BMC Systems Biology

PROCEEDINGS Open Access

# Investigating the utility of clinical outcome-guided mutual information network in network-based Cox regression

Hyun-hwan Jeong, So Yeon Kim, Kyubum Wee, Kyung-Ah Sohn*

*From* The Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015)
HsinChu, Taiwan. 21-23 January 2015

## Abstract

**Background:** Network-based approaches have recently gained considerable popularity in high- dimensional regression settings. For example, the Cox regression model is widely used in expression analysis to predict the survival of patients. However, as the number of genes becomes substantially larger than the number of samples, the traditional Cox or $L_2$-regularized Cox models are still prone to noise and produce unreliable estimations of regression coefficients. A recent approach called the network-based Cox (Net-Cox) model attempts to resolve this issue by incorporating prior gene network information into the Cox regression. The Net-Cox model has shown to outperform the models that do not use this network information.

**Results:** In this study, we demonstrate an alternative network construction method for the outcome-guided gene interaction network, and we investigate its utility in survival analysis using Net-Cox regression as compared with conventional networks, such as co-expression or static networks obtained from the existing knowledgebase. Our network edges consist of gene pairs that are significantly associated with the clinical outcome. We measure the strength of this association using mutual information between the gene pair and the clinical outcome. We applied this approach to ovarian cancer patients' data in The Cancer Genome Atlas (TCGA) and compared the predictive performance of the proposed approach with those that use other types of networks.

**Conclusions:** We found that the alternative outcome-guided mutual information network further improved the prediction power of the network-based Cox regression. We expect that a modification of the network regularization term in the Net-Cox model could further improve its prediction power because the properties of our network edges are not optimally reflected in its current form.

## Background

The Cox regression model [1] has been commonly used in survival analysis to detect important features and to predict patient survival. Due to the advance of sequencing technology, the number of genes or features in these analyses is becoming substantially larger than the number of samples. Although several variant models, such as $L_1$ −regularization [2] or $L_2$ −regularization in Hilbert space [3], have been proposed to resolve this issue, those variant models are still prone to noise and over-fitting [4]. Recently, network-based approaches have gained considerable popularity for expression quantitative trait loci (eQTL) studies and clinical outcome predictions in high-dimensional regression settings. These approaches incorporate prior network information of either the features [5] or outcomes [6], or both [7,8]. For survival analysis, Zhang *et al.* [9] recently proposed a variant of the $L_2$ −regularized Cox model called the Net-Cox model, which is a network-regularized Cox regression model. The typical types of networks used in such approaches are either coexpression networks, which are constructed by computing the correlation between every pair of variables, or static networks, such as the protein-protein interaction (PPI) network,

* Correspondence: kasohn@ajou.ac.kr
Department of Information and Computer Engineering, Ajou University, Suwon, S. Korea 443-749

which can be obtained from the existing knowledgebase. In the Net-Cox model, co-expression and functional linkage networks were incorporated in the survival analysis, and the results showed enhanced performance when compared with the conventional methods, which do not use the network information.

A potential limitation of these conventional networks is that the edges only reflect the information of within-features or within-outcomes relations, and do not consider the association between features and outcomes, which may be useful in improving the predictive power. In this study, we show that the outcome-guided mutual information network improves the performance of survival analysis in the Net-Cox regression [9]. We demonstrate the utility of this outcome-guided gene network with the analysis of a TCGA ovarian cancer dataset and compare its performance with those of survival analyses that use other types of networks.

## Methods

### Background on survival analysis using Cox regression

First, we describe the basic formulation of Cox regression [1] for survival analysis using expression and survival data. Given a gene expression profile $X$, which consists of $n$ patients and $p$ genes, the risk of an event at time $t$ for the $i$th patient with gene expression $\mathrm{X}_i = (\mathrm{X}_{\mathrm{i}1}, ..., X_{ip})'$ is defined as $h(t|X_i) = h_0(t)e^{X'_i\beta}$, where $\beta = (\beta_1, ..., \beta_p)'$ is a regression coefficient vector, and $h_0(t)$ denotes the baseline hazard function at $t$. The coefficient $\beta$ and the function $h_0$ are generally unknown and need to be estimated. In a traditional Cox regression model, the estimation of $\beta$ is based on the maximization of the partial log-likelihood, $pl(\beta)$:

$$pl(\beta) = \sum\nolimits_{i=1}^{n} \delta_i \left\{ X'_i\beta - \log\left[ \sum\nolimits_{j \in R(t_i)} e^{X'_j\beta} \right] \right\}.$$

Here, $\delta_i$ is the observed status ($\delta_i = 1$ implies observed; $\delta_i = 0$ implies censored), $t_i$ represents the event time of the $i$th patient, and $R(t_i)$ is the subset of patients who survived to time $t_i$. Once the coefficient vector $\hat{\beta}$ is obtained, a Breslow estimator can estimate the baseline hazard function $h_0$ as follows:

$$\hat{h}_0(t_i) = \frac{1}{\sum_{j \in R(t_i)} e^{X'_j\hat{\beta}}}.$$

### Regularized Cox regression in high-dimensional setting

When $p \gg 1$, the Cox regression model is prone to noise and tends to produce unreliable estimations of regression coefficients. Several solutions that shrink the coefficients have been proposed. A common solution is the $L_2$ –Cox model, which uses the penalized total log-likelihood:

$$l_{pen}(\beta, h_0) = \sum_{i=1}^{n} \left\{ -e^{X'jB} H_0(t_i) + \delta_i \left[ log\left(h_0(t_i)\right) + X'_\beta \right] \right\} - \frac{1}{2}\lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \sum\nolimits_{j=1}^{p} \beta_j^2$ is the regularization term and $\lambda$ controls the amount of shrinkage. Net-Cox regression is an extension of the $L_2$ –Cox model and uses the following penalized log-likelihood:

$$l_{pen}(\beta, h_0) = \sum_{i=1}^{n} \left\{ -e^{X'jB} H_0(t_i) + \delta_i \left[ log\left(h_0(t_i)\right) + X'_\beta \right] \right\} - \frac{1}{2}\lambda\beta'\Gamma\beta,$$

in which $\lambda\beta'\Gamma\beta$ is the penalty term and $\Gamma = (1 - \alpha) L + \alpha I$ conveys prior network information between genes. Here, $L = I - S$, where $I$ is the identity matrix and $S$ is a normalized Laplacian matrix, which comes from the gene network.

The parameter $\alpha \in (0,1]$ controls the contribution of network information to the model. When $\alpha$ is small, more network information is incorporated into the penalty term than when $\alpha$ is large. We note that when $\alpha = 1$, the network-based cox regression model reduces to the $L_2$ –Cox model that does not use any network information.

The typical network used for this type of network-based approach is the coexpression network, in which the edge weights correspond to the correlation between expression vectors of two genes [6]. In [9], a functional linkage network was constructed from the existing knowledgebase. We note that both types of networks do not reflect information about the relation between the features and the outcomes inferred from a given dataset. Therefore, we consider an alternative approach of constructing and utilizing outcome-guided mutual information network in survival analysis. The detailed procedures are explained in the following sections.

### Gene interaction networks associated with clinical outcome

Mutual information is an information-theoretic measure that reveals the dependency or association between two random variables [10] and is defined as follows:

$$I(X;Y) = H(X) + H(Y) - H(X,Y),$$

where $H(X)$ and $H(Y)$ denote the entropies of the respective variables, $X$ and $Y$, and $H(X, Y)$ denotes the joint entropy of $X$ and $Y$. It can be used to detect both linear and nonlinear relations between two random variables [11]. While many previous studies generally used mutual information to detect dependency between two features of the same type, more recent approaches in [12,13] have also extended the mutual information to assess the association between expressions of two genomic features, $X_1$ and $X_2$, and the clinical outcome $Y$ as follows:

$$I(X_1, X_2; Y) = H(X_1, X_2) + H(Y) - H(X_1, X_2, Y).$$

Note that genomic feature values, such as gene expressions or clinical outcome values, are often numeric. To compute the mutual information, we discretize the genomic feature values using a histogram-based technique as in [14]. In the case of survival data, a simple thresholding scheme converts the outcome variable of the survival month to a binary variable. In the case of the TCGA dataset used in this study, we split the patients into short-term living (≤ 3 years) and long-term living (> 3 years). Patients who are reported as *living* and have an overall survival time less than 3 years are filtered out of this study.

Computation of mutual information for every pair of genes and the clinical outcome variable produce a complete network between genes. We can further filter out less significant edges by using the permutation testing scheme proposed in [14]. Through the repeated permutation of clinical outcome labels and the re-computation of mutual information values, a threshold θ is defined as the maximum of the average mutual information values for each gene pair. This θ can be used as a base threshold to remove insignificant edges. However, this threshold still leaves a large number of edges that may not be fully beneficial to the downstream analysis. In this study, another parameter $\sigma$, which amplifies the significance level by a factor of $(1 + \sigma)$, was proposed to create a stricter cut-off. In particular, the filtered mutual information network with significance level $\sigma$ is defined by

$$\mathrm{G}_\sigma = \left\{ \left(g_i, g_j\right) / g_i, g_j \in P \text{ and } I\left(g_i, g_j; Y\right) \geq \theta\left(1 + \sigma\right) \right\},$$

where $P$ is the set of all genes, $Y$ denotes the binary survival status of patients, and $I(g_i, g_j; Y)$ is the mutual information of the gene pair $(g_i, g_j)$ and $Y$. We adopt this thresholding scheme and test the performance of different parameter values. In order to apply the network to the Net-Cox model, we also normalize $\mathrm{G}_\sigma$ and create a normalized Laplacian matrix $\mathrm{S} = \mathrm{R}^{-\frac{1}{2}} G_\sigma C^{-\frac{1}{2}}$, where $\mathrm{R}_{ii} = \sum_j \mathrm{G}_{\sigma_{ij}}$ and $\mathrm{C}_{ii} = \sum_j G'_{\sigma_{ij}}$ both of which are diagonal matrices.

## Dataset

We use genomic profiles and clinical outcome data from patients with *ovarian serous cystadenocarcinoma* in TCGA. Genomic profiles of copy number alterations (CNAs), messenger RNA (mRNA), and methylation (METH) are used in the experiments. We remove genes or patients with missing values and extract genes that are common to all three profiles. As a result, our data matrix consists of expressions and alterations of 10,022 genes and 340 patients for each of the three profiles.

## Evaluation measure

We compare the performance of the Net-Cox model using the proposed mutual information networks with those of models using a gene co-expression and gene functional linkage network as seen in [9]. In order to create a baseline, an analysis of the $L_2$ – Cox model, which does not use any network information, is performed. This is the same as setting α = 1 in the Net-Cox model.

With each run of the Net-Cox analysis, the prognostic indices $PI = X'\hat{\beta}$ are calculated as prediction markers, where $X$ is the test data not used in training, and the regression coefficient $\hat{\beta}$ is estimated from given training data, network information, and the parameters. We evaluate the prediction performance using a time-dependent area under the curve (AUC) [15].

Specifically, time-dependent sensitivity and specificity functions are defined as follows:

$$sensitivity[c, t|f(X)] = Pr\{f(X) > c|\delta(t) = 1\},$$
$$specificity[c, t|f(X)] = Pr\{f(X) > c|\delta(t) = 0\},$$

in which $c$ is the cut-off point, $t$ is the survival time, $f(X)$ are the prognostic indices, $f(X) = \mathrm{X}'\beta$ and $\delta(t)$ is the event indicator at time $t$ [16]. Upon examination of $sensitivity[c, t|f(X)]$ and $1 - specificity[c, t|f(X)]$, we can define $\mathrm{ROC}[t|f(X)]$ as receiving operating characteristic (ROC) curves at any time $t$ and $\mathrm{AUC}[t|f(X)]$ as the area under the ROC curves at any time $t$. The larger $\mathrm{AUC}[t|f(X)]$ is, the better our prediction model performs at time $t$.

In addition, we evaluate the performance, using a validation set and the log-rank test [17], to analyze whether the patients are properly classified: high-risk or low-risk. We rank patients in descending order by their prognostic indices *(PI)* and divide them into a high-risk group containing the top 40% of patients and a low-risk group containing the bottom 40%. We then test the validity of this group assignment by using the log-rank test and the true survival information.

## Cross validation

We select the optimal parameters, λ and α, that respectively control the degree of sparsity and the amount of network constraint in the Net-Cox model using a 5-fold cross validation process. We reserve 20% of all the samples in this study for validation (the validation set) and use the remaining 80% (the training set) for cross validation in order to optimize the parameters and check the cross-validation error. In the case of mutual information networks, we also choose the optimal $\sigma$ using cross-validation. The average time-dependent AUC is used as an evaluation measure for each genomic profile and the different kinds of networks.

We trained each model five times using a 5-fold cross validation process on the training set. At each run, we estimate the regression coefficient $\hat{\beta}$ and the mutual information network (or co-expression network) on four folds of the training data. The estimated coefficient $\hat{\beta}$ and the remaining fold, consisting of 20% of the training data, are then used to calculate the prognostic indices $PI = X'\beta$, which are, in turn, used to rank the patients according to their expected survival time, or month.

To examine how the mutual information network filtered by significance level $\sigma$ contributes to the performance of the Net-Cox model, we experiment with the complete mutual information network and those networks filtered by $\sigma$, where $\sigma$ = [0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3] for each profile. This corresponds to eight mutual information networks on each profile, including the complete network. We vary the parameters $\lambda = [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$ and $\alpha$ = [0.1, 0.3, 0.5, 0.7, 0.9, 1]. As a result, we choose the optimal ($\lambda$, $\alpha$) pair with the best performance, in terms of the time-dependent ROC curves, for each network and each profile.

### Enrichment analysis

In order to assess the biological significance of the results, we performed a gene list enrichment analysis for the 100 genes with the largest regression coefficients for each profile using *ToppGene* http://toppgene.cchmc.org [18]. The enrichment analysis is based on Gene Ontology [19] and Pathway and Disease [20]. The terms that have adjusted $p$-values under 0.05 when using the Benjamini-Hochberg correction [21] are considered biologically significant. We also ran the enrichment test on the genes in the network consisting of the 100 edges with the largest mutual information values for each profile.

## Results and discussion

### Optimal parameter selection using cross validation

We first examined the performance behavior of the mutual information network based Net-Cox model as a function of the parameter $\sigma$ (Figure 1). Note that as the value of $\sigma$ increases, fewer edges remain in the network. In Figure 1, the leftmost $x$-tick label "comp" represents the result obtained using the complete network, and the following $x$-ticks correspond to different $\sigma$ values. Overall, the best performance resulted not from the complete network but from the network filtered by $\sigma$. Therefore, reducing less significant information and using only significant edges seem to improve the performance of the network-regularized regression model. When the CNA and methylation profiles were examined, the best result was obtained using the largest test value of $\sigma$ = 0.3, for which 9,896 and 6,011 edges, respectively, remained, which was about 0.01% of the total number of pairs. In contrast, the mRNA profile preferred a larger network consisting of 568,486 edges (about 1%) with the optimum at $\sigma$ = 0.1.

We also optimized the parameters $\lambda$ and $\alpha$ using a 5-fold cross validation process (Table 1). We found that the optimal $\alpha$ value for the mutual information network was smaller than those of the co-expression and functional linkage networks. This implies that the mutual information network contributed more to the regression

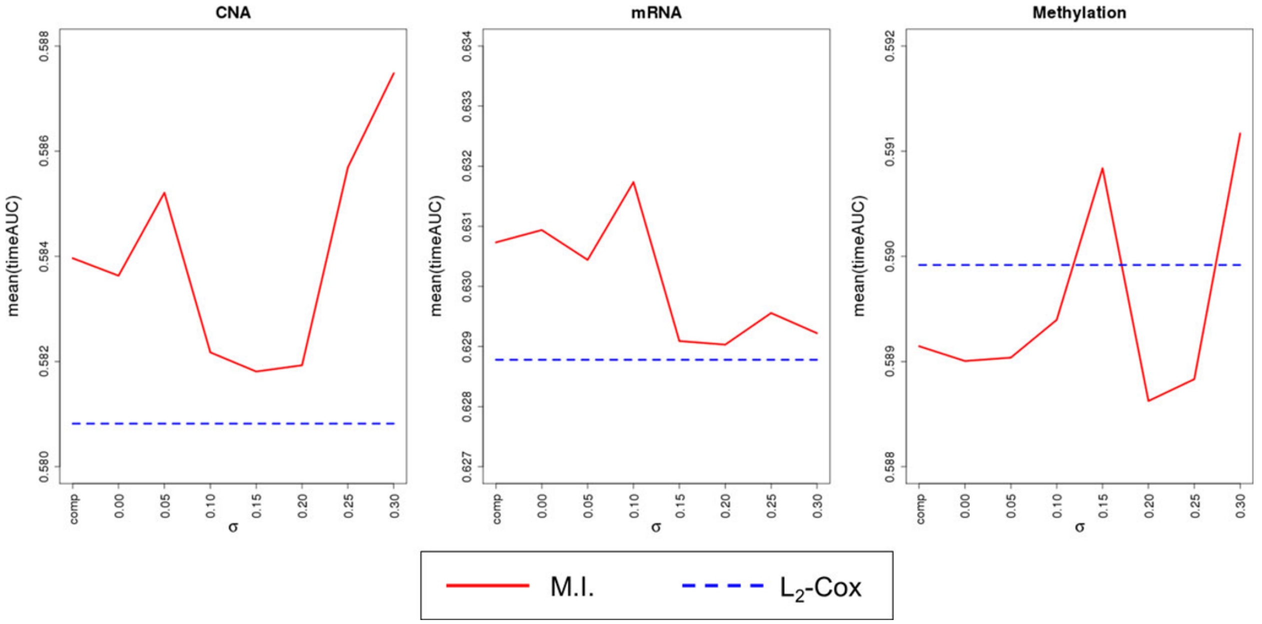

**Figure 1 Prediction accuracy of the mutual information network-based Net-Cox model as a function of the network significance parameter σ**. Average time-dependent AUC using 5-fold cross validation on training data is shown across σ.

**Table 1 Optimal parameters and averaged time-dependent AUCs of each network using 5-fold cross validation on training data**

| Profile | Network | $\sigma$ | $\lambda$ | $\alpha$ | Mean (time AUC) |
|---|---|---|---|---|---|
| **CNA** | Mutual Information | 0.30 | $10^{-3}$ | 0.1 | **0.5875** |
| | Correlation | - | $10^{-3}$ | 0.9 | 0.5817 |
| | Functional Linkage | - | $10^{-3}$ | 0.3 | 0.5786 |
| | $L_2 - \text{Cox}$ | - | $10^{-3}$ | 1.0 | 0.5810 |
| **mRNA** | Mutual Information | 0.10 | $10^{-4}$ | 0.1 | **0.6317** |
| | Correlation | - | $10^{-4}$ | 0.9 | 0.6280 |
| | Functional Linkage | - | $10^{-4}$ | 0.3 | 0.6242 |
| | $L_2 - \text{Cox}$ | - | $10^{-4}$ | 1.0 | 0.6288 |
| **METH** | Mutual Information | 0.30 | $10^{-4}$ | 0.5 | **0.5912** |
| | Correlation | - | $10^{-4}$ | 0.7 | 0.5894 |
| | Functional Linkage | - | $10^{-4}$ | 0.3 | 0.5860 |
| | $L_2 - \text{Cox}$ | - | $10^{-4}$ | 1.0 | 0.5899 |

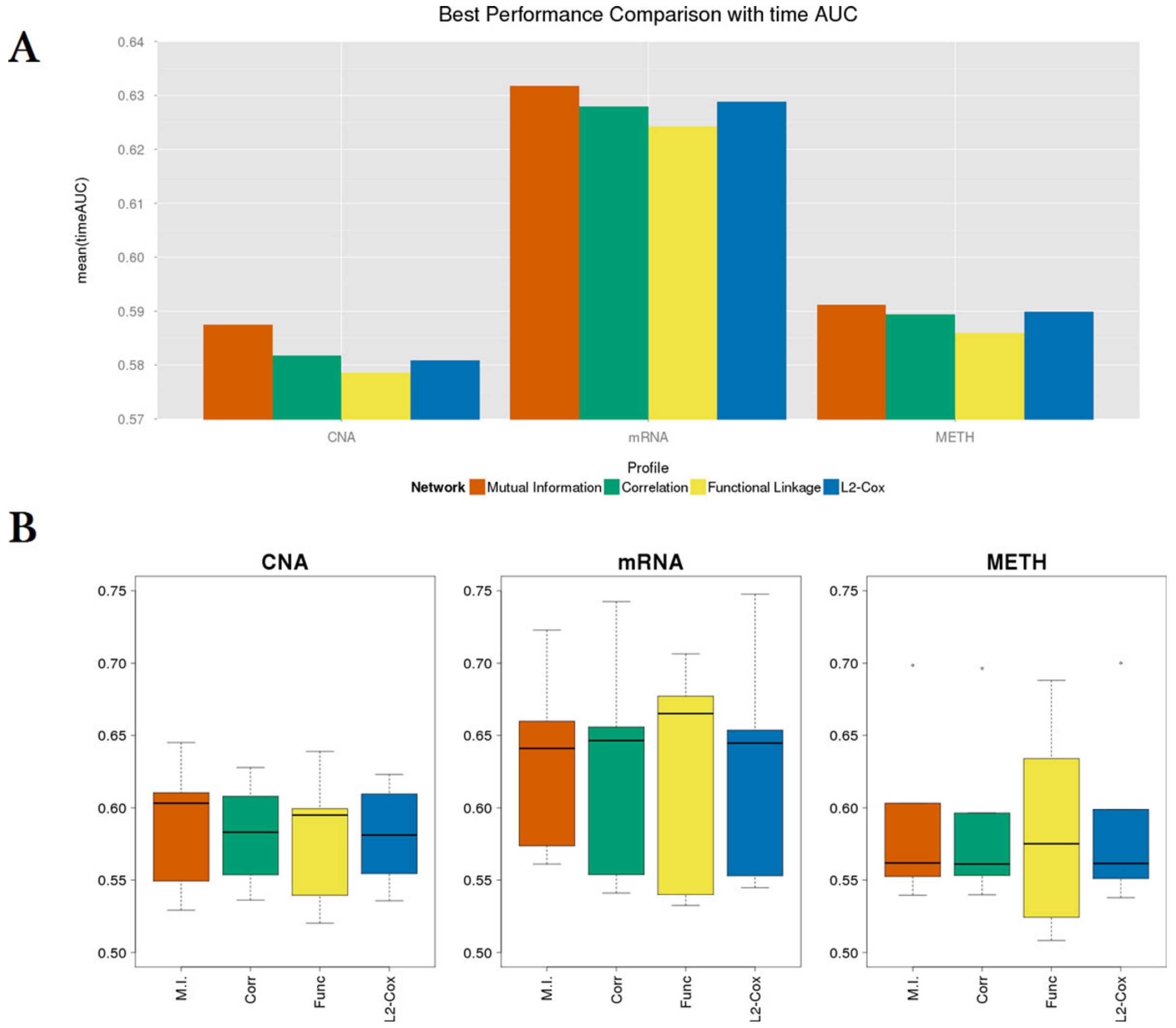

**Figure 2** Performance comparison of different network types in terms of 5-fold cross-validation accuracy.

model than the other approaches. Moreover, the average time-dependent AUC was highest when the mutual information based network was used in all profiles. However, we also note that the median time-dependent AUC was not always highest for the mutual information network, which may imply complementary properties of different types of networks. In Figure 2, the bar plot for the average time-dependent AUC and the boxplot for the distribution of the time-dependent AUC across all the time points and the 5-fold experiments are shown.

### Performance comparison on validation set

After selection of the best parameters for each network, we re-trained each model using the entire training set and then applied the obtained regression model to the holdout validation set. Figure 3 shows a comparison between the resulting time-dependent AUCs. The mutual information based network had the best performance for the mRNA profile and comparable results for the other profiles. Figure 4 shows the Kaplan-Meier survival curves and $p$-values from the log-rank tests with respect to the patient group assignment for each approach. Examining the CNA and mRNA profiles, every method revealed significant results ($p$-value < 0.05) with the log-rank test, except for the functional linkage network on the CNA profile. However, all the methods showed insignificant results for the methylation profile. It appears that the interaction effect in the

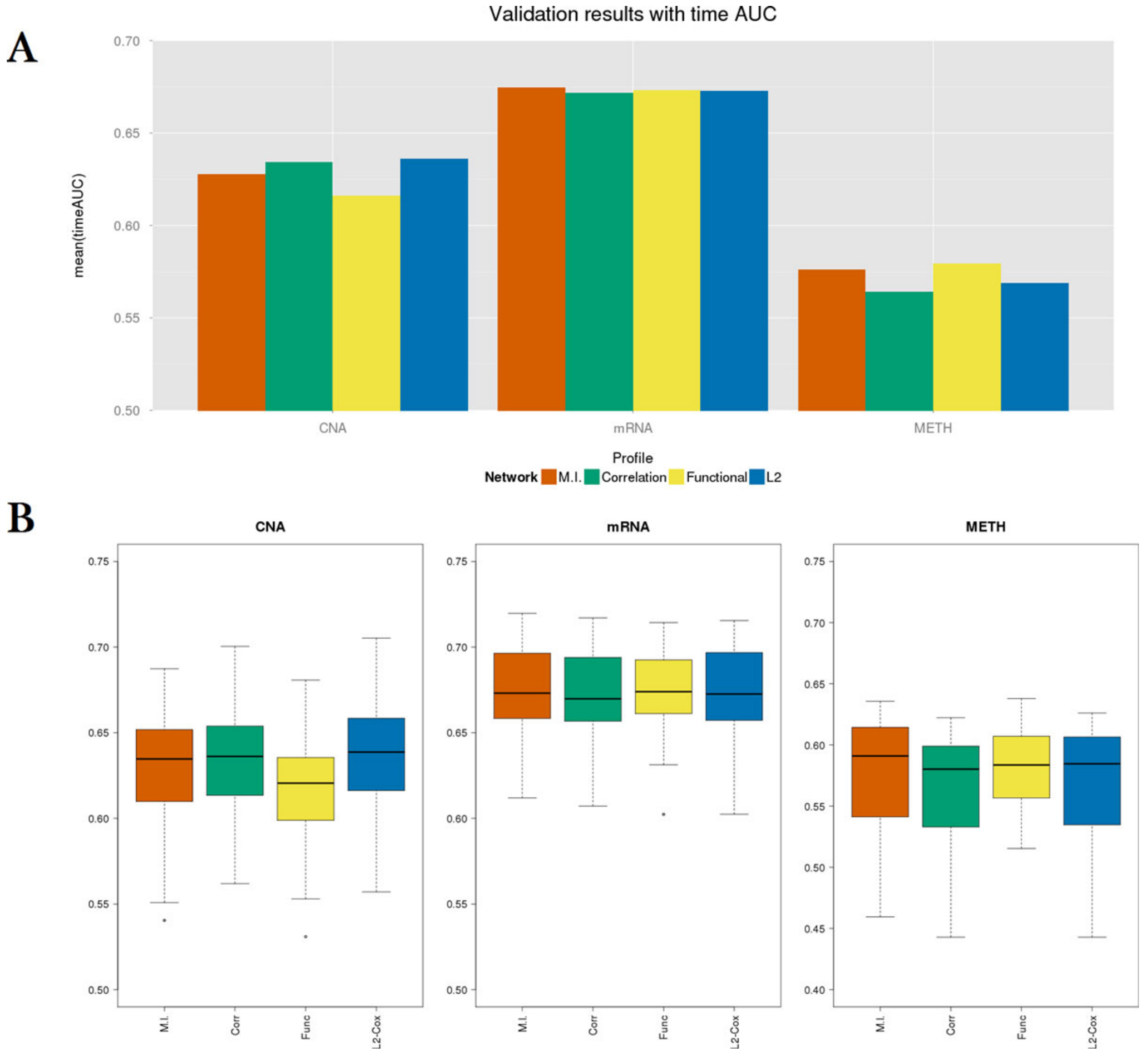

**Figure 3 Performance comparison of different network types on validation set.**

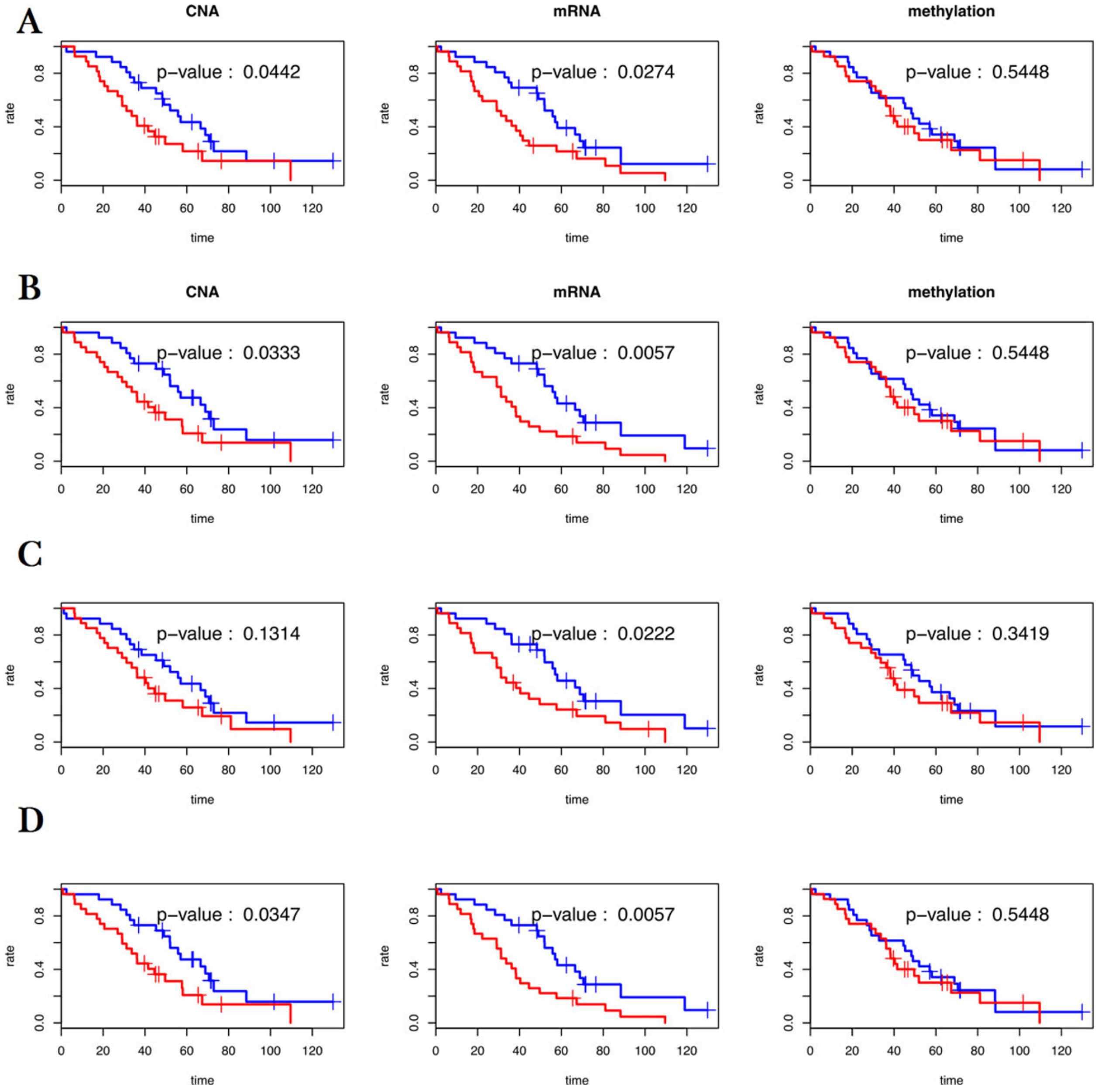

**Figure 4** Kaplan-Meier survival curves (high-risk group: red line; low risk group: blue line) and log-rank test results on validation set for each profile and network (A–Mutual information based network; B–Correlation network; C–Functional linkage network; D–$L_2$ -Cox).

methylation profile is substantially less than in other profiles.

## Signature genes for each profile

In order to examine the genes that have the strongest marginal association with survival, we displayed the five largest regression coefficients in each profile as a heatmap (Figure 5). The top five genes from each profile were all distinct with no overlap, but some genes had large coefficients in multiple profiles. The top genes from the CNA profile had smaller overall regression coefficients than the top genes in other profiles. Moreover, the top genes in the CNA profile had larger regression coefficients than those in other profiles, which suggests that the roles of these genes are more prominent in other genomic levels.

We find that the identified genes are associated with ovarian cancer in many previous studies. The gene with the largest coefficient in the CNA profile was *URI1, prefoldin-like chaperone* (URI1), which is reported to be an

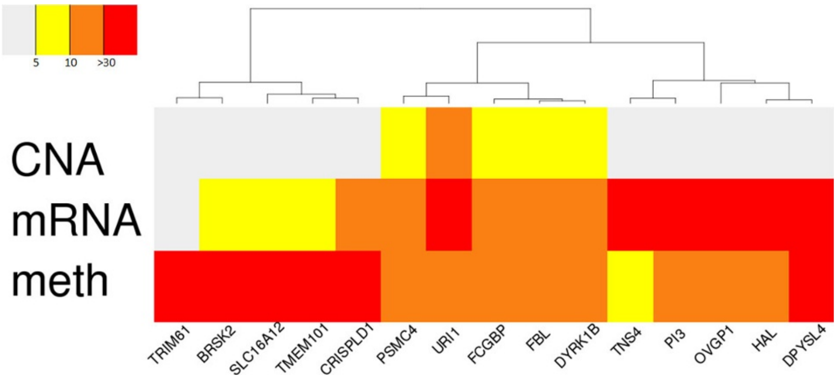

**Figure 5 Heat-map for the regression coefficients of 15 selected genes**. The five genes having the largest regression coefficients are selected for each profile and merged together.

**Table 2 Significantly enriched terms (using *ToppGene*) in the 100 genes that have the largest regression coefficients for each profile**

| Profile | Category | ID | Name | p-value | Adjusted p-value | Count | Total |
|---|---|---|---|---|---|---|---|
| **CNA** | GO:BP | GO:0048706 | embryonic skeletal system development | 5.99E-06 | 1.25E-02 | 7 | 129 |
| | GO:BP | GO:0048704 | embryonic skeletal system morphogenesis | 1.43E-05 | 1.50E-02 | 6 | 98 |
| | GO:BP | GO:0009952 | anterior/posterior pattern specification | 2.17E-05 | 1.52E-02 | 8 | 217 |
| | Disease | ctd:D010051 | Ovarian Neoplasms | 5.24E-04 | 2.84E-02 | 4 | 78 |
| | Disease | ctd:D002277 | Carcinoma | 7.42E-03 | 3.66E-02 | 4 | 161 |
| | GO:BP | GO:0031442 | positive regulation of mRNA 3'-end processing | 7.94E-05 | 4.15E-02 | 3 | 16 |
| | GO:BP | GO:0031440 | regulation of mRNA 3'-end processing | 1.15E-04 | 4.80E-02 | 3 | 18 |
| **mRNA** | GO:MF | GO:0004568 | chitinase activity | 4.80E-06 | 1.78E-03 | 3 | 7 |
| | GO:BP | GO:0006032 | chitin catabolic process | 4.83E-06 | 4.94E-03 | 3 | 7 |
| | GO:BP | GO:0006030 | chitin metabolic process | 4.83E-06 | 4.94E-03 | 3 | 7 |
| | GO:BP | GO:1901072 | glucosamine-containing compound catabolic process | 7.70E-06 | 5.24E-03 | 3 | 8 |
| | GO:MF | GO:0086080 | protein binding involved in heterotypic cell-cell adhesion | 8.12E-05 | 1.50E-02 | 2 | 3 |
| | GO:MF | GO:0098631 | protein binding involved in cell adhesion | 1.62E-04 | 1.50E-02 | 2 | 4 |
| | GO:MF | GO:0098632 | protein binding involved in cell-cell adhesion | 1.62E-04 | 1.50E-02 | 2 | 4 |
| | GO:BP | GO:0046348 | amino sugar catabolic process | 3.86E-05 | 1.97E-02 | 3 | 13 |
| | GO:MF | GO:0008061 | chitin binding | 2.69E-04 | 1.99E-02 | 2 | 5 |
| | GO:MF | GO:0030492 | hemoglobin binding | 7.45E-04 | 4.61E-02 | 2 | 8 |
| **METH** | GO:CC | GO:0030424 | axon | 9.91E-06 | 1.95E-03 | 11 | 409 |
| | GO:CC | GO:0043025 | neuronal cell body | 2.06E-04 | 1.57E-02 | 10 | 478 |
| | GO:CC | GO:0043195 | terminal bouton | 3.15E-04 | 1.57E-02 | 4 | 62 |
| | GO:CC | GO:0036477 | somatodendritic compartment | 3.19E-04 | 1.57E-02 | 12 | 705 |
| | GO:CC | GO:0043005 | neuron projection | 4.07E-04 | 1.61E-02 | 14 | 945 |
| | GO:CC | GO:0044297 | cell body | 5.10E-04 | 1.68E-02 | 10 | 536 |
| | GO:CC | GO:0032983 | kainate selective glutamate receptor complex | 7.48E-04 | 2.11E-02 | 2 | 8 |
| | GO:MF | GO:0004952 | dopamine neurotransmitter receptor activity | 2.80E-04 | 4.01E-02 | 2 | 5 |
| | GO:MF | GO:0001965 | G-protein alpha-subunit binding | 3.95E-04 | 4.01E-02 | 3 | 27 |
| | GO:MF | GO:0015075 | ion transmembrane transporter activity | 4.14E-04 | 4.01E-02 | 13 | 819 |
| | GO:MF | GO:0022857 | transmembrane transporter activity | 5.21E-04 | 4.01E-02 | 14 | 951 |
| | GO:MF | GO:0015026 | coreceptor activity | 5.98E-04 | 4.01E-02 | 3 | 31 |
| | GO:MF | GO:0022891 | substrate-specific transmembrane transporter activity | 7.68E-04 | 4.29E-02 | 13 | 875 |
| | GO:MF | GO:0022892 | substrate-specific transporter activity | 9.57E-04 | 4.58E-02 | 14 | 1012 |

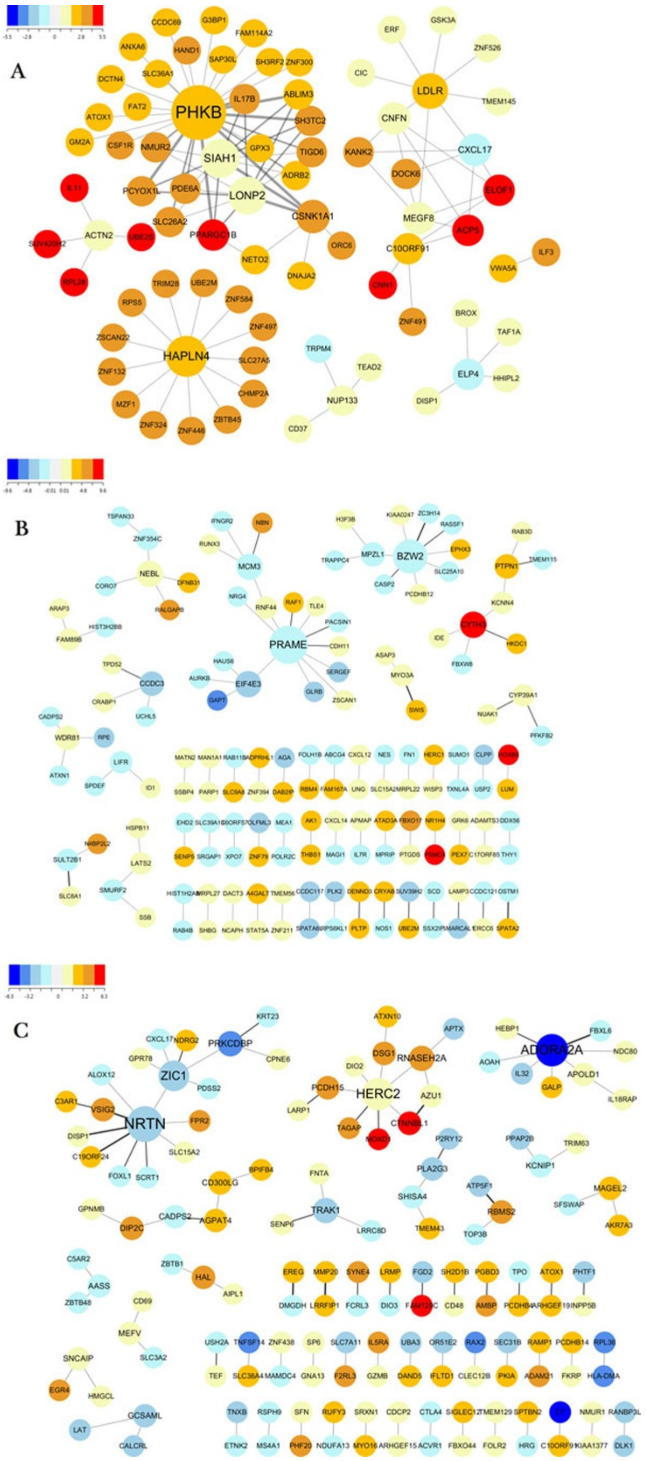

**Figure 6 Gene-gene interaction sub-network constructed by using the 100 edges with the largest mutual information values on each profile (A: CNA, B: mRNA, C: METH).**

oncogene that amplified in ovarian cancer cells [22]. *Dual-Specificity Tyrosine-(Y)-Phosphorylation Regulated Kinase 1B* (DYRK1B), which is over-expressed in a wide spectrum of ovarian cancer cell lines and human specimens [23], was the fourth highest gene. The largest coefficient gene from the mRNA profile, *Oviductal Glycoprotein 1* (OVGP1), is reported to be a more accreted detection marker than other markers of ovarian epithelial cancers

**Table 3 Network properties of mutual information sub-network for each profile**

| Properties | CNA | mRNA | METH |
|---|---|---|---|
| **Nodes** | 78 | 154 | 149 |
| **Connected components** | 7 | 54 | 50 |
| **Network centralization** | 0.312 | 0.058 | 0.052 |
| **Characteristic path length** | 2.117 | 2.187 | 2.164 |
| **Average number of neighbors** | 2.564 | 1.299 | 1.342 |
| **Network density** | 0.033 | 0.008 | 0.009 |
| **Network heterogeneity** | 1.383 | 0.828 | 0.809 |
| **$R^2$ of node degree distribution** | 0.672 | 0.922 | 0.909 |

[24]. Over-expression of the *Elafin/Peptidase Inhibitor 3* (PI3) gene from the mRNA profile is associated with poor overall survival [25]. We found no reported associations of the top five genes in the methylation profile with ovarian cancer from previous studies.

An enrichment analysis was performed for the 100 genes having the largest coefficients in each profile (Table 2). Many significantly enriched terms demonstrated an association with ovarian cancer. For example, the genes URI1, *Processing Of Precursor 4* (POP4), *Pleckstrin Homology Domain Containing, Family F Member 1* (PLEKHF1), and DYRK1B from the CNA profile were enriched in the Ovarian Neoplasms term (ctd:D010051) of Disease. For Gene Ontology, the top 100 genes in the CNA profile were primarily enriched in the Biological Process (BP) terms. The Hox family of homeobox genes was enriched in the terms related with the embryonic skeletal system (GO:0048706; GO:0048704), and these genes are critical for cell migration and DNA repair [26]. The genes in the mRNA profile that were enriched in the Chitinases-related terms were already known for their cancer indication roles [27]. Therefore, we can conclude that the genes with large regression coefficients are expected to be related to ovarian cancer.

### Network analysis

To illustrate the general topology of the mutual information network and its effect on prediction performances more closely, we constructed gene interaction sub-networks by using the 100 edges with the largest mutual information values for each profile (Figure 6). The color of each node represents the strength of its marginal effect. We also see many genes with weak marginal effects appear in the network. Topologies of the networks were analyzed by Cytoscape [28] and summarized in Table 3. The three networks reveal different network structures and topologies. The CNA network consists of a smaller number of connected components than the others (7 versus 54 and 50). It also shows a denser connection between the genes. The network centralization value is about five times higher, and its average number of neighbors is about twice as much. In addition, the overall coefficients of the genes in the CNA network were smaller than those in the other profiles. Considering that those genes have high mutual information values and, hence, strong interactive effects, this may imply that the interaction effect on survival in the CNA profile is more dominant than the marginal effect with each gene. The $R^2$ value for the power-law distribution of the mRNA and methylation networks were 0.922 and 0.909, respectively, which shows strong scale freeness [29], as is the case with many other biological networks [30-33].

Table 4 summarizes the enrichment test results for the genes in each sub-network. The network from the methylation profile was mainly enriched in GO Molecular Function (MF) terms. Networks for the CNA and mRNA profiles were not enriched in GO terms but in terms of diseases related to ovarian cancer, such as Neoplasms (ctd:D009369), Carcinoma (ctd:D002277), and Neoplasm Recurrence, Local (ctd:D009364). Further, the CNA profile network was enriched in other related disease terms, such as Obesity (ctd: D009756) and BMI (Body Mass Index) [34], and

**Table 4 Significantly enriched terms (using ToppGene) in the sub-network consisting of the gene pairs that have the 100 largest mutual information values**

| Profile | Category | ID | Name | p-value | Adjusted p-value | Count | Total |
|---|---|---|---|---|---|---|---|
| **CNA** | Disease | ctd:D006949 | Hyperlipidemias | 1.76E-05 | 2.24E-03 | 3 | 13 |
| | Disease | ctd:D009765 | Obesity | 1.98E-03 | 1.21E-02 | 4 | 132 |
| | Disease | ctd:D007333 | Insulin Resistance | 8.96E-03 | 2.42E-02 | 2 | 35 |
| | Disease | ctd:D009369 | Neoplasms | 1.05E-02 | 2.78E-02 | 2 | 38 |
| **mRNA** | Disease | ctd:D002277 | Carcinoma | 1.19E-04 | 1.23E-02 | 8 | 161 |
| | Disease | ctd:D009364 | Neoplasm Recurrence, Local | 1.25E-02 | 4.96E-02 | 3 | 52 |
| **METH** | GO:MF | GO:0008528 | G-protein coupled peptide receptor activity | 4.15E-05 | 1.21E-02 | 7 | 120 |
| | GO:MF | GO:0001653 | peptide receptor activity | 4.86E-05 | 1.21E-02 | 7 | 123 |
| | GO:MF | GO:0004942 | anaphylatoxin receptor activity | 1.78E-04 | 1.78E-02 | 2 | 3 |
| | GO:MF | GO:0004948 | calcitonin receptor activity | 1.78E-04 | 1.78E-02 | 2 | 3 |
| | GO:MF | GO:0004800 | thyroxine 5′-deiodinase activity | 1.78E-04 | 1.78E-02 | 2 | 3 |

Insulin Resistance (ctd:D007333) [35] and Hyperlipidaemia (ctd:D006949) [36].

## Conclusions

In this study, we investigated the utility of an alternative network construction approach based on mutual information in network-based Cox regression. Our results show that the mutual information based network can further improve prediction performance in survival analyses. Moreover, the permutation testing scheme used to discard insignificant pairs improved the prediction performance.

Overall, the performance gain of this alternative approach over existing methods was rather marginal. It seems due to a mismatch between the high mutual information value and the small value of the penalty term $(\psi(\beta)) = \frac{1}{2}\sum_{i,j}^{p} S_{i,j}(\beta_i - \beta_j)^2)$ in the Net-Cox model–it does not necessarily mean that the gene pairs containing high mutual information with respect to survival should have similar marginal effects because the mutual information measure is more concerned with the interaction effect. Even with this discrepancy, results based on the mutual information network are still promising. In future studies, the regularization term could be modified to better reflect the information contained in the mutual information network and, hence, further improve the performance. Another direction would be to apply this network scheme to network-based approaches in other domains.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

HJ, KW, and KS designed the study. HJ and SK implemented the idea and performed the experiments. HJ, SK, KW, and KS developed the idea and performed the analysis. HJ, SK, KW, and KS wrote the paper.

### Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Science, ICT, and Future Planning (MSIP) (No. 2010-0028631), and by the Ministry of Education (2012R1A1A2042792).

### Declarations

The publication costs of this article were funded by Ajou University, S. Korea. This article has been published as part of *BMC Systems Biology* Volume 9 Supplement 1, 2015: Selected articles from the Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015): Systems Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/9/S1

Published: 21 January 2015

### References

1. Cox D: **Regression Models and Life-Tables.** In *Breakthroughs in Statistics.* Springer New York;Kotz S, Johnson N 1992:527-541.
2. Tibshirani R: **The lasso method for variable selection in the Cox model.** *Statistics in medicine* 1997, **16(4)**:385-395.
3. Li H, Luan Y: **Kernel Cox regression models for linking gene expression profiles to censored survival data.** *Pacific Symposium on Biocomputing* 2002, 65.
4. Gui J, Li H: **Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data.** *Bioinformatics* 2005, **21(13)**:3001-3008.
5. Li C, Li H: **Network-constrained regularization and variable selection for analysis of genomic data.** *Bioinformatics* 2008, **24(9)**:1175-1182.
6. Kim S, Sohn K-A, Xing EP: **A multivariate regression approach to association analysis of a quantitative trait network.** *Bioinformatics* 2009, **25(12)**:i204-i212.
7. Cheng W, Zhang X, Guo Z, Shi Y, Wang W: **Graph-regularized dual Lasso for robust eQTL mapping.** *Bioinformatics* 2014, **30(12)**:i139-i148.
8. Lee S, Xing EP: **Leveraging input and output structures for joint mapping of epistatic and marginal eQTLs.** *Bioinformatics* 2012, **28(12)**:i137-i146.
9. 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 computational biology* 2013, **9(3)**: e1002975.
10. Steuer R, Kurths J, Daub CO, Weise J, Selbig J: **The mutual information: detecting and evaluating dependencies between variables.** *Bioinformatics (Oxford, England)* 2002, **18(Suppl 2)**:S231-240.
11. Goebel B, Dawy Z, Hagenauer J, Mueller JC: **An approximation to the distribution of finite sample size mutual information estimates.** *IEEE International Conference on Communications, 2005 ICC* 2005, **2**:1102-1106.
12. Leem S, Jeong H-h, Lee J, Wee K, Sohn K-A: **Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure.** *Computational Biology and Chemistry* 2014.
13. Hu T, Sinnott-Armstrong Na, Kiralis JW, Andrew AS, Karagas MR, Moore JH: **Characterizing genetic interactions in human disease association studies using statistical epistasis networks.** *BMC bioinformatics* 2011, **12**:364.
14. Butte AJ, Kohane IS: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.** *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2000, **426**:418-429.
15. Heagerty PJ, Lumley T, Pepe MS: **Time-dependent ROC curves for censored survival data and a diagnostic marker.** *Biometrics* 2000, **56(2)**:337-344.
16. Li H, Gui J: **Partial Cox regression analysis for high-dimensional microarray gene expression data.** *Bioinformatics* 2004, **20(Suppl 1)**: i208-215.
17. Mantel N: **Evaluation of survival data and two new rank order statistics arising in its consideration.** *Cancer chemotherapy reports Part 1* 1966, **50(3)**:163-170.
18. Chen J, Bardes EE, Aronow BJ, Jegga AG: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.** *Nucleic Acids Res* 2009, , **37 Web Server:** W305-311.
19. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**:25-29.
20. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, *et al*: **The Comparative Toxicogenomics Database: update 2013.** *Nucleic Acids Res* 2013, , **41 Database:** D1104-1114.
21. Benjamini Y: **Discovering the false discovery rate.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010, **72(4)**:405-416.
22. Theurillat JP, Metzler SC, Henzi N, Djouder N, Helbling M, Zimmermann AK, Jacob F, Soltermann A, Caduff R, Heinzelmann-Schwarz V, *et al*: **URI is an oncogene amplified in ovarian cancer cells and is required for their survival.** *Cancer Cell* 2011, **19(3)**:317-332.
23. Gao J, Yang X, Yin P, Hu W, Liao H, Miao Z, Pan C, Li N: **The involvement of FoxO in cell survival and chemosensitivity mediated by Mirk/Dyrk1B in ovarian cancer.** *International journal of oncology* 2012, **40(4)**:1203-1209.
24. Maines-Bandiera S, Woo MM, Borugian M, Molday LL, Hii T, Gilks B, Leung PC, Molday RS, Auersperg N: **Oviductal glycoprotein (OVGP1, MUC9): a differentiation-based mucin present in serum of women with ovarian cancer.** *Int J Gynecol Cancer* 2010, **20(1)**:16-22.
25. Clauss A, Ng V, Liu J, Piao H, Russo M, Vena N, Sheng Q, Hirsch MS, Bonome T, Matulonis U, *et al*: **Overexpression of elafin in ovarian carcinoma is driven by genomic gains and activation of the nuclear**

factor kappaB pathway and is associated with poor overall survival. *Neoplasia* 2010, **12(2)**:161-172.

26. Liu Y, Devescovi V, Chen S, Nardini C: **Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties.** *BMC Systems Biology* 2013, **7(1)**:14.
27. Kzhyshkowska J, Gratchev A, Goerdt S: **Human chitinases and chitinase-like proteins as indicators for inflammation and cancer.** *Biomarker insights* 2007, **2**:128.
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.** *Genome Research* 2003, **13**:2498-2504.
29. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Statistical applications in genetics and molecular biology* 2005, **4**:Article17.
30. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási aL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407**:651-654.
31. Diez D, Wheelock AM, Goto S, Haeggström JZ, Paulsson-Berne G, Hansson GK, Hedin U, Gabrielsen A, Wheelock CE: **The use of network analyses for elucidating mechanisms in cardiovascular disease.** *Molecular bioSystems* 2010, **6**:289-304.
32. Carter SL, Brechbühler CM, Griffin M, Bond AT: **Gene co-expression network topology provides a framework for molecular characterization of cellular state.** *Bioinformatics (Oxford, England)* 2004, **20**:2242-2250.
33. Reguly T, Breitkreutz A, Boucher L, Breitkreutz B-J, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, *et al*: **Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae.** *Journal of biology* 2006, **5**:11.
34. Engeland A, Tretli S, Bjørge T: **Height, Body Mass Index, and Ovarian Cancer: A Follow-Up of 1.1 Million Norwegian Women.** *Journal of the National Cancer Institute* 2003, **95(16)**:1244-1248.
35. Arcidiacono B, Iiritano S, Nocera A, Possidente K, Nevolo MT, Ventura V, Foti D, Chiefari E, Brunetti A: **Insulin resistance and cancer risk: an overview of the pathogenetic mechanisms.** *Experimental diabetes research* 2012, **2012**:789174.
36. Parazzini F, Moroni S, La Vecchia C, Negri E, dal Pino D, Bolis G: **Ovarian cancer risk and history of selected medical conditions linked with female hormones.** *European journal of cancer* 1997, **33(10)**:1634-1637.

doi:10.1186/1752-0509-9-S1-S8
**Cite this article as:** Jeong *et al.*: Investigating the utility of clinical outcome-guided mutual information network in network-based Cox regression. *BMC Systems Biology* 2015 **9**(Suppl 1):S8.