Combining pathway identification and breast cancer survival prediction via screening-network methods

A methodological and computational practice

Antonella Iuliano, Ph.D.
a.iuliano@na.iac.cnr.it

Convegno GNCS 2018
14 February 2018
Plan of talk

Project purpose

Cancer Survival Analysis
  Survival data
  Cox proportional hazards model

Screening-network Cox model
  Statistical strategies
  Variable screening for Cox’s proportional hazards model
  Variable selection after screening

Cancer Survival Applications
  Datasets information
  Main results

Summary and Future works
Project purpose

- Solve the *high-dimensional* problem with **screening techniques**.
- Solve the *high-correlation* problem with **network regression methods**.
- Good prediction and interpretability of cancer survival data.

Thanks to international projects and consortia (TCGA, EGA, ICGC, etc.) the access to the genome-wide data at single profile and multiple molecular levels has been made available by a variety of **high-throughput technologies**.
Project purpose

Goal. The development of a new multistage computational-statistical model that predicts patient clinical outcomes by screening key survival-related genes by using cancer omics data.

This new approach is based on the two following steps:

A. the REDUCTION of the high-dimensional feature space by using different types of screening:

1. biomedical-driven screening → BMD-screening;
2. data-driven screening → DAD-screening;
3. combination of BMD-and-DAD-screening → BMD+DAD-screening;
Project purpose

B. the **INCLUSION** of the a priori *network information (Gene Functional Maps)* by applying *penalized Cox regression methods*:

- to select a subset of *potential biomarkers or genes* (KEGG pathways and COSMIC investigation) that can improve the classification of high and low risk patients;

- to help doctors to take decisions regarding the patient management and therapeutic (*Personalized medicine*):
  - *early diagnoses, precise prognoses and specific treatments.*

Antonella Iuliano, Ph.D. | a.iuliano@na.iac.cnr.it
Cancer Survival Analysis
Survival data

Output variables:
- \( T_i \): Survival time \( t_i = \min \{ T_i, C_i \} \): observed survival time
- \( C_i \): Censoring time \( \delta_i = I(T_i \leq C_i) \): censoring indicator:
  - \( \delta_i = 1 \) if the survival time is observed
  - \( \delta_i = 0 \) if the survival time is censored

for \( i = 1, \ldots, n \) (patients).

Input variables:
- \( x_i = (x_{i1}, \ldots, x_{ip})^T \): \( p \)-vector for the \( i^{th} \) subject, \( i = 1, \ldots, n \).
  - Gene expression profile of the \( i^{th} \) patient over \( p \) genes.

\( \{(t_i, \delta_i, x_i), i = 1, \ldots, n\} \)
Cancer Survival Analysis
Cox proportional hazards model

**Cox Proportional hazards model** describes the *relationship between the survival times and predictors* (Cox, 1972).

- The **hazard function** at time $t$ for subject $i$ with covariates $x_i$ is
  \[ h(t|x_i) = h_0(t) \exp \left( x_i^T \beta \right), \quad \beta = (\beta_1, \ldots, \beta_p)^T. \]

  - $h_0(t)$ is a **baseline function** that describes the risk for individuals with $x_i = 0$.
  - $\exp \left( x_i^T \beta \right)$ is the **relative risk**, a proportionate increase or reduction in risk, associated with $x_i$.

- The regression coefficients $\beta$ are estimated by maximizing the **Cox’s log-partial likelihood function**:
  \[
  \ell(\beta) = \sum_{i=1}^{n} \delta_i \left\{ x_i^T \beta - \log \left[ \sum_{j \in R(t_i)} \exp(x_j^T \beta) \right] \right\}
  \]
  where $R(t_i)$ is the risk set at time $t_i$, i.e., the set of all patients who still survived prior to time $t_i$. 

Antonella Iuliano, Ph.D. | a.iuliano@na.iac.cnr.it
Screening-network Cox model
Statistical strategies

Limitations

- High-dimensionality \((p \gg n)\)
- High-correlation
- Overfitting

Variable screening methods

1. BioMedical-Driven screening (BMD)
2. DAta-Driven screening (DAD)
3. BioMedical + DAta-driven screening (BMD+DAD)

Network penalty functions

\[ \ell_{\text{pen}}(\beta) = \ell(\beta) + P_\lambda(\beta) \]

where \(P_\lambda(\beta)\) is a penalty function on the coefficient \(\beta\).
(a) A training dataset (T) is chosen to construct different types of screening: **BIO screening**, **DAD screening** and **BIO+DAD-screening**. (b) A subset $\mathcal{I}$ composed by the screened genes is identified. (c) Network-penalized methods are executed on $\mathcal{I}$. (d) BIO, DAD and BIO+DAD markers are selected. (e) A validation dataset (D) is used for prediction. (f) A survival analysis is executed to divide the patients in high-and-low risk group. (g) A pathway analysis is performed. (h) A COSMIC investigation is carried out to detect novel potential biomarkers.

Antonella Iuliano, Ph.D. | a.iuliano@na.iac.cnr.it
Let be $T$ the training set and $D$ the testing set.
Define the set $\{x_j, j \in I\}$ as the subset of the screened variables selected by using $T$ and $d = |\{x_j, j \in I\}|$.

1. **BMD-screening**: If there is enough biomedical knowledge available in literature.

   **Aim**: reduce the feature space from dimension $p$ to a moderate space $d < p$.
   - A subset of active high-risk genes $I$ is identified by using HEFalMp (Human Experimental/Functional Mapper, Huttenhower et al., 2009).
   - Each gene has a $p$-value that indicates the functional relationships between the gene and the disease of interest.
   - Only genes with $p$-value $\leq 0.05$ are selected.
Screening-network Cox model
Variable screening for Cox’s proportional hazards model

2. DAD-screening: If there is no biomedical information available. It relies only on the observed data.

► **Aim:** reduce the feature space from dimension $p$ to a moderate space $d < p$.
► The maximum marginal likelihood estimator (MMLE) $\beta^M_k$, for $k = 1, \ldots, p$, is defined as the maximizer of the log-partial likelihood with a single covariate:

$$
\beta^M_k = \arg \max_{\beta_k} \sum_{i=1}^{n} \delta_i \left\{ \mathbf{x}_{ki}^T \beta_k - \log \left[ \sum_{j \in R(t_i)} \exp(\mathbf{x}_{kj}^T \beta_k) \right] \right\}.
$$

► The set of active variables is

$$
\mathcal{I} = \{1 \leq k \leq p : |\beta^M_k| \geq \delta_n \}
$$

where $\delta_n$ is a threshold value chosen so that we pick the $d$ top ranked covariates.
► The choice of the threshold is data-driven and/or model-based.

Antonella Iuliano, Ph.D. | a.iuliano@na.iac.cnr.it
3. **BMD+DAD-screening**: If there is partial/poor biomedical knowledge available.

- It is useful to obtain novel information from the data not already known about the disease.
- Indeed, there is never a complete information about any disease.

This new type of screening explores the best model that explain the data:

- to capture *new potential markers* that the BMD-screening ignores;
- to improve the survival prediction;
- to give more accurate prognoses, diagnoses and treatments.
Screening-network Cox model
Variable selection after screening

1. By using the subset $\mathcal{I}$ obtained from the screening techniques, the **Cox penalized partial likelihood** on $T$ is computed as

$$
\ell(\beta_\mathcal{I}) = \arg \min_{\beta_\mathcal{I}} \left( \sum_{i=1}^{n} \delta_i \left\{ x_{\mathcal{I},i}^T \beta_\mathcal{I} - \log \left[ \sum_{j \in R(t_i)} \exp(x_{\mathcal{I},j}^T \beta_\mathcal{I}) \right] \right\} \right) + P_\lambda(\beta_\mathcal{I}),
$$

where $P_\lambda(\cdot)$ is the **penalty function** on the regression coefficients $\beta_\mathcal{I}$.

2. **Network-penalized approaches** on the screened genes $\mathcal{I}$ are applied:
   - to *incorporate* a priori biological knowledge into the model;
   - to further *reduce* the dimensional space $d$;
   - to *identify* genes that confer a high risk of breast cancer;
Screening-network Cox model

The Network

- $G = (V, E, W)$ is a network of the relationships among vertices:
  - $V = \{1, ..., p\}$ is the set of vertices (genes);
  - $(i, j) \in E \subset V \times V$ indicates a link between vertices $i$ and $j$;
  - $W = (w_{ij}), (i, j) \in E$ is the set of weights;
  - $d_i = \sum_{(i,j) \in E} w_{ij}$ the degree of vertex $i$.

- Each edge is weighted between $[0,1]$, i.e. the probability that two genes are functionally related.
Screening-network Cox model
Network penalty functions

\[ P_{\lambda, \alpha}(\beta_I) = \lambda [\alpha \|\beta_I\|_p + (1 - \alpha) \Phi(\beta_I)] \]

- \( \lambda > 0 \) (sparsity) and \( \alpha \in (0, 1] \) (network influence) are two regularization parameters.
- The first part is a \( \ell_p \)-norm with \( p \in \{1, 2\} \) which induces sparsity (\( \ell_1 \)) or thresholding (\( \ell_2 \)).
- The second part \( \Phi(\beta_I) = \beta_I^T L \beta_I \) where \( L \) is a Laplacian matrix calculated by using the weight matrix \( W \).
- k-fold cross-validation is performed to select the optimal tuning parameter values (\( \hat{\lambda_I}, \hat{\alpha_I} \)) on training set \( T \).
- Regularized Cox models: AdaLnet (Adaptive Laplacian net, Sun et al., 2014) \( \rightarrow \) Coxnet package in R (Li et al., 2015)
  - \( L_1 \)-norm: \( \|\beta_I\|_1 \)
  - Network:
  \[ \Phi(\beta_I) = \sum_{(i,j) \in E} w_{ij} \left( \text{sgn}(\beta_{I,i})\beta_{I,i}/\sqrt{d_i} - \text{sgn}(\beta_{I,j})\beta_{I,j}/\sqrt{d_j} \right)^2 \]
3. **Validation of prediction model** is performed by computing the *prognostic index (PI)* for each patient $i$ in $D$:

$$PI^D_i = x^D_i \hat{\beta}_i, \quad i = 1, \ldots, n.$$  

- Each patient $i$ in $D$ is assigned into the *high/low-risk* group if its $PI^D_i$ is above (or below) the optimal cutoff $PI^*_T$ selected adaptively on $T$.

4. **High-risk and low-risk groups** are identified and the relative survival curve are shown.

- Use *keplein-Meier* survival estimates to produce survival curve;
- *keplein-Meier* → estimates % survival at each time point;
- *Log-rank test* → assesses if curves differ significantly.
  - The significance level is $p$-value$< 0.05$. 

Antonella Iuliano, Ph.D. | a.iuliano@na.iac.cnr.it
Breast cancer dataset: *METABRIC* downloaded from The European Genome phenome Archive (EGA).

- It contains the profile of about 2000 samples performed on the Illumina HT-12 v3 platform.
- *Long-term Overall Survival data* are available on cBioPortal for Cancer Genomics (http://www.cbioportal.org/):
  - OS-MONTHS ($Q_1 = 60.78$, *Median* = 116.10, $Q_3 = 184.90$);
  - OS-STATUS: Died of Disease=1, Living=0, Died of Other Causes=0.
- *Profile-description:*
  - mRNA expression (microarray);
  - copy-number alterations from DNA copy.
- The samples are partitioned into two subsets:

| Omics data | Training set ($T$) | Testing set ($D$) |
|------------|---------------------|-------------------|
|            | # samples | # genes | # samples | # genes |
| mRNA       | 997      | 19151   | 960       | 19151   |
| mRNA+CNA   | 958      | 18006   | 913       | 18006   |
Cancer Survival Applications

Main results

1. BMD-screening + Network-based penalized Cox methods

i. **BMD-screening**: High-risk breast cancer genes \( b \subset p \) are selected using the tool HEFalMp with \( p\text{-value} < 0.05 \).

ii. **Network-penalized Cox methods**: BMD-genes, i.e., the predictors with regression coefficients \( \hat{\beta}_Ib \neq 0 \), are identified by using the training set (T).

iii. **Validation**: the \( p\text{-value} \) is obtained by using the testing set (D).

| Omics         | Methods | # BMD-genes | \( p\text{-value} \)   | \( \alpha \) | \( \lambda \) | \( b \) |
|---------------|---------|-------------|-----------------------|-------------|-------------|-------|
| mRNA          | Coxnet  | 65          | 8.21e-05              | 0.1         | 0.3414      | 528   |
| mRNA+CNA      | Coxnet  | 63          | 1.11e-05              | 0.1         | 0.3351      | 526   |
Cancer Survival Applications
Survival analysis

BMD-screening + Coxnet

- **Validation test**: Kaplan-Meier survival curves based on high and low risk groups.
## Cancer Survival Applications

### Main results

#### 2. DAD-screening + Network-based penalized Cox methods

1. **DAD-screening**: the $p$-dimensional space is reduced to different large scales $d < p$, for $d = 50, 100, \ldots, 700$.

2. **Network-penalized Cox methods**: for each threshold $d$ the DAD-genes, i.e., the predictors with $\hat{\beta}_I \neq 0$, are selected by using training sets (T).

3. **Validation**: the $p$-values are obtained by using testing sets (D).

#### 3. BMD+DAD-screening + Network-base penalized Cox methods

1. **BMD+DAD-screening**: the BMD-genes are combined with the genes obtained at each threshold $d$, for $d = 50, 100, \ldots, 700$.

2. **Network-penalized Cox methods**: for each threshold $(b, d)$ the BMD+DAD-genes, i.e., the predictors with $\hat{\beta}_{I(b,d)} \neq 0$, are selected by using training sets (T).

3. **Validation**: the $p$-values are obtained by using testing sets (D).
Cancer Survival Applications

Main results

The selected \textit{DAD-genes} and \textit{BMD+DAD-genes} are collected in:

\begin{enumerate}
  \item \textbf{genes-HEFaIMp-high}: genes that match the genes selected by \textit{HEFaIMp} with $p$-value $< 0.05$;
  \item \textbf{genes-HEFaIMp-low}: genes that match the genes selected by \textit{HEFaIMp} $p$-value $> 0.05$;
  \item \textbf{genes-no-HEFaIMp}: genes that are not covered by \textit{HEFaIMp}.
\end{enumerate}

The identification of the \textbf{genes-no-HEFaIMp} is important to discover \textit{new potential biomarkers} not explored by the \textit{BMD-screening}.

\textbf{COSMIC} investigation.
Cancer Survival Applications
Survival Prediction

Coxnet: $p$-value plot

Antonella Iuliano, Ph.D. | a.iuliano@na.iac.cnr.it
3. BMD+DAD-screening + Network-based penalized Cox methods

- \( (b, d) = (528, 700) = 1207 \) (mRNA)
- \( (b, d) = (526, 700) = 1198 \) (mRNA+CNA).

| Omics        | Methods | # BIO+DAD-genes | \( p \)-value  | \( \alpha \) | \( \lambda \) |
|--------------|---------|----------------|----------------|-----------|-----------|
| mRNA         | Coxnet  | 85             | 9.49e-06       | 0.1       | 0.3747    |
| mRNA+CNA     | Coxnet  | 50             | 1.03e-09       | 0.1       | 0.4430    |

| # Genes-HEFaIMp-high | # Genes-HEFaIMp-low | # Genes-no-HEFaIMp |
|----------------------|---------------------|---------------------|
| 50                   | 35                  | 5                   |
| 30                   | 18                  | 2                   |

Antonella Iuliano, Ph.D. | a.iuliano@na.iac.cnr.it
Kaplan-Meier survival curves based on high and low risk groups.
The pathway analysis based on KEGG database is performed on the training set T.

Each node into the network represents a gene.

Each edge between two nodes indicates that the two genes belong to the same KEGG pathway.

- **Orange color** → genes selected by HEFaIMp with \( p \)-value < 0.05.
- **Green color** → genes selected by HEFaIMp with \( p \)-value > 0.05.
- **Violet color** → genes that are not explored by HEFaIMp tool.

Triangular-shaped nodes indicates that the genes are identified in literature as breast-cancer associated genes.

- The *number of papers* is also reported in the triangular nodes.
Cancer Survival Applications
Pathway Analysis

BIO+DAD-screening + Coxnet

KEGG P53 signaling pathway
PI3K/AKT/mTOR pathway
AKT/PKB signaling pathway.

1. KEGG GLYCEROLIPID_METABOLISM
2. KEGG DNA_REPLICATION
3. KEGG MAPK_SIGNALING_PATHWAY
4. KEGG CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
5. KEGG CHEMOKINE_SIGNALING_PATHWAY
6. KEGG PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM
7. KEGG NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION
8. KEGG CELL_CYCLE
9. KEGG OOCYTE_MEIOSIS
10. KEGG P53_SIGNALING_PATHWAY
11. KEGG LYSOSOME
12. KEGG APOPTOSIS
13. KEGG TIGHT_JUNCTION
14. KEGG_GAP_JUNCTION
15. KEGG NEUROTROPHIN_SIGNALING_PATHWAY
16. KEGG PROGESTERONE_MEDIATED_OOCYTE_MATURATION
17. KEGG EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION
18. KEGG PATHWAYS_IN_CANCER
19. KEGG PANCREATIC_CANCER
20. KEGG GLIOMA
21. KEGG PROSTATE_CANCER
22. KEGG MELANOMA
23. KEGG CHRONIC_MYELOID_LEUKEMIA
24. KEGG SMALL_CELL_LUNG_CANCER
25. KEGG NON_SMALL_CELL_LUNG_CANCER

Antonella Iuliano, Ph.D. | a.iuliano@na.iac.cnr.it
Cancer Survival Applications
Gene investigation

Genes-no-HEFaiMp associated to Breast Cancer (BC) survival

| Genes-no-HEFaiMp | Gene description                  | Protein class                | COSMIC Mutation     |
|------------------|-----------------------------------|------------------------------|---------------------|
| ACTL9            | Actin like 9                      | Predicted intracellular      | Missense            |
|                  |                                   | proteins                    |                     |
| LRRN4            | Leucine rich repeat neuronal 4    | Predicted membrane proteins  | Missense            |

| Genes-no-HEFaiMp | Gene description                  | Protein class                | Function BC         |
|------------------|-----------------------------------|------------------------------|---------------------|
| SMIM5            | Small integral membrane protein 5 | Predicted membrane protein   | Suspicious calcifications\(^1\) |
|                  |                                   |                              | Prognostic marker in renal and cervical cancer |
| C2CD4A           | C2 calcium dependent domain       | Predicted intracellular      | Stage-dependent     |
|                  | containing 4A                     | proteins                     | genes\(^2\)         |

1. Shin et al. (2017) Gene expression profiling of calcifications in breast cancer. SCIENTIFIC REPORTS | 7: 11427 | DOI:10.1038/s41598-017-11331-9
2. Yao et al. (2015) Identification of gene-expression signatures and protein markers for breast cancer grading and staging. Plos One.
Summary

- Development and implementatation of a multistage strategy based on screening-network Cox methods for the analysis of multi-omics data.
  - The **BMD-screening** can be performed only when there is enough biological evidence available, although it is still far from being complete.
  - The **DAD-screening** is suggested when there is limited biomedical information available.
  - The **BMD+DAD-screening** is introduced to take advantage of the available biomedical knowledge and also to allow finding novel elements of investigation that can emerge from data.

- Therefore, this new statistical procedure is useful:
  - to prevent the drawbacks of the current methods;
  - to provide a computational and methodological framework in cancer research;
  - to improve prognoses, diagnoses and treatments.
Future works

- Integration of other types of *omics* data.
- Development of an user-friendly interface for the biomedical community.

Antonella Iuliano, Ph.D. | a.iuliano@na.iac.cnr.it
Thank you for attention!

Finanziamento GNCS Giovani Ricercatori 2016