Hepatocellular Carcinoma tumor stage classification and gene selection using machine learning models

Martin Palazzo[1,2,3], Pierre Beauseroy[1], and Patricio Yankilevich[2]

[1] Institut Charles Delaunay, UMR CNRS 6281 ICD/ROSAS/LM2S, Universite de Technologie de Troyes, Troyes, France
{martin.palazzo,pierre.beauseroy}@utt.fr
[2] Instituto de Investigacion en Biomedicina de Buenos Aires (IBioBA), CONICET - Partner Institute of the Max Planck Society, Buenos Aires, Argentina
{mpalazzo,pyankilevich}@ibioba-mpsp-conicet.gov.ar
[3] Universidad Tecnologica Nacional Facultad Regional Buenos Aires
mpalazzo@frba.utn.edu.ar

Abstract. Cancer researchers are facing the opportunity to analyze and learn from big quantities of omic profiles of tumor samples. Different omic data is now available in several databases and the bioinformatics data analysis and interpretation are current bottlenecks. In this study somatic mutations and gene expression data from Hepatocellular carcinoma tumor samples are used to discriminate by Kernel Learning between tumor subtypes and early and late stages. This classification will allow medical doctors to establish an appropriate treatment according to the tumor stage. By building kernel machines we could discriminate both classes with an acceptable classification accuracy. Feature selection have been implemented to select the key genes which differential expression improves the separability between the samples of early and late stages.

Keywords: Feature Selection · Kernel Learning · Cancer Genomics

1 Introduction

The intersection of Bioinformatics and Machine Learning is an important field in Biomedicine since the speed to generate biological data has been increasing and the cost of obtaining it decreases exponentially [1]. This evolution opens a big opportunity for Cancer biomedical developments. Nowadays it is possible to obtain multi-omics data [2] and different layers of biological information from a tumor sample of a patient and learn from cancer data. Multi-omic data is composed by Gene Expression, Somatic Mutations [3], Copy Number Variation, Metabolomics [4] or Proteomics data among others. This work focus on learning models from Hepatocellular carcinoma (HCC) multi-omic data. HCC is the most common liver cancer type. It is linked to viral hepatitis infection and exposure to toxins such as alcohol, both are subtypes of HCC tumors. In this study two omics are used to classify different subtypes and stages of HCC. First Simple

Somatic Mutations data or Genomics are used to discriminate between tumor subtypes [5] of HCC samples. Then in a second step the Gene expression data or Transcriptomics is used to discriminate between early and late tumor stage [6] of the virus associated tumor subtype.

1.1 Related work

Similar studies have used multi-omic data to characterize along different stages of Hepatocellular carcinoma associated to alcohol subtype using network analysis [7]. Other studies have done Early versus Late stage classification using microarray gene expression data [8]. In this study we propose to work with gene expression and simple somatic mutations data from Next Generation Sequencing (NGS) technologies in a multi-omics data analysis approach. In this section multi-omic data sources and machine learning approaches for tumor classification are introduced.

1.2 Analysis description

Machine Learning and Pattern Recognition techniques allow researchers to explore big volumes of biological data and find hidden patterns. Since Biological data includes tens of thousands of variables like genomic variants, selecting features to reducing dimensionality is a key task. Feature selection enables to improve model training and gives insight about the key genomic variants related to a tumor. In this work we perform a two step data analysis pipeline to select features and classfy tumor subtypes and stages.

The first step consists of the accurate discrimination of the HCC Virus associated tumor sub-type using a binary classification method, to discriminate it from the HCC Alcohol associated tumor subtype. Both tumor subtypes [9] belong to the same primary site Liver but the treatment varies due to the cause of the tumor. The second step of the data analysis pipeline consists in using another omic data layer: gene expression. Here the analysis of HCC Virus associated is with patients labeled by tumor stages 1, 2, 3 and 4. To simplify the problem, stages 1 and 2 are categorized as Early Stage and stages 3 and 4 are categorized as Late Stages.

In both steps of our analysis the HCC is characterized by Protein Coding Genes features (either mutation and expression). The Human Genome has approximately 20.000 genes. On the other hand the quantity of samples obtained from donors are less than 600. The sample to feature ratio is close to 0,03. This situation is known as The Curse of Dimensionality [10]. For this reason we choose to determine the subsets of genes that maximizes linear classification accuracy based on Lasso. The goal first is to obtain a reduced gene panel to help the detection of the Virus associated HCC tumor subtypes using Gene Somatic Mutations. Then another reduced gene panel is obtained to help the discrimination of the early stage in Virus Associated tumors using gene expression data.

In addition, at each step of the data analysis pipeline a similarity matrix between tumor samples is computed by a Kernel Gram Matrix. Then Kernel Target Alignment (KTA) [11] score is calculated for each Kernel, in order to understand how well the kernel separate the two classes. Our aim is to understand how the KTA evolves before and after selecting gene features and how it can be useful in the classification process. The KTA increases with the contrast between the two classes, thus a larger KTA means a higher inter-class distance and it enables to learn a better classifier. The kernel function type used is Gaussian and it is tuned by considering the improvement of the KTA. The tuned kernel is used to train a support vector classifier between the classes [12].
This work poses two objectives. The first one aims to generate gene signature for each omic by selecting the genes that improves the classification performance between HCC tumor subtypes and tumor stages. The second objective is to understand the impact of tuning the kernel function used for support vector classification considering the kernel target alignment, which means the inter-class separability.

2 Data analysis pipeline

This study consists in a data analysis pipeline with two main steps: first classification of HCC subtypes and second a HCC Virus associated tumor stages classification. Both steps are detailed in the next subsections.

2.1 Hepatocellular Carcinoma sub-type classification

The first step consists in the binary classification between virus and alcohol associated tumor sub-types from the correspondent LIRI-JP and LICA-FR studies (see Datasets section). In this first step of the pipeline the number of Simple Somatic Mutations per gene are used as features. Previous studies have discriminated between different tumor types [15] using somatic mutations although a greater challenge remains to classify between tumor sub-types within the same primary site like breast cancer sub-type classification [16]. In this pipeline step, the dataset is characterized by a matrix of 510 samples and 19990 somatic mutated genes as features. Kernel Target Alignment (KTA) is computed for a tuned Gaussian kernel. Then Support Vector Classification with the tuned kernel is trained with the two tumor sub-types classes considering the full feature set. After that the Least Absolute Shrinkage and Selection Operator (LASSO) method [17] using the L1 norm is used for feature selection. A considerably reduced gene set is obtained and then again KTA is computed and SVC is performed aiming a better classification result.

2.2 Tumor stage classification of Hepatocellular Carcinoma Virus subtype

The second step consists in the early and late stage classification of the HCC Virus subtype samples. The data is available from the LIRI-JP study within the ICGC data portal. In this step Normalized Gene Expression per gene is used as features. The dataset is characterized by a matrix of 232 samples and 22913 protein expressed genes and other transcript locations. Early stage is obtained by grouping stage I and II. Late stage is obtained by grouping stage III and IV. Early stage has a total of 142 samples and late stage a total of 90 samples. A first approach consists in a gaussian kernel tuned to improve the Kernel Target Alignment. Then a Support Vector classifier using the tuned kernel is implemented for binary classification using the full feature subset. Since the dimension of the problem is too high for the number of samples Lasso method is used to select a reduced set of expressed genes. A second approach consists in a Gaussian kernel built with the new gene subset selected by Lasso. Using the new kernel KTA and SVC is computed and compared with the results obtained on the original dataset. This step has a a critical impact on clinical diagnosis since the early detection of a tumor can lead to better and less invasive treatments. The challenge in this step is to deal with more similar classes since both belong to the same tumor subtype.

3 Datasets

In this study, we used a total of two datasets from the International Genome Cancer Consortium (ICGC) [13] [14] from the Liver primary site tumor types. For the first step of the data analysis pipeline we used the Simple Somatic Mutation data from the studies LICA-FR (Liver tumor samples associated to alcohol) and LIRI-JP (Liver tumor samples associated to Virus).
For the second step of the data analysis pipeline is used the Gene Expression data from the study LIRI-JP corresponding to HCC associated to Virus subtype. LIRI-JP data is provided to ICGC by the RIKEN National Cancer Center and the Human Genome Centre, Institute of Medical Science, University of Tokyo. LICA-FR data is provided to ICGC by the Centre National de Gnotypage (CNG) and the Institut National de la Sant et de la Recherche Mdicale (INSERM). Table 1 and 2 details the number of samples per class in each dataset. The ICGC data portal presents in the LIRI-JP project some extra samples within the SSM dataset than the Gene Expression one. All the available samples in both data sources are used at each step of the data analysis pipeline.

Table 1: HCC subtype simple somatic mutation (SSM) dataset

Tumor subtype	Number of sample	ICGC Study
Virus associated	258	LIRI-JP
Alcohol associated	252	LICA-FR

Table 2: HCC virus associated tumor stage dataset from LIRI-JP study using gene expression data

Tumor stage	Number of sample	Class
Stage I	36	Early stage
Stage II	106	Early stage
Stage III	71	Late stage
Stage IV	19	Late stage

3.1 Data preprocessing

First step: tumor subtype classification In the first step of the data analysis pipeline, preprocessing consists in removing 5% of the outliers by considering the 0.95 quantile as an upper bound of the total amount of mutations per patient. The dataset is reduced from 510 to 484 patients, leaving out the outliers with excessive number of mutations by the assumption these patients are not representative of the true distribution of HCC. Then each feature is normalized between 0 and 1.
Only protein coding genes have been considering for this step. Data is splitted in training and test set where train corresponds to 80% of the data.
Each sample is characterized by a vector where each position corresponds to one mutated gene within the full dataset. Every mutated gene of a sample is filled with the number of mutations within that gene. It is important to remark that for this reason the data matrix is highly sparse.

Second step: Early and Late stage classification Data is splitted in training and test set where train corresponds to 70% of the total amount of samples. Preprocessing consisted in an auto-scaling of the features to obtain a 0 mean and unit standard deviation.

4 Methods

The corresponding section details the statistical learning techniques and methods used along the data analysis pipeline. First Kernel Learning concepts are introduced, then the Kernel Target Alignment score, followed by Support Vector Classification and finally feature selection by Lasso.

4.1 Kernel Learning

Before jumping to classification and feature selection we will introduce kernel learning and its role in this study. Let X be a compact space. The function $k : X \times X \mapsto R$ is symmetric and describes the mapping ϕ from X to a Reproducing Kernel Hilbert Space (RKHS) H through an inner product.

$$K\left(x_i, x_j\right) = \left\langle \phi(x_i), \phi(x_j) \right\rangle_H \tag{1}$$

Here ϕ is a function that maps from X to a feature space H

$$\phi : X \mapsto \phi\left(X\right) \in H \tag{2}$$

Consider the vector of samples x_i belonging to a set S

$$S = \{x_1, , \cdots, x_m\} \tag{3}$$

The Gram Matrix of a Kernel [19] is built from S and is defined as an $m \times m$ matrix G with respective entries $G_{i,j} = \langle x_i, x_j \rangle$. The application of a Reproducing Hilbert Space Kernel function to compute the inner product between training vectors with a feature mapping ϕ allow to compute the following gram matrix:

$$G_{i,j} = \langle \phi\left(x_i\right), \phi\left(x_j\right) \rangle = k\left(x_i, x_j\right) \tag{4}$$

where each position of the Gram matrix if $G_{i,j} = 0$ corresponds to orthogonal training vectors i and j. In this work we use Gram matrices to compute similarity between samples. We refer to high similarity when a pair of training vectors correspond to a value closer to 1 and a lower similarity when the gram matrix coordinates of two pair of vectors is closer to 0.

4.2 Kernel Target Alignment

Given a set S of labeled samples

$$S = \{\left(x_1, y_1\right), \cdots, \left(x_m, y_m\right)\} \tag{5}$$

and each $x_i \subseteq R^n$ is labeled with y_i where $Y = \{+1, -1\}$, a kernel K can be built from equation (1).

We start defining the kernel target alignment (KTA) [20] as a measure of similarity between a given kernel K and an ideal kernel K_y built considering the label vector Y. Here K_y gram matrix gives a value $G_{ij} = 1$ where x_i and x_j belong to the same class. On the other hand, when the two sample vectors x_k and x_t belong to different classes $G_{kt} = 0$. The kernel target alignment of two kernels K and Ky with respect a sample S is the quantity:

$$A\left(S, K, yy'\right) = \frac{\langle K, yy' \rangle_F}{\sqrt{\langle K, K \rangle_F \langle yy', yy' \rangle_F}} = \frac{y'Ky}{S\left\|K\right\|_F} \tag{6}$$

The alignment $A\left(K, K_y\right)$ is the normalized Frobenius inner product between the Gram matrix and the target Matrix K_y. In this work we aim to maximize the KTA and thus tune the kernel for discriminant analysis [21]. Increasing the Frobenius inner product corresponds to increase the inter-cluster distance in the feature space [22]. The greater the KTA of a given kernel, the better the similarity obtained between samples of the same class and at the same time low similarity between samples from different classes. In this work KTA is computed with gram matrices built with the full feature space and with the reduced feature subset after feature selection. KTA is analyzed to understand if the feature selection methods improves KTA and thus improves the similarity of samples of the same class.

4.3 Support Vector Classification

During the two steps of the data analysis pipeline a binary classification model is built before and after feature selection. The classification task used is Support Vector Classification [23], which builds a nonlinear rule by constructing a linear boundary in a transformed and high dimensional version of the feature space. Some characteristics of support vector machines binary classifiers used in this paper are described briefly. In binary classification our goal is to estimate a function $f : R \to \{+1, -1\}$ from training data samples x_i with label Y_i. Support Vector Classification aims to estimate a hyperplane

$$x : f(x) = x^T \beta + \beta_0 = 0 \tag{7}$$

corresponding to the decision function:

$$D\left(x\right) = sign\left[x^T \beta + \beta_0\right] \tag{8}$$

The decision function D obtained corresponds to the hyperplane which maximizes the separating margin M between the two classes where $M = 1/\left\|\beta\right\|$. Now supose that both classes overlap and are not linearly separable. A set of slack variables $\xi = (\xi_1, ..., \xi_m)$ are defined to allow some miss classifications when a sample fall on the wrong side of the margin.

Then a convex optimization problem is expressed in equation 9 where the cost C parameter penalizes every miss classification.

$$min\frac{1}{2}\left\|\beta\right\|^2 + C\sum_{i=1}^{m}\xi_i \tag{9}$$

$$s.t.\xi_i \geq 0; y_i\left(x^T\beta + \beta_0\right) \geq 1 - \xi_i \tag{10}$$

After a quadratic programming solution applying Lagrange Multipliers the solution of β is expressed as

$$\hat{\beta} = \sum_{1}^{m}\hat{\alpha}_i y_i x_i \tag{11}$$

with non zero coefficient $\hat{\alpha_i}$ only for the samples lying on the edge of the margin or for the miss classified ones. These samples are known as support vectors.

4.4 Support Vector Classification and Kernels

Since the support vector classifier finds a linear boundary in the input feature space, a new feature space can be obtained by a transformation ϕ and thus make possible the resolution in complex problems where classes are highly overlapped and are not linearly separable.

The core idea is to apply a transformation/mapping to the input feature vector X and then use linear models in the new space. The transformation is denoted as ϕ where ϕ_m corresponds to the m_{th} transformation of X and $m = 1...M$. Then the decision function can be written as

$$f(x) = \phi\left(x\right)^T \beta + \beta_0 = \sum_{i=1}^{M} \alpha_i y_i \left\langle \phi\left(x\right), \phi\left(x'\right) \right\rangle + \beta_0 \tag{12}$$

where $\phi(x)$ is used only for inner products. For this reason it is not necessary to determine the transformation $\phi(x)$ but it is required to know the positive and semi-definite kernel function $K\left(x_i, x_j\right) = \langle \phi(x_i), \phi(x_j) \rangle$ responsible to compute the inner products in the transformed space.

In this study we use the Gaussian Kernel function K_g.

$$K_g(x_i, x_j) = \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right); \sigma > 0 \tag{13}$$

and the hyperparameter σ has been tuned to improve the KTA of K_g.

4.5 Feature Selection

In this work feature selection is performed by Least Absolute Shrinkage and Selection Operator (LASSO) [23]. Lasso technique aims to improve the generalization capacity and performance of a classifier by selecting a subset of features using a penalization term β_j from a full set of N features.

$$(\beta_0, \beta) = argmin\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_j \beta_j x_{ij}\right)^2\right\} \tag{14}$$

$$s.t. \sum_j |\beta_j| \leq t \tag{15}$$

The parameter t of the equation (15) defines the solution domain for β. Reducing t reduces the solution domain and force more coefficients β_i to reduce to 0. In this work the t parameter is determined by a grid search of different values of t evaluated in a 5-fold cross validation on training set where the t parameter with the minimal corresponding error is selected. In the next section Lasso method is used to select a subset genes in both steps of the data pipeline.

4.6 Computational tools

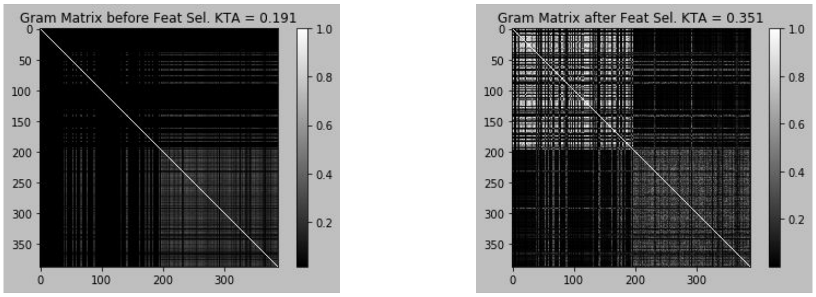
All the experiments have been executed in Python 3.6. Feature selection and support vector classification engine have been used from the Scikit-Learn Python library (https://scikit-learn.org/) [29].

5 Results and Discussion

The methods and techniques detailed in the previous section have been implemented on the datasets according to the data analysis pipeline presented in this work. Feature selection and classification results are discussed and analyzed by comparing classification performance and KTA scores obtained for each step of the data analysis pipeline.

5.1 HCC Tumor sub-type classification using Simple Somatic Mutations

As described in section 2.2 the first step of the data analysis pipeline use Simple Somatic Mutation as features for the two HCC tumor sub-types: Virus and Alcohol associated. The objective of this step is to discriminate accurately both classes and obtain a reduced feature subset that improves the classification performance. Kernel Gram Matrix is built using gaussian kernel before and after the feature selection process and KTA is computed for both matrices (Figure 2).



Before feature selection. KTA = 0.19

After feature selection. KTA = 0.35

Fig. 2: Gram matrix obtained after tuning of sigma parameter using simple somatic mutations as features. Samples are sorted by class along the matrix axis.

Lasso feature selection step identified in a subset of 70 genes with a $t = 0.00084$ after 5 folds cross validation on training data. The selected genes are associated to non-zero β_i parameters of Lasso. Then for different values of sigma a gaussian kernel is built in both scenarios. The KTA score (Table 3) and Support vector classification performance (table 4) are compared using the same kernel before and after feature selection. Given different values of sigma, Table 3 shows that the reduced feature subset produces a higher KTA than the original feature set.

Table 3: KTA score along different values of sigma before and after feature selection using SSM data.

Features	**Score**	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$
19990	KTA	0.001	0.075	0.051	0.050
70	KTA	0.001	0.031	0.3499	0.207

It is clear in Table 4 that the classifier trained in the reduced feature subset has a better performance with almost every value of sigma than the one trained in the original feature set.

Table 4: Classification performance along different values of sigma before and after feature selection using SSM data.

Features	**Score**	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$
19990	AUC	0.929	0.942	0.925	0.510
70	AUC	0.923	0.963	0.969	0.932

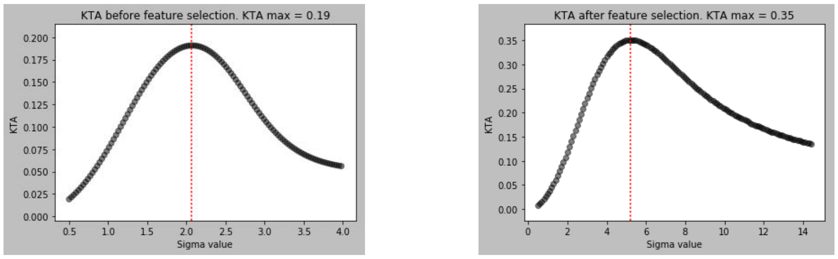
Finally the search of the sigma parameter is refined independently for both scenarios with the objective to obtain a sigma value that improves as much as possible the KTA with the corresponding feature set. Figure 4 shows the Gaussian Kernel tuning process along different values of sigma and the KTA score. Table 3 shows the classification performance using the Kernel with the highest KTA obtained from Figure 4 before and after feature selection. The results show again an improvement of the KTA using the reduced feature subset.

Table 5: Classification performance before and after feature selection using SSM data.

Num. of features	**KTA**	**AUC**	σ
19990	0.19	0.942	2.06
70	0.35	0.969	5.19

Despite the classification performance with the original feature set is acceptable, for biomedical and pattern recognition purposes it is convenient to handle less than 100 genes than a set of almost 20,000. The 70 selected genes make a biological gene signature to discriminate tumor subtypes and improves the classification performance. This signature can be further analyzed in the literature and with biological wet lab experiments. Sigma parameter of Gaussian kernel is tuned by considering the value which improves KTA score (Figure 4).
The maximum KTA score achieved by the Gaussian kernel after feature selection is 0.35 against 0.19 in the original full feature set. This let us observe that high dimensional feature spaces product of somatic mutations in genetic data with

close to 20,000 features does not help to increase the separability between the clusters of each class. In addition, the majority of mutated genes add noise to the classification problem. There is a correlation between the sample to feature ratio and the KTA score obtained. Before feature selection the sample to feature ratio is 0.025. After the feature selection process the sample to feature ratio obtained is 7.28. Figure 2 shows how samples of the same class in the symetric Gram matrix (sorted by class) are more similar using the reduced feature subset than the original one, where two clusters corresponding to both classes are clearly visible.

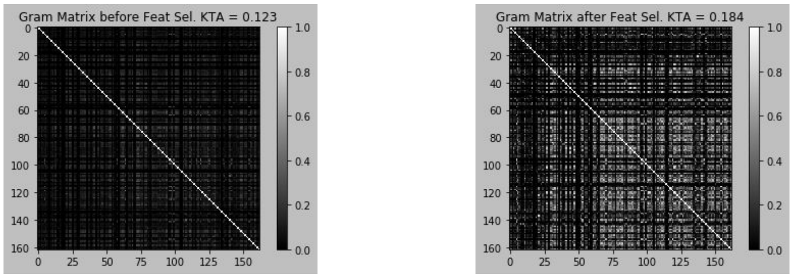


KTA before feature selection. KTA after feature selection.

Fig. 4: KTA by different values of sigma. Maximum KTA score is highlighted with the dashed line.

5.2 HCC Tumor stage classification using Gene Expression data

In this section results of the early and late stage classification for the HCC tumor subtype are shown. The importance of the results relies on the early detection of the tumor and on the selected gene expression panel that improves classification performance. These gene expression panel will help biomedicine scientists to explore the related pathways that impact on the tumor.



Before feature selection. KTA = 0.12. After feature selection. KTA = 0.18.

Fig. 6: Gram matrix obtained after tuning of sigma before and after feature selection using gene expression data.

During the feature selection process the Lasso algorithm selected 9 deferentially expressed genes and transcript locations. Lasso is tuned by a 5 folds cross validation on training set and the penalization parameter result in t = 0.252. Using the same benchmark process as the previous section for different values of sigma, the KTA (Table 6) and classification performance (Table 7) are analyzed before and after feature selection.

Table 6: KTA score along different values of sigma before and after feature selection using Gene Expression data.

Features	**Score**	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$
22913	KTA	0.078	0.078	0.078	0.078
9	KTA	0.070	0.169	0.078	0.078

The highest KTA and classification performance along the different sigma values is obtained with the reduced feature subset (KTA = 0.169).

Table 7: Classification performance along different values of sigma before and after feature selection using Gene Expression data.

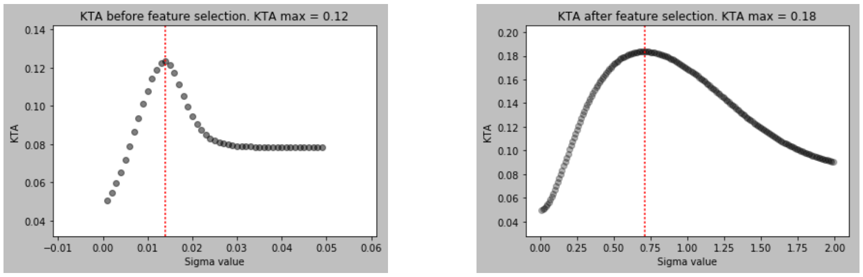
Features	**Score**	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$
22913	AUC	0.5	0.5	0.5	0.5
9	AUC	0.824	0.812	0.5	0.5

Once again, following the same analysis as the previous section, the sigma parameter is tuned at each scenario independently with the objective to obtain a sigma value that improves the KTA with the corresponding feature set. Table 8 shows the results of KTA and classification performance before and after feature selection. In this case the difference between full and reduced feature subset is very significant regarding classification performance. AUC improved by 13%.

Table 8: Classification performance before and after feature selection using SSM data.

Features	**KTA**	**AUC**	σ
22913	0.12	0.682	0.014
9	0.18	0.832	0.71

Gaussian Kernel has been tuned by the sigma parameter with the highest KTA score obtained as seen in Figure 8. Despite the improvement, the KTA is not high enough to distinguish both cluster classes in the gram matrix (Figure 6). Nevertheless the classification performance is acceptable and much better than the original scenario. It is visible in the Gram Matrix after feature selection that one of the classes at the lower right side of the matrix (early stage) is well defined but the other one at the upper left (late stage) is more spread. This can be explored by the fact that at late stage cancer cells deferentially express from one to another.



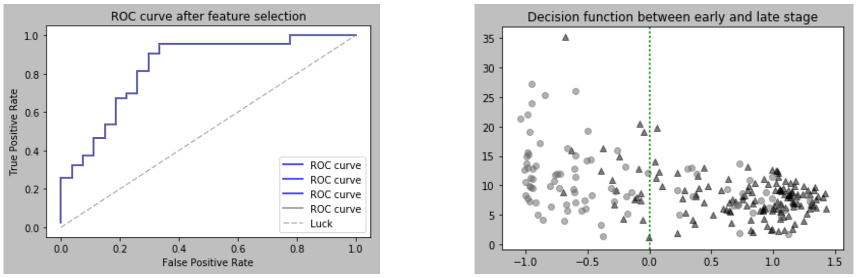
KTA before feature selection. KTA = 0.12 KTA after feature selection. KTA = 0.18

Fig. 8: KTA by different values of sigma. Maximum KTA score is detailed with a dashed line.

Despite this observation, support vector classifier could classify well according to the complexity of the data when the two classes belong to the same tumor subtype. Figure 10 details the area under the curve ROC for test set using Gene Expression data for the classification between early and late stage of HCC. We benchmark our result with Roessler et Al. [8] work where Tumor Relapse detection in Early Stage is achieved with also an AUC-ROC score of 0.83 even though both results came from different datasets. This comparison indicates that the results obtained are acceptable.

Figure 10 also describes how the full dataset samples distribute considering the hyperplane decision function learned, where one of the selected features the gene 'CDK14' is used as Y axis for visualization purposes only. The mean and standard deviation in the early stage samples (triangles and positive) of the value for the hyperplane decision boundary is 0.75 and 0.51. On the other hand for the late stage samples mean is -0.25 and standard deviation 0.73. It is visible that early stage (triangles) samples tend to be far away of the hyperplane than the late stage samples (circles) which present more dispersion and are closer to the hyperplane.

Table 9 shows 5 of the 9 deferentially expressed genes between early and late stages that were identified by other HCC virus associated studies in the literature. The gene names (Gene Symbol) are EXOC4 [26], IER3IP1 [25],CDK14 [24],ENO1 [27] and PPP2R1A [28]. A Mann-Whitney U test has been applied to the expression between classes for each gene and the corresponding p-value is reported. Despite this statistical evidence, additional biomedical validation has to be done on these genes.



Area under the ROC curve = 0.832.

Test samples and the decision function.

Fig. 10: Area under the ROC curve and distribution of samples along the hyperplane decision function for test data.

Table 9: Deferentially expressed genes: list of selected genes and analysis of significant differences between early and late stages.

Chromosome	Gene symbol	p-value
7	EXOC4	6.57e-07
18	IER3IP1	6.30e-06
7	CDK14	1.73e-05
1	ENO1	1.61e-05
19	PPP2R1A	4.01e-05

6 Conclusion and future work

We have presented an data analysis pipeline to classify Virus associated against Alcohol associated Hepatocellular carcinoma tumor sub-types using Simple Somatic Mutations per gene in tumor samples of patients. Then during the second step of the data analysis pipeline gene expression data is used to characterize early and late stages of the Virus associated tumor sub-type. In both steps of the data analysis pipeline, feature selection by Lasso and Classification with Support Vector Machines is performed. Kernel target alignment is measured with the full and reduced feature subset and KTA evolution is observed during Gaussian Kernel tuning process. This work has two main contributions. The first one is a bioinformatics and biomedical approach, where two gene signatures are identified: one for SSM and the other for Gene expression. These signatures have been obtained as biomarkers to accurately classify tumor sub-types and early stages of HCC Virus subtype. The second contribution is the clear evidence of the impact of kernel tuning applied in genomics and transcriptomics data of cancer by considering KTA the as objective function. Finally, the tumor classification and detection problem is approached from a multi-omics perspective, considering two layer of biological information. Future work should include a multi-kernel learning approach to combine different multi-omics layers like genomics, transcriptomics, proteomics and metabolomics for each patient. Also the development of new feature selection methods by kernel learning [30] is an opportunity to detect complex patterns in biomedical data.

Acknowledgements

The authors would like to thank the International Cancer Genome Consortium team, the Liver Cancer Hepatocellular carcinoma (Virus associated project) LIRI-JP by Riken in Japan and the Liver Cancer Hepatocellular carcinoma (Secondary to alcohol and adiposity) LICA-FR project by the Institut National du Cancer in France who have provided data.

Grant support

This work was supported by the doctoral program of the Universidad Tecnologica Nacional in Argentina, the UTN doctoral school of Signal Processing and the Universite de Technologie de Troyes. Also it was funded by grants from CONICET, ANPCyT and FOCEM-Mercosur.

References

1. Mardis, E. R. (2017). DNA sequencing technologies: 20062016. Nature protocols, 12(2), 213.
2. Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., ... Shi, T. (2018). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Frontiers in genetics, 9.
3. Amar, D., Izraeli, S., Shamir, R. (2017). Utilizing somatic mutation data from numerous studies for cancer research: proof of concept and applications. Oncogene, 36(24), 3375.
4. Patel, S., Ahmed, S. (2015). Emerging field of metabolomics: big promise for cancer biomarker identification and drug discovery. Journal of pharmaceutical and biomedical analysis, 107, 63-74.
5. Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W., Quackenbush, J. (2018). Cancer subtype identification using somatic mutation data. British journal of cancer, 1.
6. Bhalla, S., Chaudhary, K., Kumar, R., Sehgal, M., Kaur, H., Sharma, S., Raghava, G. P. (2017). Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. Scientific reports, 7, 44997.
7. Ressom, H. W., Di Poto, C., Ferrarini, A., Hu, Y., Ranjbar, M. R. N., Song, E., ... Zuo, Y. (2016, August). Multi-omic approaches for characterization of hepatocellular carcinoma. In Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the (pp. 3437-3440). IEEE.
8. Roessler, S., Jia, H. L., Budhu, A., Forgues, M., Ye, Q. H., Lee, J. S., ... Wang, X. W. (2010). A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer research, 70(24), 10202-10212.

9. Ally, A., Balasundaram, M., Carlsen, R., Chuah, E., Clarke, A., Dhalla, N., ... Marra, M. A. (2017). Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell, 169(7), 1327-1341.
10. Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. AMS math challenges lecture, 1(2000), 32.
11. Cristianini, N., Shawe-Taylor, J., Elisseeff, A., Kandola, J. S. (2002). On kernel-target alignment. In Advances in neural information processing systems (pp. 367-373).
12. Huang, S., Cai, N., Pacheco, P. P., Narandes, S., Wang, Y., Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics-Proteomics, 15(1), 41-51.
13. Jennings, J. L., Stein, L. D., Calvo, F. (2017). International Cancer Genome Consortium (ICGC).
14. International Cancer Genome Consortium. (2010). International network of cancer genome projects. Nature, 464(7291), 993.
15. Chen, Y., Sun, J., Huang, L. C., Xu, H., Zhao, Z. (2015). Classification of cancer primary sites using machine learning and somatic mutations. BioMed research international, 2015.
16. Vural, S., Wang, X., Guda, C. (2016). Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. BMC systems biology, 10(3), 62.
17. Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., ... Han, L. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nature biotechnology, 32(7), 644.
18. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.
19. Shawe-Taylor, J., Cristianini, N. (2004). Kernel methods for pattern analysis. Cambridge university press.
20. Cristianini, N., Kandola, J., Elisseeff, A., Shawe-Taylor, J. (2006). On kernel target alignment. In Innovations in Machine Learning (pp. 205-256). Springer, Berlin, Heidelberg.
21. Wang, T., Zhao, D., Tian, S. (2015). An overview of kernel alignment and its applications. Artificial Intelligence Review, 43(2), 179-192.
22. Ramona, M., Richard, G., David, B. (2012). Multiclass feature selection with kernel gram-matrix-based criteria. IEEE transactions on neural networks and learning systems, 23(10), 1611-1623.
23. Friedman, J., Hastie, T., Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York, NY, USA:: Springer series in statistics.
24. Huang, J., Deng, Q., Wang, Q., Li, K. Y., Dai, J. H., Li, N., ... Fei, Q. L. (2012). Exome sequencing of hepatitis B virusassociated hepatocellular carcinoma. Nature genetics, 44(10), 1117.
25. Yiu, W. H., Yeung, T. L., Poon, J. W. M., Tsui, S. K. W., Fung, K. P., Waye, M. M. Y. (2010). Transcriptional regulation of IER3IP1 gene by tumor necrosis factor and Sp family proteins. Cell Biochemistry and Function: Cellular biochemistry and its modulation by active agents or disease, 28(1), 31-37.
26. Saitta, C., Tripodi, G., Barbera, A., Bertuccio, A., Smedile, A., Ciancio, A., ... Pollicino, T. (2015). Hepatitis B virus (HBV) DNA integration in patients with occult HBV infection and hepatocellular carcinoma. Liver International, 35(10), 2311-2317.

27. Yoon, S. Y., Kim, J. M., Oh, J. H., Jeon, Y. J., Lee, D. S., Kim, J. H., ... Kim, Y. S. (2006). Gene expression profiling of human HBV-and/or HCV-associated hepatocellular carcinoma cells using expressed sequence tags. International journal of oncology, 29(2), 315-327.
28. Chen, H. F., Mai, J. R., Wan, J. X., Gao, Y. F., Lin, L. N., Wang, S. Z., ... Liao, K. (2013). Role of a novel functional variant in the PPP2R1A promoter on the regulation of PP2A-Aalpha and the risk of hepatocellular carcinoma. PLoS One, 8(3), e59574.
29. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011
30. Palazzo, M., Beauseroy, P., Koile, D., Yankilevich, P. (2018, November). Learning Kernels from genetic profiles to discriminate tumor subtypes. In IV Simposio Argentino de GRANdes DAtos (AGRANDA 2018)-JAIIO 47 (CABA, 2018).