# Applications of Network-based Survival Analysis Methods for Pathways Detection in Cancer

Antonella Iuliano[1,⋆], Annalisa Occhipinti[2,⋆], Claudia Angelini[1], Italia De Feis[1], and Pietro Lió[2]

[1] Istituto per le Applicazioni del Calcolo "Mauro Picone",
[2] Consiglio Nazionale delle Ricerche, via Pietro Castellino 111, 80131 Napoli, Italy
{a.iuliano,c.angelini,i.defeis}@na.iac.cnr.it
[3] Computer Laboratory, University of Cambridge, CB3 0FD, UK
{ao356,pl219}@cam.ac.uk

**Abstract.** Gene expression data from high-throughput assays, such as microarray, are often used to predict cancer survival. Available datasets consist of a small number of samples ($n$ patients) and a large number of genes ($p$ predictors). Therefore, the main challenge is to cope with the high-dimensionality. Moreover, genes are co-regulated and their expression levels are expected to be highly correlated. In order to face these two issues, network based approaches can be applied. In our analysis, we compared the most recent network penalized Cox models for high-dimensional survival data aimed to determine pathway structures and biomarkers involved into cancer progression.

Using these network-based models, we show how to obtain a deeper understanding of the gene-regulatory networks and investigate the gene signatures related to prognosis and survival in different types of tumors. Comparisons are carried out on three real different cancer datasets.

**Keywords:** Cancer, comorbidity, Cox model, high-dimensional data, gene expression data, network analysis, regularization, survival data.

## 1 Introduction

Cancer is a *multi-factorial disease* since it is caused by a combination of genetic and environmental factors working together in a still unknown way. Genetic screening for mutations associated with multi-factorial diseases cannot predict exactly whether a patient is going to develop a disease, but only the risk to have the disease. Hence, a woman inheriting an alteration in the BRCA2 gene can develop breast cancer more likely than other women, although she may also remain disease-free. Genetic mutation is only one risk factor among many. Lifestyle, environment and other biological factors are also involved in the study of the disease development. The integration of all this supplementary information is the key point to stress the mechanism of disease progression and identify reliable biomarkers.

⋆ These two authors contributed equally to this work.


C. di Serio et al. (Eds.): CIBB 2014, LNCS 8623, pp. 76–88, 2015.
DOI: 10.1007/978-3-319-24462-4_7

The advancement of recent biotechnology has increased our knowledge about the molecular mechanism involved into cancer progression. However, this biological knowledge is still not fully exploited since the integration of all those different types of data leads to the curse of dimensionality. Indeed, the number of covariates (molecular and clinical information) exceed the number of observations (patients). As a result, many classical statistical methods cannot be applied to analyse this kind of data and new techniques need to be proposed to cope with the high-dimensionality.

In cancer research is also important to study survival analysis, that can be used to investigate microarray gene expression data and evaluate cancer outcomes depending on time intervals. Those intervals start at a survival time and end when an event of interest occurs (a death or a relapse). The exploitation of the relationship between event distributions and gene expression profiles permits to achieve more accurate prognoses or diagnoses. The Cox regression [2] is the most popular method to analyse censored survival data. However, due to high-dimensionality, it cannot be directly applied to obtain the estimated parameters. Therefore, penalized techniques based on lasso type penalties [5,17,18] have been taken into account. Moreover, those methods perform estimation and variable selection by shrinking some parameters to zero. These methods solve the "$p \gg n$" issue but ignore the strong-correlation among variables (i.e. genes). For this reason, the elastic net method (an improved variant of the lasso for high-dimensional data, [13,21]) can be applied to achieve some grouping effects ([3,23]) and to incorporate pathway information of genes. A pathway is given by a group of genes that are involved in the same biological process and have similar biological functions. Those genes are co-regulated and their expression levels are expected to be highly correlated. The pathway structures play a biologically important role to understand the complex process of cancer progression.

The purpose of this paper is (*i*) to describe a systematic approach to compare the most recent methods based on the integration of pathway information into penalized-based Cox methods and (*ii*) to evaluate their performance. We considered three methods. *Net-Cox* [20] explores the co-expression and functional relation among gene expression features using an $L_2$-norm constrain plus a Laplacian penalty. The $L_2$-norm smooths the regression coefficients reducing their variability in the network; the Laplacian take into account the grouping effects. *Ada*ptive *L*aplacian *net* [16] uses an $L_1$-penalty to enforce sparsity of the regression coefficients and a quadratic Laplacian penalty to encorage smoothness between the coefficients of neighboring variables on network. Finally, *Fastcox* method [7] is a new fast algorithm for computing the elastic net penalized Cox model. We compare three different types of cancer by using the penalized regression methods presented before in order to provide an interesting investigation from a biological, medical and computational point of view.

The paper is organized as follows. In Section 2, we introduce the network-based regularized methods for high-dimensional Cox regression analysed in our comparisons. Cross-validation and parameter tuning are discussed in Section 3.

Real data analysis is presented in Section 4, with the main results obtained in the analysis. We conclude with a brief discussion about future works in Section 5.

## 2 Methodology

In this section, we describe the three methods for Cox's proportional hazard model that we used for our analysis. We first review the Cox model and then, we introduce the three regularization methods.

### 2.1 The Cox Model

Prediction of cancer patients survival based on gene expression profiles is an important application of gene expression data analysis. Usually it is difficult to select the most significant genes (i.e. covariates) for prediction, as these may depend on each other in a still unknown way. Because of the large number of expression values, it is easy to find predictors that perform well on the fitted data, but fail in external validation, leading to poor prediction rules.

The problem can be formulated as a prediction problem where the response of interest is a possibly censored survival time and the predictor variables are the gene expression values. The Cox Proportional hazards model [2] is used to describe the relationship between survival times and predictor covariates.

Given a sample of $n$ subjects, let $T_i$ and $C_i$ be the survival time and the censoring time respectively for subject $i = 1, \ldots, n$. Let $t_i = min\{T_i, C_i\}$ be the observed survival time and $\delta_i = I(T_i \leq C_i)$ the censoring indicator, where $I(\cdot)$ is the indicator function (i.e $\delta_i = 1$ if the survival time is observed and $\delta_i = 0$ if the survival time is censored) and $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$ be the $p$-variable vector for the $i$th subject (i.e. the gene expression profile of the $i$th patient over $p$ genes). The survival time $T_i$ and the censoring time $C_i$ are assumed to be conditionally independent given $\boldsymbol{X}_i$. Furthermore, the censoring mechanism is assumed to be non-informative. The observed data can be represented by the triplets $\{(t_i, \delta_i, \boldsymbol{X}_i), i = 1, ..., n\}$. The Cox regression model assumes that the hazard function $h(t|\boldsymbol{X}_i)$, which means the risk of death at time $t$ for the $i$th patient with gene expression profile $\boldsymbol{X}_i$ , can be written as

$$h(t|\boldsymbol{X}_i) = h_0(t)exp\left(\sum_{i=1}^{p} \boldsymbol{X}'_i\boldsymbol{\beta}\right) = h_0(t)exp(\boldsymbol{X}'\boldsymbol{\beta})$$

where $h_0(t)$ is the baseline hazard and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the column vector of the regression parameters.

Since the number of predictors $p$ (genes) is much greater than the number of observations $n$ (patients), the Cox model cannot be applied directly and a regularization approach needs to be used to select important variables from a large pool of candidates. For instance, a Lasso penalty ([17,18]), can be used to remove the not significant predictors by shrinking their regression coefficients exactly to zero. The lasso type approach solves the high dimensionality issue but

don't take into account the functional relationships among genes. For this reason, in the last years, network-based regularization methods have been introduced in order to identify the functional relationships between genes and overcome the gap between genomic data analysis and biological mechanisms. By using these network-based models, it is possible to obtain a deeper understanding of the gene-regulatory networks and investigate the gene signatures related to the cancer survival time. In this context, the regression coefficients are estimated by maximizing the penalized Cox's log-partial likelihood function

$$l_{pen}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{X}_i'\boldsymbol{\beta} - log \left[ \sum_{j \in R(t_i)} exp(\boldsymbol{X}_j'\boldsymbol{\beta}) \right] \right\} - P_\lambda(\boldsymbol{\beta}), \tag{1}$$

where $t_i$ is the survival time (observed or censored) for the $i$th patient, $R(t_i)$ is the risk set at time $t_i$ (i.e., the set of all patients who still survived prior to time $t_i$) and $P_\lambda(\boldsymbol{\beta})$ is a network-constrained penalty function on the coefficients $\boldsymbol{\beta}$.

### 2.2 Network-regularized Cox Regression Models

We assume that the relationships among the covariates (genes) are specified by a network $G = (V, E, W)$ (weighted and undirected graph). Here $V = \{1, \ldots, p\}$ is the set of vertices (genes/covariates); an element $(i, j)$ in the edge set $E \subset V \times V$ indicates a link between vertices $i$ and $j$; $W = (w_{ij})$, $(i, j) \in E$ is the set of weights associated with the edges. Each edge in the network is weighted between [0,1] and indicates the functional relation between two genes [6]. For instance, in a gene regulatory network built from data, the weight may indicate the probability that two genes are functionally connected.

*Net-Cox* [20] integrates gene network information into the Cox's proportional hazard model by the following

$$P_{\lambda,\alpha}(\boldsymbol{\beta}) = \lambda \left[ \alpha \left\| \boldsymbol{\beta} \right\|_2^2 + (1 - \alpha) \, \Phi(\boldsymbol{\beta}) \right], \tag{2}$$

where $\lambda > 0$ and $\alpha \in (0, 1]$ are two regularization parameters in the network constraint and

$$\Phi(\boldsymbol{\beta}) = \sum_{(i,j) \in E} w_{i,j} (\beta_i - \beta_j)^2. \tag{3}$$

The penalty (2) consists of two terms: the first one is an $L_2$-norm of $\beta$ that regularizes the uncertainty in the network constraint; the second term is a network Laplacian penalty $\Phi(\boldsymbol{\beta}) = \boldsymbol{\beta}'[(1-\alpha)\boldsymbol{L}+\alpha\boldsymbol{I}]\boldsymbol{\beta}$ that encourages smoothness among correlated gene in the network and encode prior knowledge from a network. In the penalty, $\boldsymbol{L}$ is a positive semi-definite matrix derived from network information and $\boldsymbol{I}$ is an identity matrix. Given a normalized graph weight matrix $\boldsymbol{W}$, by using Eq.(3), *Net-Cox* assumes that co-expressed (related) genes should be assigned similar coefficients by defining the following cost term over the coefficients $\Phi(\boldsymbol{\beta}) = \boldsymbol{\beta}'(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{L}\boldsymbol{\beta}$. More precisely, for any pair of genes connected by an high weight edge and with a large difference between their coefficients, the objective function will result in a significant cost in the network.

*AdaLnet* [16] (*Ada*ptive *L*aplacian *net*) is a modified version of a network-constrained regularization procedure for fitting linear-regression models and for variable selction [10,11] where the predictors are genomic data with graphical structures. *AdaLnet* is based on prior gene regulatory network information, represented by an undirected graph for the analysis of gene expression data and survival outcomes. Denoting with $d_i = \sum_{i:(i,j)\in E} w_{ij}$ the degree of vertex $i$, *AdaLnet* defines the normalized Laplacian matrix $\mathbf{L} = (l_{ij})$ of the graph G by

$$l_{i,j} = \begin{cases} 1, & \text{if } i = j \text{ and } d_i \neq 0, \\ -w_{ij}/\sqrt{d_i d_j}, & \text{if} (i,j) \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Note that $\mathbf{L}$ is positive semi definite. The network-constrained penalty in Eq. (1) is given by

$$P_{\lambda,\alpha}(\boldsymbol{\beta}) = \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha)\, \Psi(\boldsymbol{\beta})\right], \tag{5}$$

with

$$\Psi(\boldsymbol{\beta}) = \sum_{(i,j)\in E} w_{i,j} \left( sgn(\tilde{\beta}_i)\beta_i/\sqrt{d_i} - sgn(\tilde{\beta}_j)\beta_j/\sqrt{d_j} \right)^2. \tag{6}$$

Equation (5) is composed by two penalties. The first one is an $L_1$-penalty that induces a sparse solution, the second one is a quadratic Laplacian penalty $\Psi(\boldsymbol{\beta}) = \boldsymbol{\beta}' \tilde{\boldsymbol{L}} \boldsymbol{\beta}$ that imposes smoothness of the parameters $\beta$ between neighboring vertices in the network. Note that $\tilde{\boldsymbol{L}} = \boldsymbol{S}' \boldsymbol{L} \boldsymbol{S}$ with $\boldsymbol{S} = \text{diag}(sgn(\tilde{\beta}_1), \ldots, sgn(\tilde{\beta}_p))$ and $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)$ is obtained from a preliminary regression analysis. The scaling of the coefficients $\boldsymbol{\beta}$ respect to the degree allows the genes with more connections (i.e., the hub genes) to have larger coefficients. Hence, small changes of expression levels of these genes can lead to large changes in the response.

An advantage of using penalty (5) consists in representing the case when two neighboring variables have opposite regression coefficient signs, which is reasonable in network-based analysis of gene expression data. Indeed, when a transcription factor (TF) positively regulate gene $i$ and negatively regulate gene $j$ in a certain pathway, the corresponding coefficients will result with opposite sign.

Finally, *Fastcox* [7] computes the solution paths of the elastic net penalized Cox's proportional hazards model. In this method the penalty function in Eq. (1) is given by

$$P_{\lambda,\alpha}(\boldsymbol{\beta}) = \lambda \left[ \alpha \boldsymbol{w} \|\boldsymbol{\beta}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}\|_2^2 \right],$$

where the non-negative weights $\boldsymbol{w}$ allows more flexible estimation.

## 3 Tuning Parameters by Cross-validation

All above described methods require to set two hyper-parameters: $\lambda$ and $\alpha$ controlling the sparsity and the network influence, respectively. To determine the optimal tuning parameters $\lambda$ and $\alpha$ to use in our study, we performed five-fold

cross-validation following the procedure proposed by [22]. In the cross-validation, four folds of data are used to build a model for validation on the fifth fold, cycling through each of the five folds in turn. Then, the ($\lambda$,$\alpha$) pair that minimizes the cross-validation log-partial likelihood (CVPL) are chosen as the optimal parameters. CVPL is defined as

$$CVPL(\lambda, \alpha) = -\frac{1}{n}\sum_{k=1}^{K}\{\ell(\hat{\beta}^{(-k)}(\lambda, \alpha)) - \ell^{(-k)}(\hat{\beta}^{(-k)}(\lambda, \alpha))\}, \tag{7}$$

where $\hat{\beta}^{(-k)}(\cdot)$ is the estimate obtained from excluding the $k$th part of the data with a given pair of $(\lambda, \alpha)$, $\ell(\cdot)$ is the Cox log-partial likelihood on all the sample and $\ell^{(-k)}(\cdot)$ is the log-partial likelihood when the $k$th fold is left out.

## 4 Real Case Studies

In this section we describe the performances of the methods presented in Section 2 on three different types of cancer. In the following we first describe the datasets, then the results.

### 4.1 Datasets

We applied the three methods on three datasets containing large-scale microarray gene expression measurements from different type of cancer together with their (possible censored) survival informations (times and status). In particular, we used gene expression datasets downloaded from Gene Expression Omnibus as raw .CEL files. All the three datasets were generated by Affymetrix U133A. The raw files were processed and normalized individually by RMA package available in Bioconductor [4].

We consider the human gene functional linkage network [6] constructed by a regularized Bayesian integration system [6]. Such network contains maps of functional activity and interaction networks in over 200 areas of human cellular biology with information from 30.000 genome-scale experiments. The functional linkage network summarizes information from a variety of biologically informative perspectives: prediction of protein function and functional modules, crosstalk among biological processes, and association of novel genes and pathways with known genetic disorders [6]. The edges of the network are weighted between $[0, 1]$ and express the functional relation between two genes. Thus, the functional linkage network plays an important role in our tests since it includes more information than Human protein-protein interaction, frequently used as the network prior knowledge. It is clear that taking into account such biological knowledge helps in identifying significant genes that are functionally related in order to obtain important results biologically interpretable.

We use HEFaIMp [6] tool to identify the edge's weight of between two genes on the network. After merging probes by gene symbols and removing probes with no gene symbol, we use KEGG pathways [8,9] in order to obtain a network

consisting of a fixed number of unique genes derived from a large pool of probes and overlapped with the functional linkage network. The three datasets analysed are the following:

1. **Breast Cancer Microarray Data.** The first dataset is from Nagalla et al. [12] (accession number: GSE45255) and consist of $p = 2431$ gene expression measurements from $n = 93$ patients with breast cancer.
2. **Lung Cancer Microarray Data.** The second dataset is from Chen et al. [1] (accession number: GSE37745) and contains $p = 2259$ gene expression measurements from $n = 100$ patients with lung cancer.
3. **Ovarian Cancer Microarray Data.** The third dataset is from Zhang et al.[20] (accession number: GSE26712) and contains gene expression measurements from $N = 153$ patients with ovarian cancer. We use a list of $p = 2372$ genes.

### 4.2 Model Evaluation Criteria

In order to evaluate the three methods we first divided each dataset randomly into two parts: (*i*) *training set* consisting of about 2/3 of the patients used for estimation; (*ii*) *testing set* consisting of about 1/3 of the patients used for evaluate and test the prediction capability of the models. We denoted the parameter estimate from the training data for a given method by $\hat{\beta}_{train}$. This estimate is computed as described in Section 3 by using five-fold cross-validation to select the optimal tuning parameter values $(\hat{\lambda}_{train}, \hat{\alpha}_{train})$, and then by fitting the corresponding penalized function $P_{\hat{\lambda}_{train},\hat{\alpha}_{train}}(\hat{\beta}_{train})$ on the training set. In particular, we first set $\alpha$ to a sufficiently fine grid of values on $[0, 1]$. For each fixed $\alpha$, $\lambda$ was chosen from $\{1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 1e{-}1, 1\}$ for *Net-Cox*, while we set $\lambda$ to a decreasing sequence of values $\lambda_{max}$ to $\lambda_{min}$ automatically choosen by *AdaLnet* and *Fastcox*. Note that, when $\alpha = 1$ all the three methods listed in Section 2.2 ignore the network information. The results are given in Table 1. Interestingly, the optimal $\alpha$ is often 0.1 and 0.5, indicating the optimal CVPL is a balance of the information from gene expressions and the network. These results highlight that the network information is useful for improving survival analysis.

The estimated $\hat{\beta}_{train}$ is used to calculate the prognostic index (PI) for each patient $i$ in the training set, given by

$$PI_i^{train} = x_i^{'}\hat{\beta}_{train}, \tag{8}$$

**Table 1.** Cross-validation parameters

| | Net-Cox | | AdaLnet | | Fastcox | |
|---|---|---|---|---|---|---|
| Datasets | $\lambda$ | $\alpha$ | $\lambda$ | $\alpha$ | $\lambda$ | $\alpha$ |
| Breast | 0.001 | 0.5 | 0.16 | 0.5 | 0.22 | 0.5 |
| Lung | 0.0001 | 0.1 | 1.90 | 0.1 | 0.60 | 0.5 |
| Ovarian | 0.001 | 0.5 | 11.94 | 0.01 | 0.25 | 0.95 |

where $x_i$ is the vector of gene expression value associated to the $i$th patient. By using the $PI_i^{train}$, it is possible to divide the patients in two subgroups, i.e., *high-risk* and *low-risk* prognosis groups. Thus, the patient $i$ in the training set is assigned to the *high-risk* (or *low-risk*) group if its prognostic index $PI_i^{train}$, Eq. (8), is above (or below) the quantile selected on a grid of given values that spans from 30% to 70%. We select as $PI^*$ the optimal cutoff in terms of $PI^{test}$ corresponding to the lowest $p$-value in a log rank test. Then, we calculate the prognostic index $PI_i^{test}$ by using $\hat{\beta}_{train}$. Each patient $i$ in the testing set is assigned into the *high*-and-*low-risk* groups if its prognostic index $PI_i^{test}$ is above (or below) threshold $PI^*$ chosen as stated before. To evaluate the performance of rule, we applied a log rank test and used the $p$-value as an evaluation criterion (the significance level was set at 5%, i.e., $p < 0.05$). For each datasets, Kaplan-Meier survival curves are drawn and the log-rank test is performed to assess differences between groups. For instance, Fig.1 shows the survival probabilities for these two groups obtained for cancer ovarian patients selected in the testing set by using *AdaLnet* and *Net-Cox*, respectively. More precisely, first we look at survival time in the training set for patients in the top 45% (40%) compared to the lower 55% (60%) testing *Net-Cox* (*AdaLnet*), as described before. We determine the cutoff in terms of $PI^*$. Then, the prognostic $PI_i^{test}$ is calculated and patients are assigned into the *high*-and-*low-risk* groups by comparing with the cutoff obtained from the training set.The log-rank test on the test-set gives a $p$-value of 0.0103 for *AdaLnet* (Fig.1(a)), which means the two groups can be separeted and the selected pathways and genes are significant. In Fig.1(b), even if the log-rank test gives a $p$-value of 0.0189 for *Net-Cox*, we observe that a patient (bottom-right) of the *high-risk* group falls in the *low-risk* group. In particular, we observed that in predicting the survival probabilities, *AdaLnet* and *Net-Cox* discriminate the risk groups better than *Fastcox*.

We performed the same analysis for *high*-and-*low risk* patients in the other two datasets. In the lung cancer dataset, we noticed that even though the Kaplan-Meier survival curves generated by the three methods are well separated, the $p$-value is not significant. On the other hand, in the breast cancer dataset, the survival probabilities for *high*-and-*low risk* patients result not separated.

To further understand the role of the network information in cross-validation and to overcome the drawbacks of investigating only one split, in future studies we will split the dataset using a cross-validation based method for estimating the survival distribution of two or more survival risk groups. All the patients classified as *low-risk* and *high-risk* in every loop of the cross-validation are grouped together and a single Kaplan-Meier curve is computed for each group [14].

### 4.3 Genes and Subnetworks Selected

As mentioned in the beginning, one of the aim of this paper is to find the pathways and the genes selected by the analyzed methods in different types of cancer (breast, ovarian and lung cancer). This study is expected to produce high-quality and well-curated data because of the structure of the different methods. We applied each penalized Cox regression method to the datasets described in

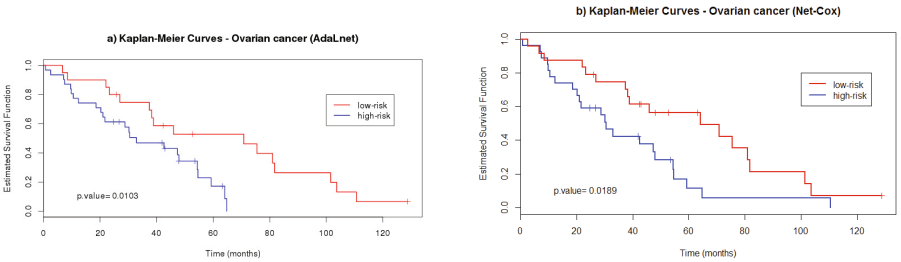



**Fig. 1. Cross-ovarian dataset survival prediction.** The patients are divided in *high-risk* and *low-risk* groups based on the selected pathways and genes. The survival probabilities of these two groups are compared using the log-rank test. a) By using *AdaLnet* the *p*-value means the two groups are well separated and the pathways and genes are significant; b) by using *Net-Cox* (functional linkage network) we note that even if the *p*-value is significant, one patient of the *high-risk* group falls in the *low-risk* group.

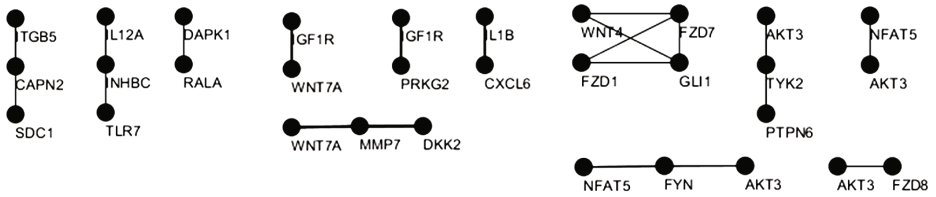
**Table 2.** Number of genes selected by the three methods.

| Datasets | Net-Cox | AdaLnet | Fastcox |
|---|---|---|---|
| Breast | 122 | 38 | 26 |
| Lung | 111 | 61 | 4 |
| Ovarian | 119 | 308 | 12 |

Section 4.1. Here, we present the KEGG networks associated to the non-isolated genes (subnetworks) simultaneously selected by the three methods (Fig.2). The number of genes selected by each method is shown in Table 2. In particular, since *Net-Cox* is a method based on ridge regression, the genes are only shrinkaged and it is necessary to fix a threshold to select the most relevant ones. We fixed the threshold at the 95*th* percentile of the regression coefficients to determine the number of genes showed in Table 2 for *Net-Cox*. We observed that *AdaLnet* identified many more genes and edges on the KEGG network than *Net-Cox* and *Fastcox* for the ovarian cancer dataset, while *Net-Cox* selected many more genes and edges than *AdaLnet* and *Fastcox* for the breast and lung cancer dataset.

In the breast cancer dataset, a subnetwork of the cancer pathway *M12868* was selected, including the *DAPK1* and *RALA* genes strictly involved in cell apoptosis and differentiation (Fig.2(a)). The other two subnetworks are part of the extracellular matrix (ECM) receptor interaction (*M7098*) and focal adhesion (*M7253*) pathways. Both of them are related to important biological processes including cell motility, cell proliferation, cell differentiation, regulation of gene expression and cell survival.

In the lung cancer dataset, *Fastcox* selected only four isolated genes (*CCL22, CSNK1D, HUWE1* and *SLC1A2*). Hence, Fig. 2(b), which reports the not isolated genes, represents the subnetworks selected only by *Net-Cox* and *AdaLnet* in the lung cancer dataset. The gene *IGF1R* appeared in *M12868* which is a




(a) Breast Cancer (b) Lung Cancer (c) Ovarian Cancer

**Fig. 2. KEGG Subnetworks.** The figure shows the subnetworks of the KEGG pathways simultaneously identified by the three algorithms. Only not isolated genes are shown. Figures (a) and (c) represent the subnetworks selected by all the three methods in the breast and ovarian cancer datasets respectively. Since *Fastcox* selected just 4 isolated genes in the lung cancer dataset, (b) shows the subnetworks simultaneously identified only by *Net-Cox* and *AdaLnet*.

well known pathway in cancer. Indeed, *IGF1R* plays an important role in cancer since it is highly overexpressed in most malignant tissues where it functions as an anti-apoptotic agent by enhancing cell survival. Gene *WNT7A*, encodes proteins that are implicated in oncogenesis [28]. The three-node subnetwork *WNT7A–MMP7–DKK2* is part of the WNT signaling pathway (*M19428*) and it is strictly related to the WNT proteins involved in cancer. Finally, the subnetwork *IL1B–CXCL6* is part of the cytokine-cytokine receptor interaction pathway (*M9809*) which is crucial for intercellular regulators and mobilizers of cells engaged in adaptive inflammatory host defenses, cell growth, differentiation, cell death and angiogenesis.

Applying the methods to the ovarian cancer dataset, 5 KEGG subnetworks were selected (Fig.2(c)). The largest connected component is part of the basal cell carcinoma pathway (*M17807*) which includes the *WNT4* gene. This gene is structurally related to genes encoding secreted signaling proteins and it has been implicated in oncogenesis and in several developmental processes, including regulation of cell fate and patterning during embryogenesis. *GLI1* is a gene that encodes a transcription activator involved in oncogene development [26]. The other two genes involved in this subpathway, *FZD1* and *FZD7*, are receptors for *WNT* signaling proteins. The most relevant subnetwork is the one including *AKT3, TYK2* and *PTPN6* genes and it is part of the Jak-STAT signaling pathway (*M17411*). This pathway is one of the core ones suggested by [27] and it is the principal signaling mechanism for a wide array of cytokines and growth factors. The subnetwork *AKT3–FZD8* is part of the cancer pathway *M12868* and both the genes are known to be regulators of cell signaling in response to growth factors. They are involved in a wide variety of biological processes including cell proliferation, differentiation, apoptosis, tumorigenesis. The other two subnetworks are related to the T and B cell receptors signaling pathway which are important components of adaptive immunity.

### 4.4 Implementation and Tools

All comparisons were performed using R and Matlab. *Net-Cox* is a Matlab package available at [24]; *Fastcox* is a R package [25]; *AdaLnet* is a R code and it was sent us upon request. We implemented the cross-validation approach presented in Section 3 for *Net-Cox* and *AdaLnet*. For *Fastcox* we used the function `cv.cocktail()` implemented in the R package [25]. For real data analysis the microarray data were preprocessed using R packages, as described in Section 4.1.

## 5 Discussion and Conclusions

A central problem in genomic research is to identify genes and pathways involved in cancer in order to create a prediction model linking high-dimensional genomic data and clinical outcomes. In cancer genomic, gene expression levels provide important molecular signatures which can be useful to predict the survival of cancer patients. Since gene expression data are characterized by a small set of samples and a large number of variables, the main challenge is to cope with the high-dimensionality and the high-correlation among genes (genes are not independent). To tackle this problem, various network penalized Cox proportional hazards models have been proposed. In this paper, we have compared three methods for the analysis of microarray gene expression data in order to better understand the disease's mechanism. Moreover a grouped/network approach [19] can help us to: (*i*) identify core pathways and significant genes within those pathways related to cancer survival; (*ii*) build a predictive model for survival of future patients based on the identification genetic signatures. Furthermore, this kind of analysis is important to understand how patients' features (i.e., age, gender and coexisting diseases-comorbidity [15]) can influence cancer treatment, detection and outcome.

The Cox model has achieved widespread use in the analysis of time-to-event data with censoring and covariates. The covariates, for example a treatment or other exposure, may change their values over time. It seems natural and appropriate to use the covariate information that varies over time in an appropriate statistical model. One method of doing this is the time-dependent Cox model. The form of a time-dependent covariate is much more complex than in Cox models with fixed (time-independent) covariates. It involves the use of a time dependent function. However, the use of time-dependent covariates offers several opportunities for exploring associations and potentially causal cancer mechanisms. The evolutionary patterns of cancer disease trajectories across different stages and cell heterogeneities provide an effective explanation of the remodulation of disease markers, i.e., the emergence of new disease markers or the change of weight of existing one inside a group of markers induced by changes in phase of the disease or the presence of comorbidity states induced by drugs/therapies or other diseases. We will investigate such problems in future studies.

**Acknowledgements.** This research was partially supported by BioforIU Project and by InterOmics Project. We would like to thank Prof. Hokeun Sun for sharing AdaLnet code.

## References

1. Chen, R., Khatri, P., Mazur, P.K., Polin, M., Zheng, Y., Vaka, D., Hoang, C.D., Shrager, J., Xu, Y., Vicent, S., Butte, A., Sweet-Cordero, E.A.: A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. Cancer Res. 74, 2892–2902 (2014). Published OnlineFirst March 20, doi: 10.1158/0008-5472.CAN-13-2775
2. Cox, D.R.: Regression models and life-tables (with discussion). J. Roy. Stat. Soc. Ser. B 34, 187–220 (1972)
3. Engler, D., Li, Y.: Survival analysis with high-dimentional covariates: An application in microarray studies. Stat. Appl. Genet. Mol. Bio. 8, Article 14 (2009)
4. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5(10), R80 (2004)
5. Gui, J., Li, H.: Penalized Cox regression analysis in the high-dimentional and low-sample size setting, with applications to microarray gene expression data. Bioinformatics 21, 3001–3005 (2005)
6. Huttenhower, C., Haley, E.M., Hibbs, M.A., Dumeaux, V., Barrett, D.R., Coller, H.A., Troyanskaya, O.G.: Exploring the human genome with functional maps. Genome Research 19(6), 1093–1106 (2009)
7. Yang, Y., Zou, H.: A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions. Statistics and Its Interface 6, 167–173 (2013)
8. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30 (2000)
9. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M.: Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199–D205 (2014)
10. Li, C., Li, H.: Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24(9), 1175–1182 (2008)
11. Li, C., Li, H.: Variable selection and regression analysis for graph-structured covariates with an application to genomics. Ann. Appl. Stat. 4, 1498–1516 (2010)
12. Nagalla, S., Chou, J.W., Willingham, M.C., Ruiz, J., Vaughn, J.P., Dubey, P., Lash, T.L., Hamilton-Dutoit, S.J., Bergh, J., Sotiriou, C., Black, M.A., Miller, L.D.: Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis. Genome Biology 14, R34 (2013)
13. Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: Regularization Paths for Coxs Proportional Hazards Model via Coordinate Descent. J. Stat. Soft. 39, 1–13 (2011)
14. Simon, R.M., Subramanian, J., Li, M.C., Menezes, S.: Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimentional data. Briefings in Bioinformatics 12, 203–214 (2011)
15. Sogaard, M., Thomsen, R.W., Bossen, K.S., Sorensen, H.T., Norgaard, M.: The impact of comorbidity on cancer survival: a review. Clinical Epidemiology 5, 3–29 (2013)

16. Sun, H., Lin, W., Feng, R., Li, H.: Network-Regularized high-dimensional cox regression for analysis of genomic data. Statistica Sinica 24, 1433–1459 (2014)
17. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58, 267–288 (1996)
18. Tibshirani, R.: The lasso method for variable selection in the Cox model. J. Roy. Stat. Med. 16, 385–395 (1997)
19. Wu, T.T., Wang, S.: Doubly regularized Cox regression for high-dimensional survival data with group structures. Statistics and Its Interface 6, 175–186 (2013)
20. Zhang, W., Ota, T., Shridhar, V., Chien, J., Wu, B., Kuang, R.: Network-based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment. PLoS Comput. Bio. 9(3), e1002975 (2013). doi:10.1371/journal.pcbi.1002975
21. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. Ser. B 67, 301–320 (2005)
22. van Houwelingen, H.C., Bruinsma, T., Hart, A.A.M., van't Veer, L.J., Wessels, L.F.A.: Cross-validated Cox regression on microarray gene expression data. Stat. Med. 25, 3201–3216 (2006)
23. Wu, Y.: Elastic net for Cox's proportional hazards model with a solution path algorithm. Statist. Sinica 22, 271–294 (2012)
24. `http://compbio.cs.umn.edu/Net-Cox/`
25. `http://code.google.com/p/fastcox/`
26. Liu, C.Z., Yang, J.T., Yoon, J.W., Villavicencio, E., Pfendler, K., Walterhouse, D., Iannaccone, P.: Characterization of the promoter region and genomic organization of GLI, a member of the Sonic hedgehog-Patched signaling pathway. Gene 209(1-2), 1–11 (1998)
27. Jones, S., Zhang, X., Parsons, D.W., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S.M., Fu, B., Lin, M.T., Calhoun, E.S., Kamiyama, M., Walter, K., et al.: Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321(5897), 1801–1806 (2008)
28. Ikegawa, S., Kumano, Y., Okui, K., Fujiwara, T., Takahashi, E., Nakamura, Y.: Isolation, characterization and chromosomal assignment of the human WNT7A gene. Cytogenetic and Genome Research 74(1-2), 149–152 (1996)