 - Cox model result with clinical data only. (Feature, β estimate)

The execution of the Cox model with the clinical data resulted in a C-index of 0.6348. This value shows that using only clinical data, in this case the subtype, stage and age, did not deliver a very assertive model, since in the C-index values close to 0.5 are matched to a random approach, while models with value 1.0 hit all the pairs defined in the metric.

In the proposed methodology we will have the execution of the *Elbow* method, so that it is possible to have an estimate of the number of groups, and later of the *k-means* for the selection of genes. The *Elbow* method returned an estimate of 24 groups, a reduction from 15959 to 24 genes, representing a reduction of more than 99%. We can see in Figure 5 the result given by the *Elbow* method, the execution of which ranged from 2 to 150 groups. Anyway, for the execution of the model we will reach other values of *k* so that it is possible to determine if the *Elbow* method in this application is effective.

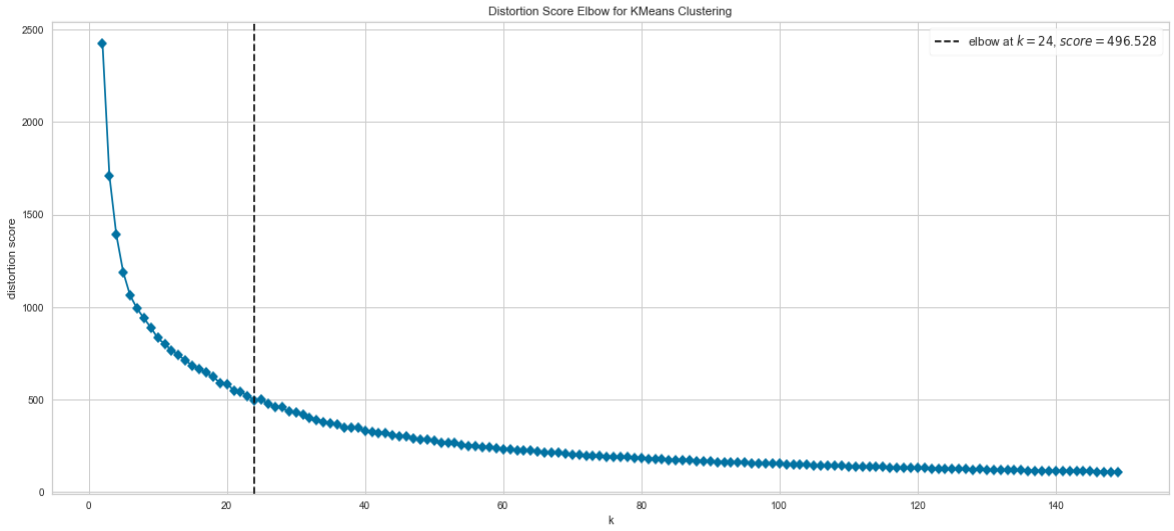

Figure 5 - Elbow results in gene clustering.

Running *k-means* with different *k´*s, according to the *Elbow* result we estimate an interval between the values 15 to 100. In this way, we will have from 15 to 100 genes selected. After grouping and unifying the genes that are closest to the centroids of each group, we can see that in Figure 6 the results refer to the values of C-index. For each execution we had a number of selected genes and with the clinical characteristics, these values were obtained in the test set.

The best C-index given by the Cox model with the *k-means* in the test set was 0.8341, a higher value than the Cox model using only clinical data. The value of *k* in this result was 25, showing that beyond the range the *Elbow* method estimated the value of *k* effectively for this application.

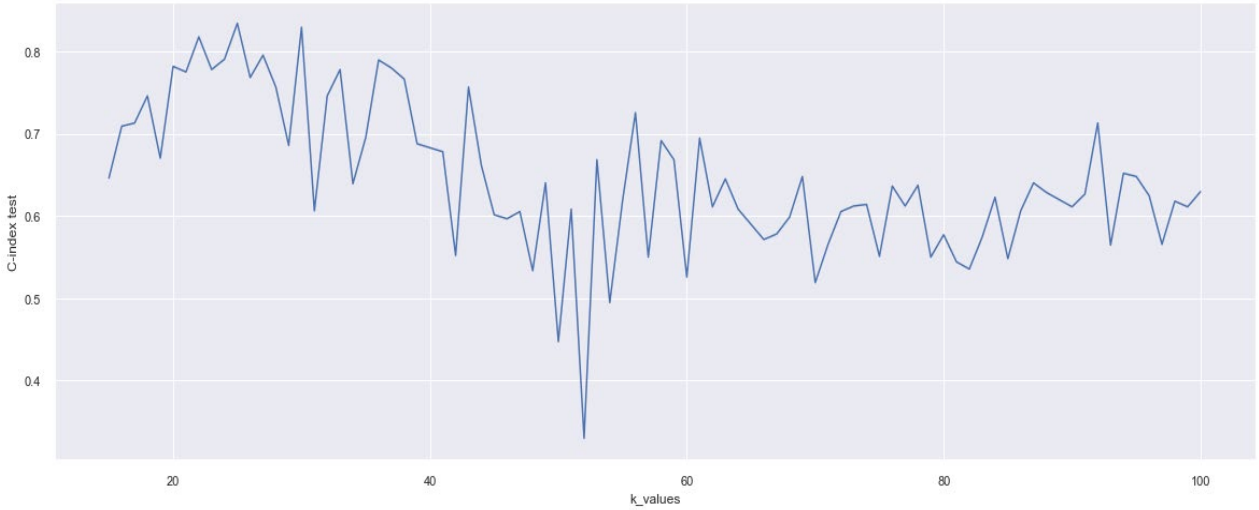

Figure 6 - Cox model result test data with genes selected by K-means. C-index in Y axis and K value of K-means in X axis.

In Figure 7 we see the information of the best Cox model, using 25 groups. In addition, we can see the values of β associated with the features. For the prediction of survival, the gene **C2orf42** is essential. Its expression denotes a significant likelihood of death for the individual. This gene is the one that showed the highest risk in the study, with a β value above 7. The other two genes with highest values, beta-actin (**ACTB**) and (**TARDBP**) encoded by **TDP-43** genes, the genes already associated with invasiveness and metastasis of several cancers, including breast [19,20].

Age 0.085636	Her2 2.039186	LumA 2.263498	LumB 2.648414	Normal 3.551025	Stage_II 1.727527
C2orf42 7.395532	TARDBP 0.502694	SPARC 0.007653	MT-CO1 0.000720	MT-CO2 0.006154	MT-CO3 0.002915
RPS11 0.006733	ACTB 0.012016	PIP 0.003219	TMSB10 0.002437	CPB1 0.003342	FTL 0.000653
MT-ND5 0.003453	MT-ND4 0.001699	MT-ATP6 -0.011350	ACTG1 -0.009520	SCGB1D2 -0.001598	MUCL1 -0.000491
TFF1 -0.011001	MT-ND2 -0.000596	HLA-B -0.015582	RPL10 -0.051796	HNRNPC -0.388986	OST4 -0.142086
		MT-ND3 -0.004236			

Figure 7 - Result of the Cox model with the genes selected by K-means, with 25 groups. In red features that bring a high risk to survival and in green the healing effect. (Feature, β estimate).

In contrast, the gene **OST4** brings a healing effect to patients. Therefore, patients with high expression of this gene have an improvement effect, which makes sense, since this gene has a fundamental role in quality control and its deficiency develops a greater number of disorders [21,22].

In addition to C-Index, we used the log-likelihood test, the value was 56.63 with a P-value 0.003, representing that the full model improves the prediction.

Conclusion

The proposal shows that inserting the genes improves the Cox model, with an improvement of almost 0.2 over using only clinical data. This value for the C-index metric demonstrates that the Cox model with the genes selected by k-means is considered good. Furthermore, the Cox method delivers its β value, making it possible to interpret the model.

Taken together, our results demonstrate that proposed feature selection was able to identify groups of molecular features related to prognosis in breast cancer. Pertinent factors of study are the metrics in the *k-means* algorithm. In this work, we used the Euclidean metric, but others could be considered. We will consider in future works different distance metrics, with the possibility of bringing metrics linked to the genetic map for the selection analysis. The computational framework and data used in this paper are freely available and described at https://github.com/khennedy/genes-clustering-kmeans-cox.

References

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 71(3), 209–249 (2021).

2. Lei, S., Zheng, R., Zhang, S., Chen, R., Wang, S., Sun, K., Zeng, H., Wei, W., He, J.: Breast cancer incidence and mortality in women in china: temporal trends and projections to 2030. Cancer biology & medicine 18(3), 900–909 (2021).

3. Vogelstein, B., Kinzler, K.W.: Cancer genes and the pathways they control. Nat Med 10(8), 789–799 (Aug 2004)

4. Mardis, E.R.: The Impact of Next-Generation Sequencing on Cancer Genomics: From Discovery to Clinic. Cold Spring Harb Perspect Med 9(9) (09 2019).

5. Abadi, A., Yavari, P., Dehghani-Arani, M., Alavi-Majd, H., Ghasemi, E., Amanpour, F., Bajdik, C.: Cox models survival analysis based on breast cancer treatments. Iranian journal of cancer prevention 7(3), 124 (2014)

6. Bellera, C.A., MacGrogan, G., Debled, M., de Lara, C.T., Brouste, V., Mathoulin-P´elissier, S.: Variables with time-varying effects and the cox model: Some statistical concepts illustrated with a prognostic factor study in breast cancer. BMC Medical Research Methodology 10(1) (Mar 2010).

7. Chen, Y., Zeng, W., Zhu, D.: Cox regression analysis on the survival rate of breast cancer patients. In: Yin, H.M., Chen, K., Me˘strovi´c, R., Oliveira, T.A., Lin, N. (eds.) International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021). vol. 12163, pp. 195 – 203. International Society for Optics and Photonics, SPIE (2022).

8. Husain, H., Thamrin, S.A., Tahir, S., Mukhlisin, A., Apriani, M.M.: The application of extended cox proportional hazard method for estimating survival time of breast cancer. Journal of Physics: Conference Series 979, 012087 (mar 2018).

9. Jiang, Q.: Cancer Classification and Gene Selection with Machine Learning Method, p. 122–127. Association for Computing Machinery, New York, NY, USA (2020)

10. Wang, W., Liu, W.: Integration of gene interaction information into a reweighted Lasso-Cox model for accurate survival prediction. Bioinformatics 36(22-23), 5405–5414 (12/2020).

11. Xie, G., Dong, C., Kong, Y., Zhong, J.F., Li, M., Wang, K.: Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features. Genes 10(3) (2019).

12. Zeng, D., Zhou, R., Yu, Y., Luo, Y., Zhang, J., Sun, H., Bin, J., Liao, Y., Rao, J., Zhang, Y., Liao, W.: Gene expression profiles for a prognostic im-
Genes clustering selection to survival prediction in breast cancer patients 13
munoscore in gastric cancer. British Journal of Surgery 105(10), 1338–1348 (04 2018).

13. De Bin, R.: Boosting in cox regression: A comparison between the likelihoodbased and the model-based approaches with focus on the r-packages coxboost and mboost. Comput. Stat. 31(2), 513–531 (jun 2016).

14. Ching, T., Zhu, X., Garmire, L.X.: Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. PLOS Computational Biology 14(4), 1–18 (04 2018).

15. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. The Annals of Applied Statistics 2(3), 841 – 860 (2008).

16. Network, T.C.G.A.: Comprehensive molecular portraits of human breast tumours. Nature 490(7418), 61–70 (Oct 2012)

17. Harrell, Frank E., J., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the Yield of Medical Tests. JAMA 247(18), 2543–2546 (05 1982).

18. M´enard, S., Fortis, S., Castiglioni, F., Agresti, R., Balsari, A.: HER2 as a prognostic factor in breast cancer. Oncology 61 Suppl 2, 67–72 (2001)

19. Guo, C., Liu, S., Wang, J., Sun, M.Z., Greenaway, F.T.: Actb in cancer. Clinica chimica acta 417, 39–44 (2013)

20. Ke H, Zhao L, Zhang H, et al. Loss of TDP43 inhibits progression of triple-negative breast cancer in coordination with SRSF3. *Proc Natl Acad Sci U S A*. 2018;115(15):E3426-E3435. doi:10.1073/pnas.1714573115

21. Dumax-Vorzet, A., Roboti, P., High, S.: Ost4 is a subunit of the mammalian oligosaccharyltransferase required for efficient n-glycosylation. Journal of cell science 126(12), 2595–2606 (2013)

22. Harada, Y., Ohkawa, Y., Kizuka, Y., Taniguchi, N.: Oligosaccharyltransferase: A gatekeeper of health and tumor progression. International journal of molecular sciences 20(23), 6074 (2019)