Network-based survival analysis methods for pathway detection in cancer

Antonella Iuliano [(1,§)], Annalisa Occhipinti [(2,§)], Haoming Xu[(2)], Claudia Angelini [(1)], Italia De Feis [(1)], Pietro Lió [(2)]

(1) Istituto per le Applicazioni del Calcolo "Mauro Picone"
Consiglio Nazionale delle Ricerche, via Pietro Castellino 111, 80131 Napoli, Italy, a.iuliano@na.iac.cnr.it, {i.defeis, c.angelini}@iac.cnr.it

(2) Computer Laboratory
University of Cambridge, CB3 0FD, UK, {ao356, pl219}@cam.ac.uk, hmingxu@gmail.com

§: Both authors contributed equally to this work.

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

Abstract. We compare three penalized Cox regression methods for high-dimensional survival data in order to identify the pathways involved into cancer occurrence and progression. We analyze each method with three gene expression datasets including breast, lung and ovarian cancer. More precisely, we focus on cancer survival prediction and on top signature genes. The goal of this study is to gain a deeper insight of the benefits and drawbacks of the regression techniques in order to find the pathways involved in a specific type of cancer and identify cancer biomarkers useful for prognosis, diagnosis and treatment.

1 Scientific Background

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 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. The 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 of such analysis in order to stress the mechanism of disease progression and identify reliable biomarkers. 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 generates the high-dimensionality problem. Indeed, gene expression data share a common scenario: 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 analyze this kind of data and new techniques need to be proposed to cope with the high-dimensionality problem.

Cancer research is also based on survival analysis, which is usually applied to study 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). By using this technique and exploiting the relationship between event distributions and gene expression profiles, it is possible to achieve more accurate prognoses or diagnoses. Moreover, including other variables and comorbidity into a survival model can be also useful to identify genes that are significant for the events of interest. Due to the high-dimensionality of gene expressions data, the Cox proportional hazard model [2] is usually used in survival analysis combined with penalties techniques. Specifically, regression methods such as L_2-Cox, L_1- Cox and *elastic net* Cox

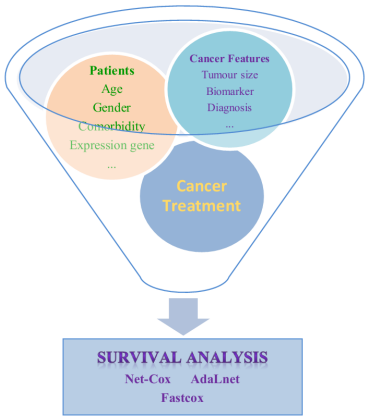


Figure 1: **Cancer survival model.** Three penalized Cox regression methods (Net-Cox, AdaLnet and Fastcox) are compared for survival analysis with high dimensional data including different patients and cancer features.

models (an improved variant of the lasso for high-dimensional data [9]) have been introduced to incorporate the gene pathways information. A pathway is a group of genes that are involved in the same biological process or 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 occurrence and progression.

Our aim is to compare the most recent methods based on the integration of pathway information into network-based survival analysis (as illustrated in Figure 1) in order to identify common pathways between different kind of cancer and overcome the limitations of the existing methodologies for survival analysis with high-dimensional data.

2 Materials and Methods

2.1 *Introduction*

A crucial point in genomic research is to identify a list of genes and pathways involved in cancer and use gene expression data to predict different molecular and clinical outcomes. The Cox model [2] is the most popular survival model used to describe the relationship between survival times and predictor variables. 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 ith subject (i.e. the gene expression profile of the ith patient over p genes). The survival time T_i and the censoring time C_i are assumed to be conditionally independent given X_i. Furthermore, the censoring mechanism is assumed to be non-informative. The observed data can be represented by the triplets $\{(t_i, \delta_i, X_i), i = 1, ..., n\}$. The Cox regression model assumes that the hazard function $h(t|\mathbf{X}_i)$, which means the risk of death at time t for the ith patient with gene expression profile $\mathbf{X}_i$, is given by

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

where $h_0(t)$ is the baseline hazard and $\beta = (\beta_1, \ldots, \beta_p)'$ is the column vector of the regression parameters. In the analysis of microarray gene expressions, if 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 can be used to remove not significant predictors by shrinking their regression coefficients exactly to zero. Due to the high correlation among variables (genes), 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. The regression coefficients are estimated by maximizing the penalized Cox's log-partial likelihood function

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

where t_i is the survival time (observed or censored) for the ith 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 $P_\lambda(\beta)$ is a network-constrained penalty function on the coefficients β.

2.2 *Network-constrained Cox regression*

In this work, we analyze three penalized Cox regression methods for high-dimensional survival data in order to determine pathway structures involved into cancer disease. We assume that the relationships among the covariates (genes) are specified by a network $G = (V, E, W)$ (weighted and undirected graph), where $V = \{1, ..., p\}$ is the set of vertices (covariates), an element (i, j) in the edge set $E \subset V \times V$ indicates a link between vertices i and j, and $W = (w_{ij})$, $(i, j) \in E$ 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 [10]. The Functional Linkage graph plays an important role in our tests since it includes more information than Human protein-protein interaction, frequently used as the network prior knowledge.

The first method used in this analysis defines a network-based Cox regression model, called *Net-Cox* [8]. It integrates gene network information into the Cox's proportional hazard model to explore the co-expression and functional relation among high-dimensional gene expression features. The network penalty function in Eq. (1) is given by

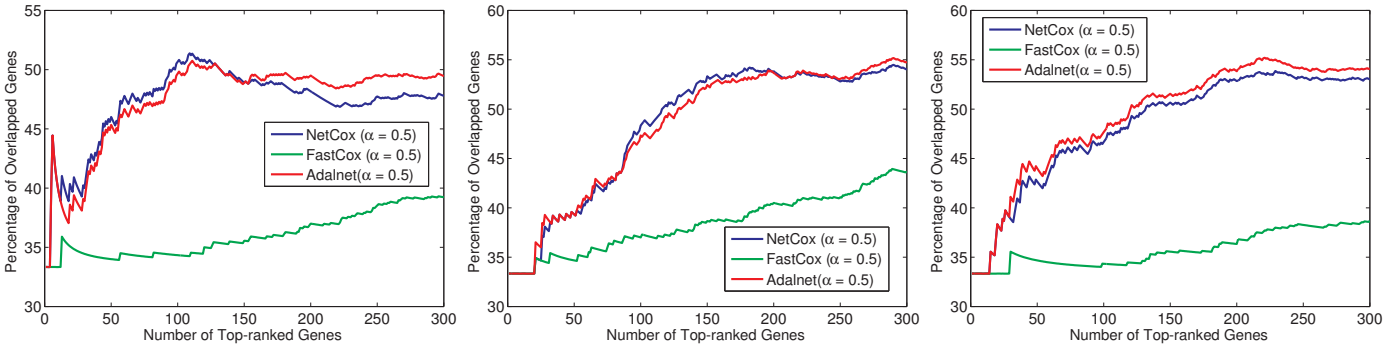
$$P_{\lambda,\alpha}(\beta) = \lambda \left[\alpha|\beta|^2 + (1-\alpha)\Phi(\beta) \right], \tag{2}$$

where λ and $\alpha \in (0, 1]$ are two regularization parameters in the network constraint and $\Phi(\beta) = \sum_{(i,j)\in E} w_{i,j}(\beta_i - \beta_j)^2$. The penalty (2) consists of two parts: the first term is an L_2-norm of β that regularizes the uncertainty in the network constraint; the second term is a quadratic Laplacian penalty $\Phi(\beta) = \beta'\mathbf{L}\beta$ that encourages smoothness among correlated gene in the network, where $\mathbf{L}$ is a positive semi-definite matrix derived from network information. 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.

The second method, called *AdaLnet* [7] (*Ada*ptive *L*aplacian *net*), introduces a network-based regularization method for high-dimensional Cox regression in order to incorporate network information into the analysis of the genomic data. AdaLnet is based on prior gene regulatory network information, represented by an undirected graph, for the analysis of genomic data and survival outcomes. Indicating with $d_i = \sum_{i:(i,j)\in E} w_{ij}$ the degree of vertex i, the network-constrained penalty in Eq. (1) is given by

$$P_{\lambda,\alpha}(\beta) = \lambda \left[\alpha|\beta|_1 + (1-\alpha)\Psi(\beta) \right], \tag{3}$$

with $\Psi(\beta) = \sum_{(i,j)\in E} w_{i,j} \left(sgn(\tilde{\beta}_i)\beta_i/\sqrt{d_i} - sgn(\tilde{\beta}_j)\beta_j/\sqrt{d_j} \right)^2$. The equation (3) is composed by two penalties. The first one is an L_1-penalty that induces a sparse solution,



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

Figure 2: Overlapped Genes. The graphs show the number of common genes among the top 300 genes for each dataset and method tested with the Functional Linkage matrix. The x-axis is the number of selected genes ranked by each method. The y-axis is the percentage of the overlapped genes between the selected ones across the different datasets.

the second one is a quadratic Laplacian penalty $\Psi(\beta) = \beta'\tilde{\mathbf{L}}\beta$ that imposes smoothness of the parameters β between neighbor vertices in the network. Note that $\tilde{\mathbf{L}} = \mathbf{S}'\mathbf{L}\mathbf{S}$ with $\mathbf{S} = \text{diag}(sgn(\tilde{\beta}_1), \ldots, sgn(\tilde{\beta}_p))$ and $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)$ is obtained from a preliminary regression analysis. The matrix $\mathbf{L}$ is always a positive semi-definite matrix derived from network prior knowledge. The scaling by degree of the coefficients β 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.

The third method, *Cocktail algorithm* [3], computes the solution paths of the elastic net penalized Cox's proportional hazards model. In this algorithm the penalty function in Eq. (1) is given by

$$P_{\lambda,\alpha}(\beta) = \lambda \left[\alpha w_j |\beta|_1 + \frac{1}{2}(1-\alpha)\beta_j \right],$$

where the non-negative weights w_j allows more flexible estimation. The Cocktail algorithm is a mixture of three optimization methods: the *Coordinate descent*, the *Majorization-minimization principle* and the *Strong rule*. [3] presents an `R` package, called *Fastcox*, that runs the cocktail algorithm.

2.3 *Datasets*

In this analysis, we considered three case studies involving different types of cancer. In particular, we used gene expression datasets downloaded from Gene Expression Omnibus as raw .CEL files. The raw files were processed and normalized individually by RMA package and library files provided by the Bioconductor project. The details about each dataset are shown in Table 1.

Accession Number	Reference	Cancer Type	Sample Number	Platform
GSE45255	Nagalla et al. [4]	Breast	139	Affymetrix U133A
GSE37745	Chen et al. [1]	Lung	196	Affymetrix U133 Plus 2.0
GSE26712	Zhang et al.[8]	Ovarian	195	Affymetrix U133A

Table 1: Details of the datasets used in the analysis.

3 **Results**

We compare the methods listed in Section 2.2 by analyzing three datasets in order to compare the different performances. Each method takes into account the Functional Linkage networks previously created for each dataset. It is interesting to analyze each dataset separately in order to determine how gene expression data are processed by the each method.

Breast Cancer			Lung Cancer			Ovarian Cancer		
Net-Cox	Fastcox	AdaLnet	Net-Cox	Fastcox	AdaLnet	Net-Cox	Fastcox	AdaLnet
AKR1C3	GJA8	EPHA3	NEFL	FGF8	PDHA2	S100A2	ELL	IREB2
CLCA2	EDN1	RPS6KB2	PTPRR	CSF3R	OTC	EEF1A2	HPR	RAD51
GRIA2	COL5A2	CRY1	MMP7	AOX1	SSH1	WNT4	MYOD1	AFF1
FGG	GALR2	TXNRD1	NEFH	CSF1	ABCD4	MYBL1	PF4	JMJD5
TXNRD1	CDC25C	CAPN2	GNAI1	F9	GRIA4	LCN2	CDH2	ACVR2A
CRY1	KRT81	ST3GAL4	NPY1R	C1S	CRH	ZBTB16	PRKACA	MSH3
COMP	TUBB1	STX7	NOS3	CD70	GABARAP	IGFBP3	MDM4	CDC6
ITGB6	LLGL1	CXXC4	BTC	CYP2C19	SLC2A4	KRT17	MDS1	NRP1
COL11A1	MPZL1	INHBC	TNFRSF11B	CCR7	PRKCE	HOXA9	CITED1	F8
FGA	CACNA1D	AGK	ITGB6	CSF3	NADSYN1	RAB25	E2F2	NCOA1
FADD	CACNG3	LTA	THBD	CTLA4	ACSL6	MAGEA1	MAP2K5	CDH13
DDIT4	MMP14	GADD45B	SPP1	ARHGEF2	FPR1	PLAGL1	ETS1	TYRO3
KRT81	PTGDR	GOT1	GPX2	CHRM4	DDIT3	HOXA5	FGFR1	BCKDHA
F12	APBA1	IDH3A	BCAT1	GABRA1	STX8	TFRC	EDN2	CDC25B
FGB	BCL2	ATP6V0A4	ABCC2	MTNR1B	PIGK	CDC6	FKBP8	HOXA1

Table 2: Top-15 signature genes for each dataset and method tested with the Functional Linkage matrix. The table lists the most significant genes selected by the three algorithms for each dataset.

In particular, we analyzed the percentage of common genes selected by each method from each dataset. This investigation assumes that the genes that are selected by multiple methods are more likely to be true signature genes. Therefore, the higher the overlapping across the methods, the higher the quality in the gene selection.

In Figure 2, we report the percentage of common genes in the rank lists identified by Net-Cox, Fastcox and AdaLnet for each dataset. We plot the percentage of overlapped genes among the first k (up to 300) genes in the gene ranking lists for the breast, lung and ovarian datasets (Figure 2 (a), (b) and (c) respectively). By setting the α parameter equals to 0.5, we observed that both Net-Cox and AdaLnet identify more common genes than Fastcox through all the datasets.

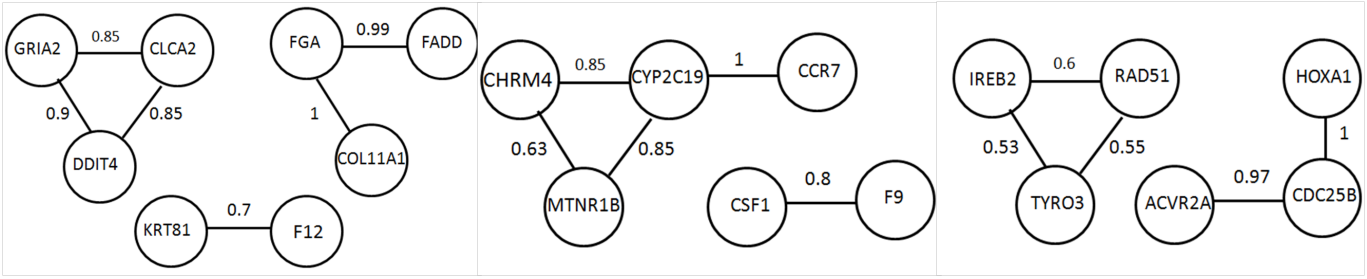
We also analyzed the signature genes identified by the different methods and in Table 2 we present the Top-15 signature genes for each algorithm and dataset. From each list of the Top-15 genes, we extracted the networks among the genes and Figures 3 (a), (b) and (c) report the pathways for three of the lists in Table 2 (Net-Cox with Breast cancer dataset, FastCox with Lung cancer dataset and AdaLnet with Ovarian cancer dataset respectively). In each graph, the edge labels indicate the strength of the interaction between a pair of genes on a scale from 0 to 1. These weights are based on the Functional Linkage interaction which summarizes information from prediction of protein function and functional modules, cross-talk among biological processes and association of genes and pathways with known genetic disorders. All calculations have been carried out in R statistical environment and Matlab.

We also tested each algorithm using the correlation matrix instead of the Functional Linkage one. We calculated the Top-15 lists and the plots of the overlapped genes but we omitted these results due to lack of space.

By using these methods, we are able to select significant pathways and genes among the ones related to breast, ovarian and lung cancer. Moreover, we can develop a predictive model for patient survival based on specific genetic groups. Consequently, the network information is useful to improve the accuracy of survival prediction and to increase the consistency in identifying signature genes across all the three independent datasets.

4 Discussions 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



(a) Breast Cancer - Net-Cox (b) Lung Cancer - FastCox (b) Ovarian Cancer - AdaLnet

Figure 3: Subnetworks identified by (a) Net-Cox algorithm in the Breast Dataset, (b) FastCox algorithm in the Lung Dataset and (c) AdaLnet algorithm in the Ovarian Dataset. The graphs represent the pathways among the Top-15 genes selected by each algorithm (Table 2) and the edges' weights indicate the strength of the interaction between two genes (in a scale from 0 to 1).

gene data are characterized by a small set of samples and a large number of genomic data, the main challenge of gene expression data analysis is the high-dimensionality. To tackle this problem, a variety of penalized Cox proportional hazards models has 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. Furthermore, this kind of analysis is important to understand how patients' features (i.e. age, gender and coexisting diseases-comorbidity) can influence cancer treatment, detection and outcome. Therefore, the future aim will be to highlight the impact of comorbidity on cancer survival analysis [5].

Acknowledgements

This research was partially supported by BioforIU Project and by InterOmics Project.

References

[1] R. Chen, P. Khatri , P.K. Mazur, M. Polin, Y. Zheng , D. Vaka, C.D. Hoang, J. Shrager, Y. Xu, S. Vicent, A. Butte and E.A. Sweet-Cordero."A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma". *Cancer Research*, Published OnlineFirst March 20, 2014.

[2] D. R. Cox. "Regression models and life-tables (with discussion)".*Journal of the Royal Statistical. Society*, Series B 34, pp. 187-220, 1972.

[3] Y. Yang and H. Zou. "A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions". *Statistics and Its Interface*, vol.6, pp. 167173, 2013.

[4] S. Nagalla, J.W. Chou, M.C. Willingham, J. Ruiz, J.P. Vaughn, P. Dubey, T.L. Lash, S.J. Hamilton-Dutoit, J. Bergh, C. Sotiriou, M.A. Black and L.D. Miller. "Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis". *Genome Biology*, 14:R34, 2013.

[5] M. Sogaard, R.W. Thomsen, K.S. Bossen, H. T. Sorensen, M. Norgaard. "The impact of comorbidity on cancer survival: a review". *Clinical Epidemiology* , vol.5, pp3-29, 2013.

[6] N. Simon, J. Friedman, T. Hastie and R. Tibshirani. "Regularization Paths for Coxs Proportional Hazards Model via Coordinate Descent". *Journal of Statistical Software*, vol.39, pp 1-13, 2011.

[7] H. Sun, W. Lin, R. Feng and H. Li. "Network-Regularized high-dimensional cox regression for analysis of genomic data". *Statistica Sinica*, in press, 2014.

[8] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu and R. Kuang. "Network-based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment". *PLoS Comput Biol*, 9(3): e1002975. doi:10.1371/journal.pcbi.1002975, 2013.

[9] H. Zou and T. Hastie. "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical. Society*, Series B 67, pp. 301320, 2005.

[10] Huttenhower, Curtis, Erin M. Haley, Matthew A. Hibbs, Vanessa Dumeaux, Daniel R. Barrett, Hilary A. Coller, and Olga G. Troyanskaya. "Exploring the human genome with functional maps." Genome research 19, no. 6 (2009): 1093-1106.