Prediction of Resistance to Chemotherapy in High Grade Serous Ovarian Adenocarcinoma

Sara Sansone
sara.sansone@mail.polimi.it
Track CSE - Data, Web and Society
Introduction to the Research Project:
A joint collaboration

Sara Sansone
Computer Science and Engineering

Giada Lalli
Biomedical Engineering
Introduction to the Research Project: A joint collaboration

Sara Sansone  
Computer Science and Engineering

Giada Lalli  
Biomedical Engineering

Prof. Stefano Ceri  
Supervisor

Dr. Arif Canakoglu, Dr. Pietro Pinoli  
Co-supervisors

Prof. Francesca Ieva (MOX)  
Co-supervisor
Introduction to the Research Project:
A joint collaboration

Sara Sansone
Computer Science and Engineering

Giada Lalli
Biomedical Engineering

Sergio Marchini
Biologist

Luca Beltrame
Bioinformatician

Prof. Stefano Ceri
Supervisor

Dr. Arif Canakoglu, Dr. Pietro Pinoli
Co-supervisors

Prof. Francesca Ieva (MOX)
Co-supervisor
Introduction to the Research Project:
Problem under study

- Ovarian cancer
Introduction to the Research Project:
Problem under study

- Ovarian cancer

**Relative 5-year survival for invasive epithelial ovarian cancer**

- Graph showing survival rates for different stages of ovarian cancer.
Introduction to the Research Project: Problem under study

- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):
Introduction to the Research Project: Problem under study

- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):
  - Rapidly growing carcinoma
Introduction to the Research Project: Problem under study

- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):
  - Rapidly growing carcinoma
  - High chromosomal instability
Introduction to the Research Project:
Problem under study

- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):
  - Rapidly growing carcinoma
  - High chromosomal instability
  - All the patients have a relapse
Introduction to the Research Project: Problem under study

- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):
  - Rapidly growing carcinoma
  - High chromosomal instability
  - All the patients have a relapse
  - They become progressively resistant to the treatment
Treatment:
Surgery and cytoreduction followed by platinum-based chemotherapy
Introduction to the Research Project: Why is it relevant?

Treatment:
Surgery and cytoreduction followed by platinum-based chemotherapy

Patient’s relapse timing:

- Platinum-refractory
- Therapy < 1 month
- ~ 20%
Introduction to the Research Project: Why is it relevant?

Treatment:
Surgery and cytoreduction followed by platinum-based chemotherapy

Patient’s relapse timing:
- Platinum-refractory: < 1 month (~ 20%)
- Platinum-resistant: 1-6 months (~ 80%)
Introduction to the Research Project:
Why is it relevant?

Treatment:
Surgery and cytoreduction followed by platinum-based chemotherapy

Patient’s relapse timing:
- Platinum-refractory: < 1 month
- Platinum-resistant: 1-6 months
- Platinum-sensitive: ≥ 6 months

~ 20% ~ 80%
Introduction to the Research Project: Why is it relevant?

Treatment:
Surgery and cytoreduction followed by platinum-based chemotherapy

Patient’s relapse timing:
- **Platinum-refractory** < 1 month: ~ 20%
- **Platinum-resistant** 1-6 months
- **Platinum-sensitive** ≥ 6 months: ~ 80%

Sensitive short: < 32 months
Introduction to the Research Project: Why is it relevant?

Treatment:
Surgery and cytoreduction followed by platinum-based chemotherapy

Patient’s relapse timing:

- **Platinum-refractory**
  - Therapy: < 1 month
  - ~ 20%

- **Platinum-resistant**
  - 1-6 months

- **Platinum-sensitive**
  - Sensitive short: < 32 months
  - Sensitive long: ≥ 32 months
  - ≥ 6 months
  - ~ 80%
Exploit computational methods to identify a molecular signature that allows to:
Introduction to the Research Project: 
Aim of the work

Exploit computational methods to identify a molecular signature that allows to:

• Predict the response to therapy (resistant / sensitive)
Introduction to the Research Project: 
**Aim of the work**

Exploit computational methods to identify a *molecular signature* that allows to:

- Predict the response to therapy (resistant / sensitive)
- Understand the cause of chemoresistance
Introduction to the Research Project: Aim of the work

Exploit computational methods to identify a molecular signature that allows to:

- Predict the response to therapy (resistant / sensitive)
- Understand the cause of chemoresistance

Genomic regions that differ between resistant and sensitive patients
Introduction to the Research Project: Aim of the work

Exploit computational methods to identify a **molecular signature** that allows to:

- Predict the response to therapy (resistant / sensitive)
- Understand the cause of chemoresistance

**Genomic regions that differ between resistant and sensitive patients**
Data Description:
Genomic data used

Copy Number Alteration (CNA)

Gene expression

miRNA expression

DNA methylation
Data Description:
Genomic data used

Copy Number Alteration (CNA)

Gene expression

DNA methylation
Data Description: Genomic data used

- A genomic region has normally two copies in the DNA, originating from the zygote formation.

- CNAs alter this occurrence in two different ways: amplification and deletion.

- The main focus is on CNA data:
  - Early events
  - May be a signal of the resistance to chemotherapy.
Data Description:
Genomic data used

Copy Number Alteration (CNA)

Gene expression

miRNA expression

DNA methylation

Sansone Sara

POLITECNICO MILANO 1863
Data Description: Genomic data used

- A gene is the basic physical and functional unit of heredity
- The information encoded in the genes are used in the synthesis of functional products, such as proteins
- The process by which it is done is called gene expression
- We are mostly interested in **protein coding** genes:
  - They are related to many cellular functions and biological activities
Data Description:
Genomic data used

Copy Number Alteration (CNA)

Gene expression

miRNA expression

DNA methylation
Data Description: Genomic data used

- Genomic data used:
  - Gene expression
  - miRNA expression
  - DNA methylation

- microRNAs (miRNAs) are small non-coding RNA molecules
- They target multiple genes and can either up-regulate or down-regulate their expression
- They have a causal role in tumorigenesis
Data Description:
Genomic data used

Copy Number Alteration (CNA)

Gene expression

DNA methylation

miRNA expression

miRNA

Sansone Sara
Data Description:
Genomic data used

- DNA methylation is an epigenetic process by which methyl groups are added to the DNA molecule.
- It can change the function of each portion of the genome, by making it more or less accessible.
Data Description:
Datasets

TCGA
THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute
Data Description:

Datasets

TCGA
The Cancer Genome Atlas
National Cancer Institute
National Human Genome Research Institute

GMQL
GenoMetric Query Language
Data Description: Datasets

- TCGA
  - National Cancer Institute
  - National Human Genome Research Institute

- GMQL
  - GenoMetic Query Language

Obtain for each patient data about:
- CNA
- miRNA
- Gene expression
- DNA methylation
Data Description:

Datasets

- **CNA**

<table>
<thead>
<tr>
<th>patient</th>
<th>chrom</th>
<th>start</th>
<th>stop</th>
<th>num_mark</th>
<th>seg_mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>3301764</td>
<td>16104539</td>
<td>7169</td>
<td>0.2480</td>
</tr>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>16108231</td>
<td>16162328</td>
<td>29</td>
<td>0.7084</td>
</tr>
</tbody>
</table>

*Segmented mean*: the $\log_2$ ratio of observed intensity of alteration over reference intensity
Data Description:

Datasets

- **CNA**

<table>
<thead>
<tr>
<th>patient</th>
<th>chrom</th>
<th>start</th>
<th>stop</th>
<th>num_mark</th>
<th>seg_mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>3301764</td>
<td>16104539</td>
<td>7169</td>
<td>0.2480</td>
</tr>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>16108231</td>
<td>16162328</td>
<td>29</td>
<td>0.7084</td>
</tr>
</tbody>
</table>

*Segmented mean*: the $log_2$ ratio of observed intensity of alteration over reference intensity

- **Gene expression**

<table>
<thead>
<tr>
<th>patient</th>
<th>chrom</th>
<th>start</th>
<th>stop</th>
<th>gene_symbol</th>
<th>fpkm</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>11868</td>
<td>14409</td>
<td>DDX11L1</td>
<td>0.000000</td>
</tr>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>14403</td>
<td>29570</td>
<td>WASH7P</td>
<td>23648.321087</td>
</tr>
</tbody>
</table>

*FPKM (Fragments Per Kilobase Million)*: the value of expression, normalized for sequencing depth and gene length
Data Description:

Datasets

- **miRNA expression**

<table>
<thead>
<tr>
<th>patient</th>
<th>chrom</th>
<th>start</th>
<th>stop</th>
<th>mirna_id</th>
<th>rpm</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>17368</td>
<td>17436</td>
<td>hsa-mir-6859-1</td>
<td>0.000000</td>
</tr>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>30365</td>
<td>30503</td>
<td>hsa-mir-1302-2</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

*RPM (Reads Per Million):* the value of expression, normalized for sequencing depth
Data Description:

Datasets

- **miRNA expression**

<table>
<thead>
<tr>
<th>patient</th>
<th>chrom</th>
<th>start</th>
<th>stop</th>
<th>mirna_id</th>
<th>rpm</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>17368</td>
<td>17436</td>
<td>hsa-mir-6859-1</td>
<td>0.000000</td>
</tr>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>30365</td>
<td>30503</td>
<td>hsa-mir-1302-2</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

*RPM (Reads Per Million):* the value of expression, normalized for sequencing depth

- **DNA methylation**

<table>
<thead>
<tr>
<th>patient</th>
<th>chrom</th>
<th>start</th>
<th>stop</th>
<th>gene_symbol</th>
<th>beta_value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>924804</td>
<td>924806</td>
<td>SAMD11</td>
<td>0.009892</td>
</tr>
<tr>
<td>R0_TCGA-13-0720</td>
<td>chr1</td>
<td>925936</td>
<td>925938</td>
<td>SAMD11</td>
<td>0.007828</td>
</tr>
</tbody>
</table>

*Beta value:* the ratio of intensities between methylated and unmethylated alleles
First approach to solve the problem:

Use only CNA data
Steps performed

1. Data preprocessing

2. Feature selection

3. Methods: Classification vs Survival Regression
Data Preprocessing:

CNA profiles

Problem

- A genome wide analysis is needed to identify regions with different CNA between the classes

Solution
Data Preprocessing:
CNA profiles

Problem

- A genome wide analysis is needed to identify regions with different CNA between the classes

Solution

- We create, for each patient, two CNA profiles (for amplification alteration and for deletion alteration)
Data Preprocessing: CNA profiles

**Problem**

- A genome wide analysis is needed to identify regions with different CNA between the classes
- The genome contains 3 billions of base pairs

**Solution**

- We create, for each patient, two CNA profiles (for amplification alteration and for deletion alteration)
Data Preprocessing: CNA profiles

Problem 🔄

• A genome wide analysis is needed to identify regions with different CNA between the classes

• The genome contains 3 billions of base pairs

Solution💡

• We create, for each patient, two CNA profiles (for amplification alteration and for deletion alteration)

• We use bins of size n, i.e., we consider one position as the average of the values of n positions
Data Preprocessing:
CNA profiles

Amplification profiles, resolution of 10Kb

- Red: Resistant
- Green: Sensitive_short
- Blue: Sensitive_long
Data Preprocessing:
CNA profiles

Deletion profiles, resolution of 10Kb
Steps performed

1. Data preprocessing

2. Feature selection

3. Methods: Classification vs Survival Regression
We tried two different approaches to extract relevant CNA regions:

1. Use GISTIC2.0, the state-of-the-art for CNA analysis. GISTIC2.0 is a module able to find regions of the genome that are significantly amplified or deleted in a certain population.

2. Compare CNA profiles of patients of different classes and compute the p-values for the regions using statistical tests:
   - Search for the more suitable test
   - Implementation of a permutation test
   - Use two different thresholds to select the p-values: 0.05, 0.005
We tried two different approaches to extract relevant CNA regions:

1. Use **GISTIC2.0**, the state-of-the-art for CNA analysis

   *GISTIC2.0 is a module able to find regions of the genome that are significantly amplified or deleted in a certain population*
We tried two different approaches to extract relevant CNA regions:

1. Use **GISTIC2.0**, the state-of-the-art for CNA analysis

   *GISTIC2.0 is a module able to find regions of the genome that are significantly amplified or deleted in a certain population*

2. Compare **CNA profiles** of patients of different classes and compute the p-values for the regions using statistical tests:
   - Search for the more suitable test
   - Implementation of a permutation test
   - Use two different thresholds to select the p-values: 0.05, 0.005
Steps performed

1. Data preprocessing

2. Feature selection

3. Methods: Classification vs Survival Regression
Methods:
Classification with CNA data

1. Choose the most suitable classification algorithm

2. Choose the best set of features

3. Evaluate the model

• We tried different classification algorithms
  • The ones giving the best performances were:
    ▪ **KNN**, when using features from GISTIC2.0
    ▪ **SVM**, in all the other cases
Methods:
Classification with CNA data

1. Choose the most suitable classification algorithm

2. Choose the best set of features
   - Take the set of features obtained with the different features selection methods
   - Compute for each of them precision, recall, accuracy and AUC through a 10-fold cross validation
   - Select the features giving the best performances

3. Evaluate the model
Methods: Classification with CNA data

1. Choose the most suitable classification algorithm
2. Choose the best set of features
3. Evaluate the model

- We did not achieve good results
- The best performances obtained for Resistant vs Sensitive were:
  - Average precision: $0.51 \pm 0.10$
  - Average recall: $0.61 \pm 0.19$
  - Average accuracy: $0.68 \pm 0.07$
  - Average AUC: $0.72 \pm 0.11$
Methods: Survival Regression

- **Progression Free Survival (PFS):**
  the interval from the date of surgery to the date of progression, date of recurrence, or date of last known contact.
Methods:
Survival Regression

- **Progression Free Survival (PFS):** the interval from the date of surgery to the date of progression, date of recurrence, or date of last known contact

- Censored data: patients who did not have the relapse up to the last contact

\[ \sim 68\% \]
\[ \sim 8\% \]
Methods: Survival Regression

- **Progression Free Survival (PFS):** the interval from the date of surgery to the date of progression, date of recurrence, or date of last known contact

- Censored data: patients who did not have the relapse up to the last contact

- How to predict PFS?
  - Cox Regression Model
Methods:
Survival Regression

- **Progression Free Survival (PFS):** the interval from the date of surgery to the date of progression, date of recurrence, or date of last known contact

- Censored data: patients who did not have the relapse up to the last contact

- How to predict PFS?
  - **Cox Regression Model**

- What features did we use?
  - The ones obtained through the **permutation test**
Results: Survival Regression

- We were not able to correctly predict the PFS times of the patients
- The best *concordance index* we got was equal to 0.58
Second approach to solve the problem:

Use four types of genomic data
Steps performed

1. Feature selection for the other three types of data

2. Classification
Feature Selection:
Gene expression, miRNA and DNA methylation data

- Compute the p-values, for the different genomic elements, using **Mann-Whitney** test (for each binary comparison)
Feature Selection:
Gene expression, miRNA and DNA methylation data

- Compute the p-values, for the different genomic elements, using **Mann-Whitney** test (for each binary comparison)

- Try different thresholds for the p-values: 0.05, 0.005, 0.0005
Feature Selection:  
Gene expression, miRNA and DNA methylation data

- Compute the p-values, for the different genomic elements, using **Mann-Whitney** test (for each binary comparison)

- Try different thresholds for the p-values: 0.05, 0.005, 0.0005

- Try different correction for multiple testing:
  - **Bonferroni correction:**
    
    \[ p_{value_{corrected}} = p_{value} \cdot n_{tests} \]
  
  - **Benjamini-Hochberg correction:**
    
    \[ p_{value_{corrected}} = p_{values} \cdot \frac{n_{tests}}{\text{ranking}} \]

- Standard version: \( n_{tests} = \text{total number of tests} \)
- Mild version: \( n_{tests} = \text{number of patients of the two classes} \)
Steps performed

1. Feature selection for the other three types of data

2. Classification
Methods:
Classification with four types of genomic data

Select the best features for each type of genomic data
Methods:
Classification with four types of genomic data

Select the best features for each type of genomic data

Merge the four datasets

Normalize

Classify using SVM
Best computational results:

ROC curves for Resistant vs Sensitive

Merging of all genomic data

- CNA
- Gene expression
- miRNA
- Methylation
Best computational results:  
For Resistant vs Sensitive

<table>
<thead>
<tr>
<th>Type of data</th>
<th>N features</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNA</td>
<td>225</td>
<td>0.51 ± 0.10</td>
<td>0.61 ± 0.19</td>
<td>0.68 ± 0.07</td>
<td>0.72 ± 0.11</td>
</tr>
<tr>
<td>Gene expression</td>
<td>20</td>
<td>0.71 ± 0.20</td>
<td>0.37 ± 0.10</td>
<td>0.77 ± 0.10</td>
<td>0.79 ± 0.11</td>
</tr>
<tr>
<td>miRNA</td>
<td>11</td>
<td>0.77 ± 0.30</td>
<td>0.37 ± 0.20</td>
<td>0.75 ± 0.10</td>
<td>0.72 ± 0.15</td>
</tr>
<tr>
<td>Methylation</td>
<td>65</td>
<td>0.79 ± 0.30</td>
<td>0.35 ± 0.10</td>
<td>0.78 ± 0.10</td>
<td>0.78 ± 0.09</td>
</tr>
<tr>
<td>Merge</td>
<td>311</td>
<td>0.68 ± 0.18</td>
<td>0.74 ± 0.11</td>
<td>0.80 ± 0.10</td>
<td>0.82 ± 0.09</td>
</tr>
</tbody>
</table>

- A single genomic data is not enough to distinguish the two main classes: *resistant* and *sensitive*

- Four genomic signals together allow to achieve good performances ⇒ the recall is significantly better
Best computational results:
ROC curves for the other binary comparisons

All genomic data

Resistant vs Sensitive Long

Resistant vs Sensitive Short
Best computational results:

Consideration

- The method is satisfying: it allows to achieve good results for all the performance measures, i.e., precision, recall, accuracy and AUC of the ROC curves

- **Innovation**: use four different genomic data-types and be able to classify the patients with good performances
Biological Results:
Relevant features for Resistant vs Sensitive

From 137 CNA amplification regions

183 genes
Protein coding

24 related to HGS-OC
Biological Results:
Relevant features for Resistant vs Sensitive

From 137 CNA amplification regions

183 genes
Protein coding

24 related to HGS-OC

8 for the biology

8 for drug resistance

8 for the etiology
Biological Results:
Relevant features for Resistant vs Sensitive

From 137 CNA amplification regions

183 genes
Protein coding

24 related to HGS-OC

8 for the biology

8 for drug resistance

8 for the etiology

2 activators of the Notch pathway (DLL1 e CTBP2)

5 related to resistance to platinum-based chemotherapy (CA9, ELAV1, HTRA1, RNASET2, URI1)

1 related to the response to chemotherapy (BID)
Biological Results:
Relevant features for Resistant vs Sensitive

• We further analyzed the 8 genes related to drug-resistance
Biological Results:

Relevant features for Resistant vs Sensitive

• We further analyzed the 8 genes related to drug-resistance

• For 5 (DLL1, CTBP2, BID, CA9, HtrA1) of them, resistant and sensitive have:
  
  ▪ **Different** CNA values distribution (at the time of diagnosis)
  
  ▪ **Not different** Gene expression distribution (at the time of diagnosis)

  ▪ **Different** Gene expression distribution (after therapy)
Biological Results:
Relevant features for Resistant vs Sensitive

• We further analyzed the 8 genes related to drug-resistance

• For 5 (DLL1, CTBP2, BID, CA9, HtrA1) of them, resistant and sensitive have:
  
  ▪ **Different** CNA values distribution (at the time of diagnosis)
  
  ▪ **Not different** Gene expression distribution (at the time of diagnosis)
  
  ▪ **Different** Gene expression distribution (after therapy)

N.B.: The last information is known from literature and need experimental confirmation
Exploiting computational methods we identified a **molecular signature** that allows to:
Conclusions:
Main contributions

Exploiting computational methods we identified a **molecular signature** that allows to:

- Predict the response to therapy (resistant / sensitive)
- Understand the cause of chemoresistance
Conclusions:
Main contributions

Exploiting computational methods we identified a **molecular signature** that allows to:

- Predict the response to therapy (resistant / sensitive)
- Understand the cause of chemo-resistance

The goal of the project is accomplished
Conclusions:
Main contributions

• We built a classifier with satisfying performances integrating four types of genomic data
Conclusions:
Main contributions

• We built a classifier with satisfying performances integrating four types of genomic data

• With our model, we discovered 137 CNA regions of amplification (less than 1% of the genome) as discriminatory for the two main classes, *resistant* and *sensitive*
Conclusions:
Main contributions

- We built a classifier with satisfying performances integrating four types of genomic data.

- With our model, we discovered 137 CNA regions of amplification (less than 1% of the genome) as discriminatory for the two main classes, *resistant* and *sensitive*.

- These regions contain 24 genes related to HGS-OC, 8 of which are directly connected to chemoresistance.
Conclusions: Main contributions

- We built a classifier with satisfying performances integrating four types of genomic data

- With our model, we discovered 137 CNA regions of amplification (less then 1% of the genome) as discriminatory for the two main classes, resistant and sensitive

- These regions contain 24 genes related to HGS-OC, 8 of which are directly connected to chemoresistance

- Two of the 8 genes belongs to the Notch Signaling Pathway
Conclusions: Future Works

• The results obtained lead to an interesting theory:

*Enhanced drug-resistance could be a direct consequence of the activation of the pathway, due to the alteration of the expression of the identified genes, which in turn occurs as a consequence of their greater replication at diagnosis within these genomic segments.*
Conclusions: Future Works

- The results obtained lead to an interesting theory:

  *Enhanced drug-resistance could be a direct consequence of the activation of the pathway, due to the alteration of the expression of the identified genes, which in turn occurs as a consequence of their greater replication at diagnosis within these genomic segments.*

- Interesting therapeutic options for resistant patients may be developed by targeting the Notch Signaling pathway
Conclusions:

Future Works

- The results obtained lead to an interesting theory:

  *Enhanced drug-resistance could be a direct consequence of the activation of the pathway, due to the alteration of the expression of the identified genes, which in turn occurs as a consequence of their greater replication at diagnosis within these genomic segments.*

- Interesting therapeutic options for resistant patients may be developed by targeting the Notch Signaling pathway

- An efficient test for copy number alterations at diagnosis could be performed using ad-hoc probes on a small set of genes
Thanks for your attention!