Integrative multi-omics data-driven approach for metastasis prediction in cancer

Carlos Fernandez-Lozano
Computer Science Department, University of A Coruna. Faculty of Computer Science, A Coruña, Spain

Jose Liñares Blanco
Computer Science Department, University of A Coruna. Faculty of Computer Science, A Coruña, Spain
j.linares@udc.es

Marcos Gestal
Computer Science Department, University of A Coruna. Faculty of Computer Science, A Coruña, Spain
marcos.gestal@udc.es

Julian Dorado
Computer Science Department, University of A Coruna. Faculty of Computer Science, A Coruña, Spain
julian@udc.es

Alejandro Pazos
Computer Science Department, University of A Coruna. Faculty of Computer Science, A Coruña, Spain
apazos@udc.es

ABSTRACT
Nowadays biomedical research is generating huge amounts of omic data, covering all levels of genetic information from nucleotide sequencing to protein metabolism. In the beginning, data were analyzed independently losing a great deal of essential information in the models. Even so, complex metabolic routes and genetic diseases could be determined. In the last decade, there has been an ever-increasing number of research projects that follow a systemic biological approach by integrating multiple omic datasets obtaining more complex, powerful and informative models that provide a deeper knowledge about the genotype-phenotype interactions. These models greatly contributed to the study of complex multi- factorial diseases such as cancer. The onset and development of any type of cancer can be influenced by multiple variables. Integrate as many as possible omic datasets is therefore the best approach to extract all the underlying knowledge. A significant factor in the mortality of this disease is the metastatic process. The identification of the factors involved in this cell behavior may be helpful in the diagnosis and hopefully in the disease prevention.
The development of novel integrative multiomics approaches is an opportunity to fill the gaps between our ability to generate data and the difficulties to understand the biology behind them. In this work we propose a methodology pipeline for analyze multi-omics data using machine learning.

CCS CONCEPTS
Computing methodologies; Artificial intelligence; Machine learning; Modeling and simulation

KEY WORDS
Multi-Omics, Cancer, Data-Driven Modelling, Data Integration, Machine Learning, Prediction, Data Fusion

1 INTRODUCTION

Omics in Molecular Biology refers to the study of the whole of something. Omics aims at the collective characterization and quantification of sets of biological molecules that are translated into the structure, function and dynamics of an organism or several organisms at the same time.

In the beginning, these different sources of omic data were analyzed separately by means of conventional statistical techniques to contrast hypotheses in order to obtain those variables that have the most biological significance. Nowadays, Artificial Intelligence algorithms are being presented as an alternative to this analysis. These techniques, which basically learn through examples, give researchers the ability to analyze and extract knowledge of huge amounts of data at the same time, independently of the dimension of the problem. Furthermore, this would not be possible with conventional statistical approaches in which the usual tests are not designed for data sets of such high dimensionality [11].

Due to the emergence of Next Generation Sequencing (NGS) techniques and moreover to the drastic reduction of the economical costs of the process, there is for example a growing amount of open access data for the study and characterization of each type of cancer and its different related biological processes. Therefore, more efforts are necessary for the analysis of these sources of omic data.

These models greatly contribute to the study of complex multifactorial diseases such as cancer. The onset and development of any type of cancer can be influenced by multiple factors so the consideration the as much of these sources as possible is probably the best way to go. Furthermore, a critical factor in the mortality of this disease is the metastatic process. Therefore, the identification of some of the factors significantly associated to this cell behavior may be of great help in the diagnosis and the characterization of subtypes of the same cancer.

Metastasis is the most fatal characteristic of malignancy tumor, which is directly related with more than 90% of tumor-related mortality. Distant organ or tissue metastasis is a sign of poor prognosis in patients with cancer. Cancer cells spreads to other parts of the body through bloodstream or lymph system. Either way, most of the escaped cancer cells die or are killed before they can start growing somewhere else. But one or two might settle in a new area, begin to grow, and form new tumors. The development of a cancer staging system help clinicians with prognosis and allow them to design a treatment plan for individual patients. Therefore, be able to classify a patient according to the cancer stage, will improve early diagnose methods and treatments. Regarding cancers that already have some type of metastasis the TNM Staging System is the most used system for diagnosis. TNM is based on the extent of the tumor (T), the extent of spread to the lymph nodes (N), and the presence of metastasis (M). Therefore the T category describes the original (primary) tumor, the N category describes whether or not the cancer has reached nearby lymph nodes and M category tells whether there are distant metastases (spread of cancer to other parts of the body).

In this context, Machine Learning (ML) techniques can help the clinicians in the diagnosis and treatment of cancer patients. Because of the complexity of the problem, it is necessary to have the as much sources of omic data as possible, integrating them in the same analytic process. In recent years there has been an increase in the number of scientific research proposals that follow a systemic biologic approach, integrating different sources of omic data [4]. Thereby, we are able to obtain more powerful models that give us greater knowledge about the interaction between phenotype and genotype [7].

Unfortunately, for this type of analysis there is no standardized methodology. Therefore, it is necessary to detail and specify a coherent analysis methodology for this type of integrative bioinformatic analysis [2] using machine learning and artificial intelligence.

In this work we detail a methodology to address these problems with biological data hoping that will improve the reproducibility of the experiments.

2 MATERIALS AND METHODS

2.1 Methodology design

The main aim of this work is to present a methodology, Figure 1 of analysis of multiomics data for the design of experiments by means of Machine Learning algorithms. This methodology will provide a standardized solution for the study of metastatic stages of different types of cancer. The ML problems addressed here consist of supervised learning.

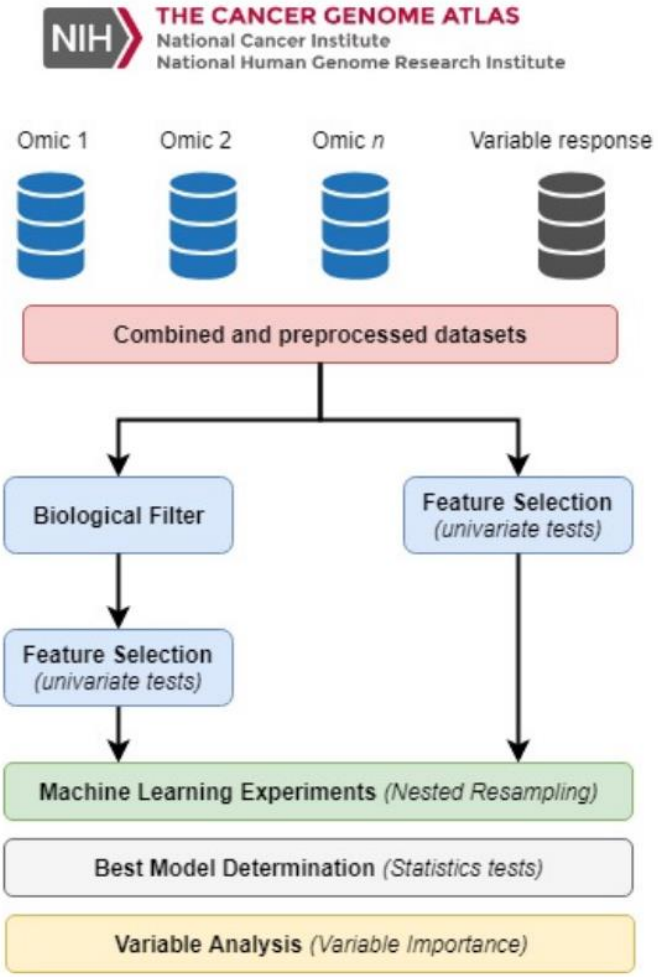

Figure 1: Overall methodology scheme

Generally speaking, most used data sources in cancer research are among others: expression, mutation, methylation, genomic sequence, microbiome, protein and metabolome.

Initially, omic data sources have a high number of dimensions and ML algorithms are not especifically design to deal with that huge number of them. In fact, those algorithms in general do not achieve good performance scores with high-dimensional data or correlated factors. We proposed two different approaches for the reduction of the number of factors for each omic source: biological and mathematical filter.

Using the first filter, our proposal is to reduce the number of genes using biological previous published information. In the case of non genetic data sources such as Methylation, Single Nucleotide Polymorphism (SNP) or Copy Number Aberration (CNA) data we propose the use of an eQTL (Expression quantitative trait loci) analysis to obtain the cis CpGs islands that are higher correlated with the filtered genes. For example, in the case of miRNA data, specific databases, such as miRBase are available to obtain miRNA molecules of interest. Finally, with the remaining biological related factors is time to use a mathematical filter. To this aim, our proposal is to use a univariate filter Feature Selection (FS) approach mainly because its faster and independent of the ML algorithm.

Therefore, once the dataset is generated according with biological and mathematical information, several different algorithms must be trained in order to perform a fair comparison of the results, this step should be planned carefully. For the training phase, a nested resampling must be used. This process is proposed in two levels: an independent external cross-validation to evaluate the generalization capability of the model and avoid overfitting and an internal cross-validation for the selection of the best hyperparameters for each algorithm. The performance of the models can be evaluated through several measures such as, area under curve ROC (AUC), accuracy, sensitivity, specificity, etc. AUC is one the most robust performance measures because is independent of the binarization threshold and considers at the same tiem Type I and Type II errors.

2.2 Omic data sources

There are a multitude of free access omic data sources repositories, where researchers can download the data to perform their own ML experiments. In this paper we will focus on The Cancer Genome Atlas (TCGA) [1], since it is the most complete repository to date of cancer-related data. Big data repositories such as TCGA, an open database with multiple sources of omic data from different types of cancer, provide to researchers the ability and ease of downloading a lot of different omic data types [5, 6]. These platforms allow researches to develop and test bioinformatic pipelines with real data.

TCGA initiative has 33 types of cancer and more than seven different data sources. Among them are expression, methylation, mutation, CNA, SNP, miRNA, RNAseq and protein data and also clinical information. Therefore we are able to define the patients through T, N or M Metastatic Stages in metastasis research. These stages determine the metastasis stage of the patients. Thus, we can differentiate patients without metastasis (n0, t0 or m0) from those with metastasis (n1, n2, n3; t1, t2, t3 and m1, m2, m3).

2.3 Data pre-processing

First, those patients with a defined TNM Stage must be selected. Therefore those patients with an NA stage are deleted. Because within each stage there are different substates, these should be grouped in the common stage. For instance, patients with stages 'n0 (i-)' or 'no (i +)' are grouped in stage n0. Subsequently, only patients with primary tumor samples were selected. In this step data visualization (Principal Component Analysis or Multidimensional Scale) is recommended to detect outliers patients and decide if these patients are to be included or not in the analysis. In addition, prior to ML experiments, data should be scaled and constant variables removed.

2.4 Dimensionality reduction

In a high dimensional dataset, some features may be redundant or correlated and some noise can occur. To take better advantage of the space of the solutions, the quantity of useless factors must be reduced to the maximum. In this work, two different approaches are proposed to this aim. The first one is the use of a biologic data-driven approach, while the second was performed only with mathematical information, without any biological information of the factors.

The objective of the biological filter is to extract a subset of factors CpGs islands, etc.) related to the biological problem, in this case, the metastasis process. This subset of features can be obtained from different initiatives such as The Human Protein Atlas (HPA) [10], Kyoto Encyclopedia of Genes and Genomes (KEGG) [3], Wikipathways [9], etc. These initiatives classify genes through location, metabolic process, function, etc.

After this reduction of dimensionality, the number of factor is potentially much higher than the number of observations. In ML, there are three main approaches for FS: filter, wrapped and embedded [8]. The main difference between the filter approach and the others is that the filter is independent of the algorithm. In this work we proposed the use of a univariate filter feature selection approach due also to its speed and simplicity. There are other approaches for dimensionality reduction in ML but FS retains the meaning of the factors.

2.5 Machine Learning

Machine learning is the field of study interested in the development of computational algorithms capable of transforming data into intelligent actions. This field is extensive in several areas, as it helps to explain and extract specific knowledge from a set of observations that humans would not be able to easily perform by themselves. There are mainly two types of learning: supervised and unsupervised. The principal difference between them is that in the former, learning occurs via labeled observations, while in the latter, the examples are not labeled, and the algorithm seeks to cluster the data into different groups.

The number of ML algorithms is increasing day by day, and each of them has a particular set of hyperparameters that should be tuned to find the best possible combination and, consequently, the best solution to the problem. ML algorithms are very powerful techniques, but the training process is critical. This kind of algorithms learn with samples, so the same samples should not be used for learning, validation or hyperparameter tuning.

Hence, a two levels nested resampling is recommended. As mentioned before, this type of resampling consists in an external cross- validation and an independent internal cross-validation for hyper- parameter tuning.

2.6 Best model determination

The final step in the proposed methodology is the statistical significance comparison of the performance of the machine learning models. Winner model should be identified using statistical tests to evaluate the statistical significance. Parametric tests have more power than non-parametric tests but, unfortunately, can be used only under certain circumstances (independence, normality and heterocedasticity). If one of the conditions is not met, a non-parametric test must be performed. In addition, a post hoc procedure must be used to correct and adjust p-values for multiple comparisons.

2.7 Analysis of the importance of the solution

Finally, an analysis of the importance of each factor in the model must be performed. The main objective is to understand the weight of each factor in the best model. Final signature should be compared with previous findings in the scientific literature. With our approach, another point of interest is to analyze which factors were selected by the biological and the mathematical filters.

3 CONCLUSIONS

In very high dimensional omic problems, the number of factors need to be reduced as much as possible without renouncing to the underlying biological information of the issue under analysis. Besides, the use of open databases of omic data sources to valorize the economic expenditure and design new pipelines/analytical techniques is vitally important. The genes associated with the type of cancer under study can be filtered with a biological approach considering specific functions and/or locations. In addition, other data types such as SNP, CNA or methylation can be filtered through eQTL analysis to reduce the number of variables to only those directly associated with the genes of interest.

The integration of biologically-based omic data sources and the subsequent selection of factors using automated learning techniques improve the understanding and identification of cancer metastasis. These ML techniques give us more knowledge about analyzed problem than conventional statistical techniques.

ACKNOWLEDGMENTS

This work is supported by "Collaborative Project in Genomic Data Integration (CICLOGEN)" PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013-2016 and the European Regional Development Funds (FEDER) - "A way to build Europe". This project was also supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. ED431G/01, ED431D 2017/16) and the "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23), and finally by the Spanish Ministry of Economy and Competitiveness for its support with the funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union and the "Juan de la Cierva" fellow- ship program supported by the Spanish Ministry of Economy and Competitiveness (Carlos Fernandez-Lozano, Ref. FJCI- 2015-26071).

REFERENCES

[1] Uma R Chandran, Olga P Medvedeva, M Michael Barmada, Philip D Blood, Anish Chakka, Soumya Luthra, Antonio Ferreira, Kim F Wong, Adrian V Lee, Zhihui Zhang, et al. 2016. TCGA expedition: a data acquisition and management system for TCGA data. PloS one 11, 10 (2016), e0165395.

[2] Carlos Fernandez-Lozano, Marcos Gestal, Cristian R Munteanu, Julian Dorado, and Alejandro Pazos. 2016. A methodology for the design of experiments in computational intelligence with multiple regression models. PeerJ 4 (2016), e2721.

[3] Minoru Kanehisa and Susumu Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28, 1 (2000), 27–30.

[4] Minseung Kim, Navneet Rai, Violeta Zorraquino, and Ilias Tagkopoulos. 2016. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nature communications 7 (2016), 13090.

[5] Yang Liu, Nilay S Sethi, Toshinori Hinoue, Barbara G Schneider, Andrew D Cherniack, Francisco Sanchez-Vega, Jose A Seoane, Farshad Farshidfar, Reanne Bowlby, Mirazul Islam, et al. 2018. Comparative molecular analysis of gastrointestinal adenocarcinomas. Cancer Cell 33, 4 (2018), 721–735.

[6] Cancer Genome Atlas Research Network et al. 2017. Integrated genomic characterization of oesophageal carcinoma. Nature 541, 7636 (2017), 169.
[7] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. 2015. Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics 16, 2 (2015), 85.
[8] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. 2007. A review of feature selection techniques in bioinformatics. bioinformatics 23, 19 (2007), 2507–2517.
[9] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L Coort, Daniela Digles, et al. 2017. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic acids research 46, D1 (2017), D661–D667.
[10] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. 2015. Tissue-based map of the human proteome. Science 347, 6220 (2015), 1260419.
[11] Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M Hoffman. 2018. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. Information Fusion (2018).