Research Project Proposal: Integrative analysis of transcriptional, mutational and DNA structural profiles in ovarian cancer of chemotherapy sensitive vs. resistant patients

Sara Sansone, sara.sansone@mail.polimi.it

1. Introduction to the problem

The problem targeted by this research is positioned in the field of Genomic computing. Genomic computing is a new science focused on understanding the functioning of the genome in order to make fundamental discoveries in biology and medicine. The challenge is to answer to relevant questions for biological and clinical research, e.g. how cancer arises and develops, how driving mutations occur, how much complex disorders such as cancer depend on environmental factors or genetic predisposition; and then use individual genomic information for personalized/precision medicine.This work is focused in particular on the integrative analysis of one specific disease, i.e. ovarian cancer.
Ovarian cancer is the deadliest gynecologic malignancy, with a 5-year survival rate of approximately 47%, a percentage that has remained constant over the past two decades. Most ovarian cancers are epithelial in origin and their treatment prioritizes surgery and cytoreduction followed by cytotoxic platinum and taxane chemotherapy. While most tumors will initially respond to this treatment, recurrence is likely to occur within a median of 16 months for patients who have the disease at an advanced stage. Thus, this research project is concerned about stage III and IV of ovarian cancer and in particular, high-grade serous ovarian adenocarcinoma[1](HGS-OC). It is a rapidly growing carcinoma believed to have tubal origin with a high chromosomal instability; its peculiarity stands in the relapse timing of the patients affected by it. Indeed, patients can be recognized and differentiated into three classes, according to the time elapsed from the end of the first line therapy to relapse:

- relapse within 6 months since the end of treatment: *resistant*;
- relapse after 12 months since the end of treatment: *sensitive*;
- relapse after 36 months since the end of treatment: *sensitive long term*.

HGS-OC generally responds to platinum-based chemotherapy, but 80% of the patients relapse within 18 months from the diagnosis and progressively becomes resistant to treatment, up to becoming incurable: less than 20% of the patients survive after five years from the initial diagnosis.
For this reason, new treatment options separate from traditional chemotherapy, which consider achievements in understanding of the pathophysiology of ovarian cancer are needed to improve outcomes. Moreover, it is crucial to find a mechanism that allows to identify and discriminate resistant and sensitive patients, at the time of diagnosis. Hence, this study involves the analysis of resistance to chemotherapy in ovarian cancer patients, based on their transcriptional, mutational, and DNA structural profiles, in particular of the molecular differences between patients that are sensitive to therapy compared to patients that are resistant, using both in-house data and data sets from the TCGA (The Cancer Genome Atlas).
The ultimate aim is the identification of a molecular signature that could be used to predict the response to therapy (sensitive / resistant) at the time of diagnosis, starting from the Copy Number Alteration (CNA) profiles of the

[1]Adenocarcinoma: malignant epithelial tumor that originates specifically from cells of the glandular epithelium

patients.

2. Main related works

Studies have been done on chemo-resistance, either considering the problem in general or analyzing it with respect to ovarian cancer. For example, it has been implemented an algorithm for classification of cell line chemosensitivity based on gene expression profiles alone [3].
Also, the lack of successful treatment strategies led The Cancer Genome Atlas researchers to measure comprehensively genomic and epigenomic abnormalities on clinically annotated HGS-OvCa samples to identify molecular abnormalities that influence pathophysiology, affect outcome and constitute therapeutic targets [2].
Moreover, integrated analysis of DNA methylation and gene expression [1] revealed signaling pathways related to platinum resistance in ovarian cancer.
Nevertheless, none of this works tried to use the Copy Number Altered regions of the genome to implement a classifier able to identify chemo-resistance. Moreover, even if comprehensive analysis of the genomic profiles of ovarian cancer patients have been carried out, none of them led to a solution to the problem of predicting the resistance at the time of diagnosis.

3. Research plan

The goal of the research is to study the possibility of building a classifier able to predict the chemotherapy resistance of a patient affected by high serous ovarian cancer. In particular, provided as input the CNA profile of a patient, it has to predict if she will be sensitive or resistant to the therapy.
The hope is to find a molecular signature that could be used to predict the response to therapy (sensitive / resistant) at the time of diagnosis with an accuracy of at least 80%.

The project plan is divided in four different tasks, that will be described in the following.

3.1. Data extraction

The bulk of the analyses is carried out on the data from TCGA (project TCGA-OV). There are no information on the drug resistance in the Genomic Data Commons, however the original publication from 2011 (The Cancer Genome Atlas, Nature, 2011; Supplementary data at https://tcga-data.nci.nih.gov/docs/publications/ov_2011/) contained the chemotherapy status (sensitive / resistant) for 90 resistant samples, 156 sensitive samples and 32 sensitive long term samples.

Samples are selected from metadata repository as follows:

1. Locate the barcode (TCGA-XXX-YYY) for samples marked as “Sensitive” or “Resistant” in the original XLS file (https://tcga-data.nci.nih.gov/docs/publications/ov_2011/TCGA-OV-Clinical-Table_S1.2.xlsx) in the “Platinum Status” column. Then the Progression-free Survival column is checked to discriminate sensitive and sensitive long term (PFS months > 36). The barcodes correspond to the clinical__shared__bcr_patient_barcode column in the GMQL data sets.

2. Obtain three lists, one for Sensitive, one for Sensitive long term and one for Resistant patients, after executing three different query on GMQL (GenoMetric Query Language).

3.2. Data analysis

Before implementing the classifier itself, a visual analysis of the data is carried out. The aim of this task is to understand more deeply the problem and to figure it in which regions of the genome the three groups of patients

differ the most. In this way, it will be possible to build a classifier which will take into account only the relevant copy number alterations.

3.3. Implementation

Different kind of classifiers can be implemented and tested. Indeed, it is possible to use either only CNA data or to use relevant CNA regions in order to identify a set of genes, whose expression will then be used to classify patients. In particular, the possibility to use a tool called GISTIC to identify those relevant regions is considered. In fact, the GISTIC module identifies regions of the genome that are significantly amplified or deleted across a set of samples.
After creating the data set, some known classifier might be used, e.g. Random Forest or K-Nearest Neighbours. In order to identify the best model, a 10-fold cross validation will be executed for each proposed classifier.

3.4. Validation

At the end, a validation of the obtained model will be done using in-house data, which are never used during the training face.

A Giantt diagram of the tasks is showed in Figure 1.

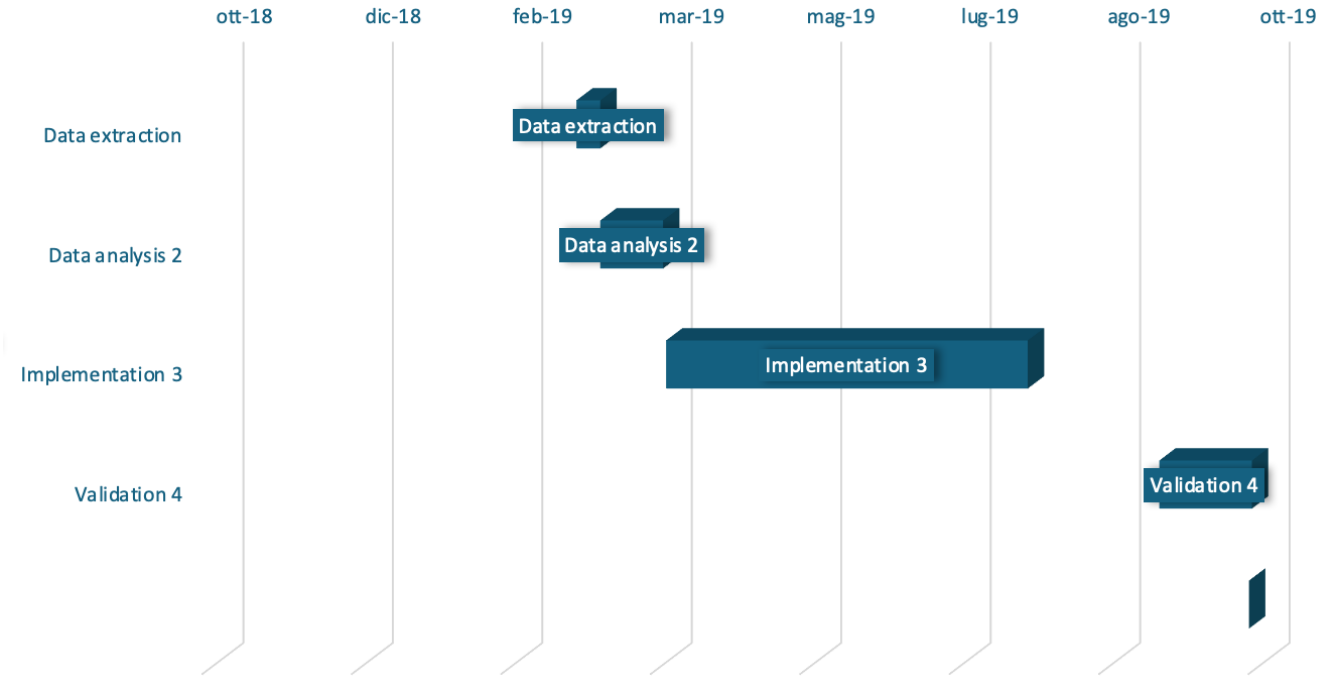

Figure 1: Gantt diagram of the tasks

Finally, the classical metrics for classification problems will be used to evaluate the output of the research. In particular, precision, recall, f1-score and accuracy will be computed. Also, some visual metric as precision-recall curve will be provided.

References

[1] Li, M., Balch, C., Montgomery, J. S., Jeong, M., Chung, J. H., Yan, P., Huang, T. H., Kim, S., and Nephew, K. P. Integrated analysis of dna methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC Medical Genomics 2*, 1 (Jun 2009), 34.

[2] Network, T. C. G. A. R., and Bell, e. A. Integrated genomic analyses of ovarian carcinoma. *Nature 474* (06 2011), 609 EP –.

[3] Staunton, J. E., Slonim, D. K., Coller, H. A., Tamayo, P., Angelo, M. J., Park, J., Scherf, U., Lee, J. K., Reinhold, W. O., Weinstein, J. N., Mesirov, J. P., Lander, E. S., and Golub, T. R. Chemosensitivity prediction by transcriptional profiling. *Proceedings of the National Academy of Sciences 98*, 19 (2001), 10787–10792.