Bioinformatics, 36(6), 2020, 1814–1822
doi: 10.1093/bioinformatics/btz809
Advance Access Publication Date: 5 November 2019
Original Paper

Systems biology

Network-based multi-task learning models for biomarker selection and cancer outcome prediction

Zhibo Wang[1,2], Zhezhi He[3], Milan Shah[4], Teng Zhang[5], Deliang Fan[3] and Wei Zhang[1,2,*]

[1]Department of Computer Science and [2]Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL 32816, USA, [3]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA, [4]Department of Computer Science, Duke University, Durham, NC, 27708, USA and [5]Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received on April 11, 2019; revised on October 6, 2019; editorial decision on October 27, 2019; accepted on October 30, 2019

Abstract

Motivation: Detecting cancer gene expression and transcriptome changes with mRNA-sequencing or array-based data are important for understanding the molecular mechanisms underlying carcinogenesis and cellular events during cancer progression. In previous studies, the differentially expressed genes were detected across patients in one cancer type. These studies ignored the role of mRNA expression changes in driving tumorigenic mechanisms that are either universal or specific in different tumor types. To address the problem, we introduce two network-based multi-task learning frameworks, NetML and NetSML, to discover common differentially expressed genes shared across different cancer types as well as differentially expressed genes specific to each cancer type. The proposed frameworks consider the common latent gene co-expression modules and gene–sample biclusters underlying the multiple cancer datasets to learn the knowledge crossing different tumor types.

Results: Large-scale experiments on simulations and real cancer high-throughput datasets validate that the proposed network-based multi-task learning frameworks perform better sample classification compared with the models without the knowledge sharing across different cancer types. The common and cancer-specific molecular signatures detected by multi-task learning frameworks on The Cancer Genome Atlas ovarian, breast and prostate cancer datasets are correlated with the known marker genes and enriched in cancer-relevant Kyoto Encyclopedia of Genes and Genome pathways and gene ontology terms.

Availability and implementation: Source code is available at: https://github.com/compbiolabucf/NetML.

Contact: wzhang.cs@ucf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Powered by high-throughput genomic technologies, it is now common practice to perform large-scale experiments for measuring mRNA expressions for cancer studies. Correlating these high-dimensional genomic features with cancer phenotype as molecular signatures (i.e. biomarkers) to detect gene expression changes can possibly improve cancer diagnosis and treatment over current clinical measures for risk assessment of patients (Van't Veer *et al.*, 2002; Weinstein *et al.*, 2013; Zhang *et al.*, 2017). Discovery of biomarkers is typically modeled as a feature selection problem. Many biomarker selection techniques have been proposed in the last few decades (Danaee *et al.*, 2017; Saeys *et al.*, 2007; Way and Greene, 2018). These techniques can be categorized into three groups: (i) *Univariate biomarker selection techniques* (Baldi and Long, 2001; Breitling *et al.*, 2004; Jafari and Azuaje, 2006; Thomas *et al.*, 2001), this category includes parametric methods (e.g. *t*-test and ANOVA) and model-free methods (e.g. Rank products and Wilcoxon rank sum). Both of them do not consider the interaction with the classifiers and also ignore the feature dependencies. (ii) *Multivariate biomarker selection techniques* Ding and Peng (2005) and Xing *et al.* (2001), the methods in this category also ignore the interactions with the classifier. However, they introduce a number of multivariate filer techniques, which is aiming at the incorporation of feature dependencies to some degree (Saeys *et al.*, 2007). (iii) *Embedded biomarker selection techniques* (Díaz-Uriarte and De Andres, 2006; Guyon *et al.*, 2002), the biomarker selection methods in this category search for an optimal subset of features, which is built into the classifier construction. The methods in this category can be seen as a search in the combined space of a subset of features and sample labels. The recently developed autoencoder-based biomarker selection methods can also be categorized into this group (Danaee *et al.*, 2017; Way and Greene, 2018). There are two major limitations of these popular methods. First, most biomarker selection methods rank the features by their individual correlation with the label (class) information of observations, and thus relations among features, e.g. the highly correlated features play some functions together. It has been shown that the signature genes identified by univariate feature selection technique from high dimension, low-sample-size genomic data are not consistent and robust due to statistical randomness and high levels of experimental noise. Second, previous studies have shown that the differentially expressed genes found in one cancer type can also be found in other cancer types (Cao and Zhang, 2016; Makhijani *et al.*, 2018). There is no unified mathematical model to simultaneously detect the differential expression events common or specific to multiple cancer types from gene expression datasets.

To address these limitations, we propose two learning frameworks, *Net*work-based *M*ulti-task *L*earning model (NetML) and *Net*work-based *S*emi-supervised *M*ulti-task *L*earning model (NetSML), to detect differentially expressed genes and classify unlabeled cancer patient samples from gene expression datasets of multiple cancer types, in which each cancer type can be regarded as one domain in multi-task learning. Multi-task learning uses common knowledge or structures among different domains to enhance multiple learning tasks (Petegrosso *et al.*, 2016; Zhang *et al.*, 2013). In the frameworks, we introduce both cancer-specific features and cancer common features in network-based learning models. The first multi-task learning model, NetML, runs a network-based algorithm on several gene graphs, where each graph represents co-expression information between pair of genes in one cancer type. The framework integrates gene co-expression and common gene modules across cancer datasets for biomarker selection. To identify the biomarkers from high-throughput gene expression datasets, *Lasso* (Tibshirani, 1996) is applied to both cancer-specific genes and common latent genes to preserve the sparsity. The second Multi-task learning framework, NetSML, is a semi-supervised learning model, which runs a network-based algorithm on sample-feature bipartite graphs to identify biomarkers and classify cancer samples across different cancer types. It also explores bi-clusters between patients and features to find biomarkers specific to subsets of patient samples. By using alternating optimization to solve both models, common molecular signatures and cancer-specific signatures could be detected from multiple cancer types. When compared with the baseline methods without knowledge crossing different domains, the proposed network-based learning models are more robust and identify more accurate signatures for cancer outcome prediction.

2 Materials and methods

In this section, we first introduce the mathematical notations and then a base network-based learning model that is widely used for biomarker selection from cancer genomic data. We next introduce a NetML model to discover common molecular signatures shared across different cancer types and cancer-specific signatures from high-dimensional gene expression data. Then we further extend the NetSML model for phenotype predictions and molecular signature identification.

2.1 Notations

The notations to define the models are summarized in Table 1. Let m be the number of genes, n be the number of samples in one cancer study and $\boldsymbol{W} \in \mathbb{R}^{m\times m}$ be the gene correlation network based on the absolute value of the Pearson's correlation coefficients between the pair of genes, where $\boldsymbol{W}_{ij}$ is the correlation between the two vectors in $\mathbb{R}^n$ that represent the ith and the jth genes. Then the gene correlation network $\boldsymbol{W}$ is used to construct a normalized graph Laplacian $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{S}$, where $\boldsymbol{S} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$ and $\boldsymbol{D}$ is a diagonal matrix with the column-sum of $\boldsymbol{W}$ on the diagonal entries.

Table 1. Notations for NetML model

Notation	Definition
δ	No. of domains (e.g. cancer types)
n_d	No. of samples in domain $d \in [1, \delta]$
m	No. of genes
$\boldsymbol{I} \in \mathbb{R}^{m\times m}$	Identity matrix
$\boldsymbol{W}_d \in \mathbb{R}^{m\times m}$	Adjacency matrix of gene correlation network
$\boldsymbol{D}_d \in \mathbb{R}^{m\times m}$	Diagonal matrix: $\boldsymbol{D}_d(i,i) = \sum_j \boldsymbol{W}_d(i,j)$
$\boldsymbol{S}_d \in \mathbb{R}^{m\times m}$	Normalized adjacency matrix $\boldsymbol{S}_d = \boldsymbol{D}_d^{-\frac{1}{2}}\boldsymbol{W}_d\boldsymbol{D}_d^{-\frac{1}{2}}$
$\boldsymbol{L}_d \in \mathbb{R}^{m\times m}$	Normalized graph Laplacian: $\boldsymbol{L}_d = \boldsymbol{I} - \boldsymbol{S}_d$
$\boldsymbol{f}_d \in \mathbb{R}^{m\times 1}$	Coefficients of domain-specific genes
$\boldsymbol{f}_c \in \mathbb{R}^{m\times 1}$	Coefficients of common genes
$\boldsymbol{y}_d \in \mathbb{R}^{m\times 1}$	Correlation coefficients between gene expressions and labels of samples in domain d
$\alpha, \gamma_d, \gamma_c \in \mathbb{R}_+$	Hyper-parameters

2.2 Standard network-based learning model

As a base model, we first introduce a network-based learning model that was applied successfully to identify molecular signatures in several variations (Winter *et al.*, 2012; Zhang *et al.*, 2010, 2012). Given a gene correlation network (e.g. gene co-expression network), the objective of the network-based learning model is to learn an assignment function $\boldsymbol{f} \in \mathbb{R}^{m\times 1}$, which represents the importance of each gene in one cancer study. The initial labeling is $\boldsymbol{f}^0 = \boldsymbol{y}$, i.e. the Pearson's correlation coefficients between gene expressions and the case/control labeling of the samples. The high absolute value indicates the differentially expressed gene. The network-based learning model assumes that genes should be assigned similar importance scores if they are co-expressed, which leads to the following objective function to be minimized:

$$\mathcal{L}(\boldsymbol{f}) = \alpha\boldsymbol{f}^T\boldsymbol{L}\boldsymbol{f} + (1-\alpha)||\boldsymbol{f} - \boldsymbol{y}||_2^2, \tag{1}$$

where $\alpha \in (0, 1)$ is a parameter to balance the contributions of the two terms in Equation (1), the first of which is the Laplacian term encouraging assigning similar importance scores to strongly connected vertices in the gene correlation network and the second term is the fitting term, which encourages consistency between the importance score and the initial score. The first term can be rewritten as

$$\alpha\boldsymbol{f}^T\left(\boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}\right)\boldsymbol{f},$$

where $\boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$ is the normalized graph Laplacian, which is positive semi-definite. Thus, Equation (1) is a quadratic optimization problem with a closed-form solution. The corresponding set of the eigenvalues or spectrum of the $\boldsymbol{L}$ matrix reflects many properties of the gene correlation graph (Chung and Graham, 1997; Li and Li, 2008).

2.3 NetML model

To introduce multi-task learning among different cancer datasets, we extend the standard model in Equation (1) on multiple cancer types and present a network-based method NetML to learn the common molecular signatures $\boldsymbol{f}_c$ shared across different cancer types and cancer-specific signatures $\boldsymbol{f}_d$ to each cancer type d. Figure 1 shows a schematic of the experimental workflow of NetML. The gene correlation network is built up for each individual cancer type based on the gene expression generated from RNA-seq or array-based data. The molecular signatures learned from the model will be used for cancer outcome prediction. The objective function of NetML is defined as follow:

$$\begin{aligned}\mathcal{L}(\boldsymbol{f}_c, \{\boldsymbol{f}_d\}_{d=1}^{\delta}) = \textstyle\sum_{d=1}^{\delta}[&\alpha(\boldsymbol{f}_c + \boldsymbol{f}_d)^T\boldsymbol{L}_d(\boldsymbol{f}_c + \boldsymbol{f}_d)\\ &+(1-\alpha)||\boldsymbol{f}_c + \boldsymbol{f}_d - \boldsymbol{y}_d||_2^2 + \gamma_d||\boldsymbol{f}_d||_1]\\ &+\gamma_c||\boldsymbol{f}_c||_1,\end{aligned} \tag{2}$$

where $\boldsymbol{L}_d$ is the normalized graph Laplacian to represent the gene correlation network for cancer type d. $\boldsymbol{y}_d$ is the correlation coefficients between the label of the samples in cancer type d and the expression value

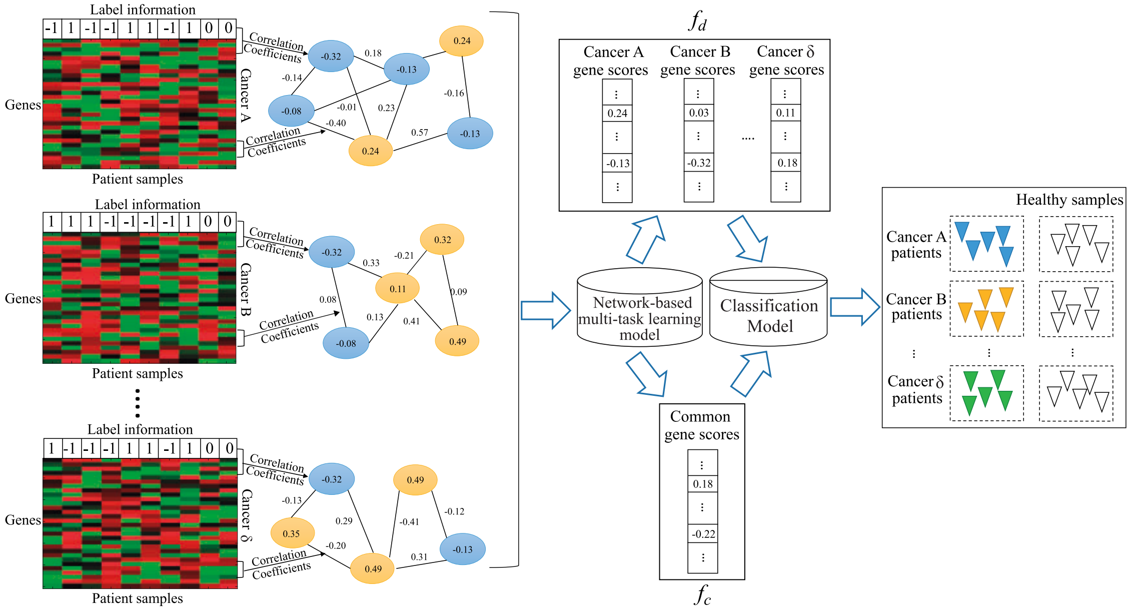

Fig. 1. Running NetML model on gene correlation graphs. A gene correlation graph is constructed from gene expression data for each cancer type. The vertices are then initialized by the correlation between each individual gene expression and the labels. The proposed NetML model re-ranks all the genes to get cancer-specific gene scores and common gene scores across all the cancer types for biomarker discovery. The selected biomarkers can be used as molecular signatures for cancer outcome prediction

of each gene. The cancer-specific signatures f_d and common signatures f_c are learned together in the model. When compared with Equation (1), the multi-task learning model is regularized with a *Lasso* penalty (L_1-norm) on f_d and f_c, which induces a sparse solution for feature selection (Tibshirani, 1996). $\gamma_d > 0$ and $\gamma_c > 0$ are two regularization parameters. The hyper-parameter α balances the Laplacian term $(f_c + f_d)^T L_d (f_c + f_d)$ and the fitting term $||f_c + f_d - y_d||_2^2$ in each cancer type (domain) introduced in the standard model. The common gene score vector f_c is shared across all the cancer types, and only the significant molecular signatures in all the domains will be selected (nonzero elements in f_c).

2.3.1 Alternative optimization algorithm

The objective function defined by Equation (2) can be solved by alternative optimization of f_d and f_c. The minimization with respect to f_d and f_c are equivalent to solving two *Lasso*-type optimization problems.

Computation of f_d

If we fix variable f_c, solving Equation (2) with respect to each f_d, where $d \in [1, \delta]$, can be simplified as:

$$\begin{aligned}\mathcal{L}(f_d) = \alpha(f_c + f_d)^T \boldsymbol{L}_d (f_c + f_d) \\ +(1-\alpha)||f_c + f_d - \boldsymbol{y}_d||_2^2 + \gamma_d ||f_d||_1.\end{aligned} \tag{3}$$

Let $\boldsymbol{Y}_d = -[\alpha f_c^T \boldsymbol{X}_d^T - (1-\alpha)\boldsymbol{y}_d^T \boldsymbol{X}_d^{-1}]^T$ and $\boldsymbol{X}_d \in \mathbb{R}^{m \times m}$ satisfies $\boldsymbol{X}_d^T \boldsymbol{X}_d = [\alpha \boldsymbol{L}_d + (1-\alpha)\boldsymbol{I}]$. In this implementation, we let the eigenvalue decomposition of $[\alpha \boldsymbol{L}_d + (1-\alpha)\boldsymbol{I}]$ be $\boldsymbol{U}_d \boldsymbol{S}_d \boldsymbol{U}_d^T$ and $\boldsymbol{X}_d = \boldsymbol{S}_d^{\frac{1}{2}} \boldsymbol{U}_d^T$. Then Equation (3) can be rewritten as:

$$\mathcal{L}(f_d) = ||\boldsymbol{Y}_d - \boldsymbol{X}_d f_d||_2^2 + \gamma_d ||f_d||_1 + C_1, \tag{4}$$

where C_1 is a constant. Therefore, to minimize the objective function $\mathcal{L}(f_d)$ is equivalent to solve the above *Lasso* problem. The Python package *scikit-learn* can be used to solve the problem. The detailed derivation from Equations (3) to (4) can be found in Supplementary Material.

Computation of f_c

Similar to updating f_d, solving Equation (2) with respect to each f_c if f_d is fixed can be simplified as:

$$\begin{aligned}\mathcal{L}(f_c) = \textstyle\sum_{d=1}^{\delta}[\alpha(f_c + f_d)^T \boldsymbol{L}_d (f_c + f_d) \\ +(1-\alpha)||f_c + f_d - \boldsymbol{y}_d||_2^2] + \gamma_c ||f_c||_1.\end{aligned} \tag{5}$$

Let $\boldsymbol{Y}_c = -[\sum_{d=1}^{\delta}[\alpha f_d^T \boldsymbol{L}_d + (1-\alpha)(f_d - \boldsymbol{y}_d)^T]\boldsymbol{X}_c^{-1}]^T$ and $\boldsymbol{X}_c \in \mathbb{R}^{m \times m}$ satisfies $\boldsymbol{X}_c^T \boldsymbol{X}_c = \sum_{d=1}^{\delta}[\alpha \boldsymbol{L}_d + (1-\alpha)\boldsymbol{I}]$. Then Equation (5) can be rewritten as:

$$\mathcal{L}(f_c) = ||\boldsymbol{Y}_c - \boldsymbol{X}_c f_c||_2^2 + \gamma_c ||f_c||_1 + C_2, \tag{6}$$

where C_2 is a constant. Therefore, to minimize the objective function $\mathcal{L}(f_c)$ is also equivalent to solving a *Lasso* problem. The detailed derivation from Equations (5) to (6) can be found in Supplementary Material.

Algorithm 1. NetML model.

```
Input: $\boldsymbol{L}_d$, $\boldsymbol{y}_d$, $\alpha$, $\gamma_c$, $\gamma_d$ and $\delta$
Output: $f_d, f_c$
Initialization: $f_d, f_c = \left[\frac{1}{m}; \frac{1}{m}; \ldots, \frac{1}{m}\right]^m$, maxIter = 100,
    $\epsilon$ = 1e−7
for $d = 1 \rightarrow \delta$ do
    Let the eigenvalue decomposition of $[\alpha \boldsymbol{L}_d + (1-\alpha)\boldsymbol{I}]$ be
    $\boldsymbol{U}_d \boldsymbol{S}_d \boldsymbol{U}_d^T$
    $\boldsymbol{S}_d^{\frac{1}{2}} \boldsymbol{U}_d^T \rightarrow \boldsymbol{X}_d$
end for
for $t = 1 \rightarrow$maxIter do
    for $d = 1 \rightarrow \delta$ do
        $-[\alpha f_c^T \boldsymbol{X}_d^T - (1-\alpha)\boldsymbol{y}_d^T \boldsymbol{X}_d^{-1}]^T \rightarrow \boldsymbol{Y}_d$
        $\arg\min_{f_d}(||\boldsymbol{Y}_d - \boldsymbol{X}_d f_d||_2^2 + \gamma_d ||f_d||_1 + C_1) \rightarrow f_d$
    end for
    $\sum_{\mathrm{d}=1}^{\delta}[\alpha \mathbf{L}_d + (1-\alpha)\mathbf{I}] = \mathbf{X}_c^T \mathbf{X}_c \rightarrow \mathbf{X}_c$
    $-[\sum_{\mathrm{d}=1}^{\delta}[\alpha \mathbf{f}_d^T \mathbf{L}_d + (1-\alpha)(f_d - \boldsymbol{y}_d)^T]\mathbf{X}_c^{-1}]^T \rightarrow \mathbf{Y}_c$
    $\arg\min_{f_c}(||\boldsymbol{Y}_c - \boldsymbol{X}_c f_c||_2^2 + \gamma_c ||f_c||_1 + C_2) \rightarrow f_c$
    $f_c \rightarrow f_c^{(t)}$
    if$||f_c^{(t)} - f_c^{(t-1)}||_1 < \epsilon$
        break
    end if
end for
return: $f_d, f_c$
```

The complete NetML algorithm is outlined in Algorithm 1. In the algorithm, the for-loop between Lines 4–7 calculates $\boldsymbol{X}_d$ for each cancer type d. The outer for-loop between Lines 8–20 performs multiple passes for updating f_d and f_c. The inner for-loop between Lines 9–12 scans through each cancer type to update each f_d for all δ cancer types. Lines 13–16 update f_c. The convergence of the algorithm is checked at Line 17. After convergence, the non-zero elements in f_d and f_c are cancer-specific genomic features and common genomic features for phenotype prediction as illustrated in Figure 1. In the real experiment, if the coefficients for one feature in all the f_ds and f_c are non-zero, we set the values in f_ds to be 0 and keep the original coefficient in f_c since these are common features cross domains identified by the model.

2.4 NetSML model

We next extend the framework in Equation (2) for semi-supervised multi-task learning on sample-feature bipartite graphs. The model NetSML is formulated for cancer-specific and common molecular signatures discoveries and also for cancer diagnosis/prognosis as a semi-supervised learning problem. For each cancer type d, a sample-gene bipartite graph is constructed as shown in Figure 2. Each edge connecting sample vertex and gene vertex is weighted by the corresponding gene expression in the sample as illustrated in Figure 2. The sample vertices in the bipartite graph are labeled with +1/−1/0 (tumor/healthy/unlabeled) as illustrated by the blue rectangle and the gene vertices are initialized with zeros. NetSML will assign values to all the blue vertices.

Let $y_{d(u)}$ and $y_{d(v)}$ denote the initial values in sample vertices and gene vertices, respectively, and u and v indicate the samples and genes. The normalized gene expression value is denoted by S_d in cancer type d. $S_d = \boldsymbol{D}_{d(v)}^{-\frac{1}{2}} \boldsymbol{W}_d \boldsymbol{D}_{d(u)}^{-\frac{1}{2}}$, where $\boldsymbol{W}_d$ is the raw gene expression data in cancer type d. $\boldsymbol{D}_{d(v)}$ and $\boldsymbol{D}_{d(u)}$ are diagonal matrices, and the elements on the diagonals represent row sums and column sums of $\boldsymbol{W}_d$. The notations to define NetSML are summarized in Table 2. In this context, the cost function over the bipartite graphs is defined as:

$$\mathcal{L}(f_{c(v)}, \{f_{d(v)}, f_{d(u)}\}_{d=1}^{\delta}) = \sum_{d=1}^{\delta}[||f_{c(v)} + f_{d(v)}||_2^2 + ||f_{d(u)}||_2^2 \\ -2(f_{c(v)} + f_{d(v)})^T S_d f_{d(u)} + \alpha_1 ||f_{d(v)} + f_{c(v)} - y_{d(v)}||_2^2 \\ +\alpha_2 ||f_{d(u)} - y_{d(u)}||_2^2 + \gamma_d ||f_{d(v)}||_1] + \gamma_c ||f_{c(v)}||_1, \qquad (7)$$

where $f_{d(v)}$ and $f_{c(v)}$ are the coefficients of cancer-specific genes and common genes, respectively, $f_{d(u)}$ is the label of the samples in cancer type d. All three variables are learned together by minimizing the Equation (7). $\gamma_d > 0$ and $\gamma_c > 0$ are two regularization parameters. The hyper-parameters α_1 and α_2 balance three different parts in the objective function. The first part, $||f_{c(v)} + f_{d(v)}||_2^2 + ||f_{d(u)}||_2^2 - 2(f_{c(v)} + f_{d(v)})^T S_d f_{d(u)}$, constrains new scores/labels assigned to the variables to be consistent between the connected sample–gene pairs in each bipartite graph. The second part, $||f_{d(v)} + f_{c(v)} - y_{d(v)}||_2^2$, is a fitting term which keeps the total score (cancer-specific score and common score across all cancer types) assigned to each gene consistent with the initial score. The third part, $||f_{d(u)} - y_{d(u)}||_2^2$, is used in the same spirit to constrain the cost on the sample vertices. For the unlabeled sample vertices with $y_{d(u)} = 0$, the fitting term is used to regularize these $f_{d(u)}$ such that the total cost is constrained.

2.4.1 Alternative optimization algorithm

Similar to Algorithm 1 for solving the model in Equation (2). The objective function defined by Equation (7) can also be solved by alternative optimization of $f_{d(v)}$, $f_{c(v)}$ and $f_{d(u)}$.

Computation of $f_{d(v)}$

If we fix variables $f_{c(v)}$ and $f_{d(u)}$, solving Equation (7) with respect to each $f_{d(v)}$ can be simplified as solving a *Lasso* problem:

Table 2. Notations for NetSML model

Notation	Definition
δ	No. of domains (e.g. cancer types)
n_d	No. of samples in domain $d \in [1, \delta]$
m	No. of genes
$\boldsymbol{W}_d \in \mathbb{R}^{m \times n_d}$	Gene expression in domain d
$S_d \in \mathbb{R}^{m \times n_d}$	Normalized gene expression in domain d
$f_{d(v)} \in \mathbb{R}^{m \times 1}$	Coefficients of domain-specific genes
$f_{c(v)} \in \mathbb{R}^{m \times 1}$	Coefficients of common genes
$f_{d(u)} \in \mathbb{R}^{n_d \times 1}$	Predicted labels of samples in domain d
$y_{d(u)} \in \mathbb{R}^{n_d \times 1}$	Initial labels of samples in domain d
$y_{d(v)} \in \mathbb{R}^{m \times 1}$	Initial scores of genes in domain d
$\alpha_1, \alpha_2, \gamma_d, \gamma_c \in \mathbb{R}_+$	Hyper-parameters

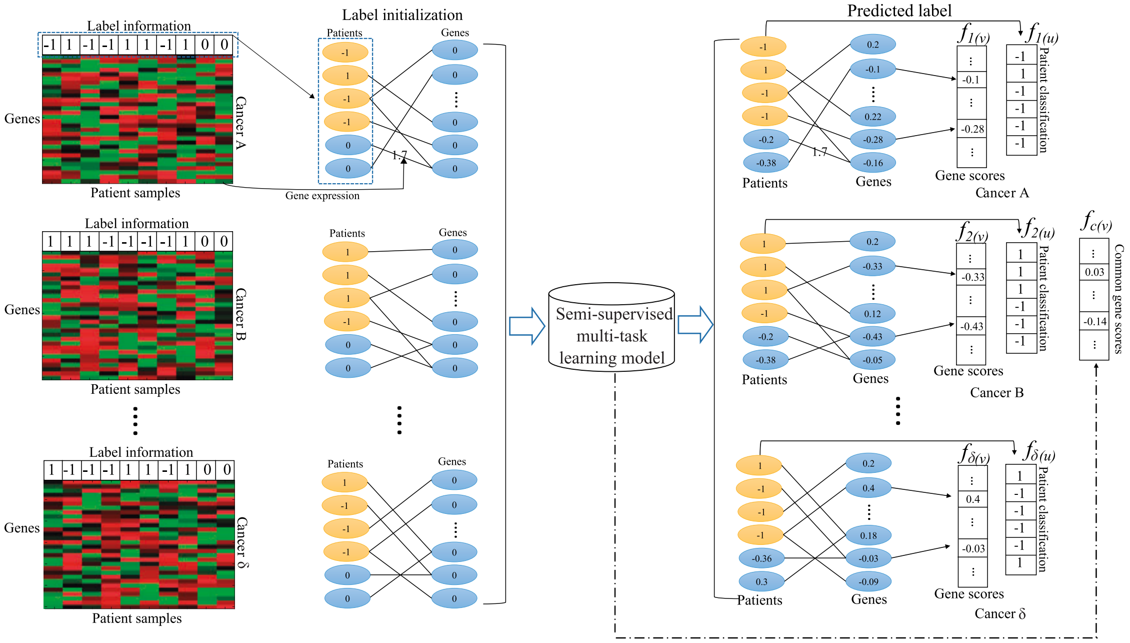

Fig. 2. Running NetSML model on sample-feature bipartite graphs. Gene expression data in each cancer study are modeled as sample-feature bipartite graphs. The sample vertices are initialized by the case/control labels and the gene vertices are initialized with 0. The sample vertices are connected to all the gene vertices and the edges are weighted by the gene expression value. Semi-supervised multi-task learning model classifies the unlabeled samples in each cancer type and ranks the genes based on their cancer-specific gene scores $f_{d(v)}$ and common gene scores $f_{c(v)}$

$$\mathcal{L}(f_{d(v)}) = ||\mathbf{Y}_{d(v)} - f_{d(v)}||_2^2 + \gamma'_d||f_{d(v)}||_1,$$

where $\gamma'_d = \frac{\gamma_d}{1+\alpha_1}$ and $\mathbf{Y}_{d(v)} = \frac{S_d f_{d(u)} + \alpha_1 \mathbf{y}_{d(v)}}{1+\alpha_1} - f_{c(v)}$.

Computation of $f_{c(v)}$

Similarly, solving Equation (7) with respect to $f_{c(v)}$ can be simplified as solving the following problem if $f_{d(v)}$ and $f_{d(u)}$ are fixed.

$$\mathcal{L}(f_{c(v)}) = ||\mathbf{Y}_{c(v)} - f_{c(v)}||_2^2 + \gamma'_c||f_{c(v)}||_1,$$

where $\mathbf{Y}_{c(v)} = \sum_{d=1}^{\delta}\left[\frac{\mathbf{S}_d \mathbf{f}_{d(u)} + \alpha_1 \mathbf{y}_{d(v)}}{(1+\alpha_1)\delta} - \frac{\mathbf{f}_{d(v)}}{\delta}\right]$ and $\gamma'_c = \frac{\gamma_c}{(1+\alpha_1)\delta}$.

Computation of $f_{d(u)}$

If $f_{c(v)}$ and $f_{d(v)}$ are fixed, solving Equation (7) with respect to each $f_{d(u)}$ can be simplified as solving the following *Lasso* problem:

$$\mathcal{L}(f_{d(u)}) = ||\mathbf{Y}_{d(u)} - f_{d(u)}||_2^2,$$

where $\mathbf{Y}_{d(u)} = \frac{S_d^T(f_{d(v)} + f_{c(v)}) + \alpha_2 \mathbf{y}_{d(u)}}{1+\alpha_2}$.

The complete derivation and outline of NetSML algorithm to solve Equation (7) is available in Supplementary Material. The algorithm iteratively updates $f_{c(v)}$ and $f_{d(v)}$, $f_{d(u)}$ for each cancer type. After convergence, the non-zero elements in $f_{c(v)}$ and $f_{d(v)}$ are common genomics features and cancer-specific features. The unlabeled samples in cancer type d are classified to different groups based on the sign of the final score for each sample in $f_{d(u)}$.

3 Experiments

In the experiments, we first generated four artificial datasets to test the NetML model on detecting common latent features and dataset specific features in coefficients vectors. Next, we performed three experiments on real cancer gene expression datasets (mRNA-sequencing data and microarray gene expression data). The first experiment was a cross-dataset analysis on breast cancer to show that the multi-task learning model can utilize information from other similar studies to improve outcome prediction. The second experiment was a cross-domain analysis on two major subtypes of lung cancer, lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD). The third experiment was also a cross-domain analysis on breast cancer, ovarian cancer, and prostate cancer.

3.1 Simulation

In this section, we generated four artificial datasets with the same number of features to test NetML of detecting common and data-specific discriminative features. The synthetic datasets were constructed as follows: we let $\delta = 4$, $\{n_d\}_{d=1}^{\delta} = 200$ and $m = 600$. If we represent the features in the dth domain by a matrix $\mathbf{Z}_d \in \mathbb{R}^{n_d \times m}$, then

$$Z_1(i,j) \sim \begin{cases} N(90,50), \text{if } 1 \le i \le 100 \text{ and } 1 \le j \le 150 \\ N(100,50), \text{if } 1 \le i \le 100 \text{ and } 551 \le j \le 600 \\ N(60,20), \text{otherwise,} \end{cases}$$

$$Z_2(i,j) \sim \begin{cases} N(90,50), \text{if } 1 \le i \le 100 \text{ and } 101 \le j \le 250 \\ N(100,50), \text{if } 1 \le i \le 100 \text{ and } 551 \le j \le 600 \\ N(60,20), \text{otherwise,} \end{cases}$$

$$Z_3(i,j) \sim \begin{cases} N(90,50), \text{if } 1 \le i \le 100 \text{ and } 201 \le j \le 350 \\ N(100,50), \text{if } 1 \le i \le 100 \text{ and } 551 \le j \le 600 \\ N(60,20), \text{otherwise,} \end{cases}$$

$$Z_4(i,j) \sim \begin{cases} N(90,50), \text{if } 1 \le i \le 100 \text{ and } 301 \le j \le 450 \\ N(100,50), \text{if } 1 \le i \le 100 \text{ and} 551 \le j \le 600 \\ N(60,20), \text{otherwise,} \end{cases}$$

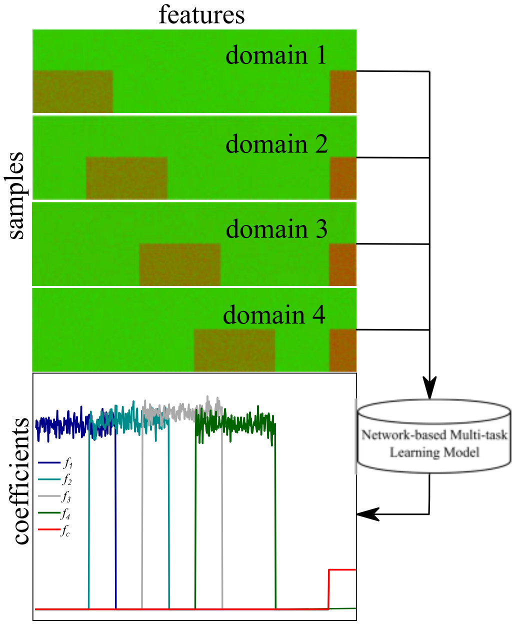

Fig. 3. Simulation experiments. The top panel shows the heat map of four artificial datasets. The bottom panel shows the corresponding coefficients vectors learned in NetML

The generated data are visualized in Figure 3. There are 600 features and 200 samples, and the first 100 samples and the second 100 samples are in two different classes in all the 4 datasets. The first 150 are true discriminative features in the first domain, the features between 101 and 250 are the true discriminative features in the second domain, the features between 201 and 350 are the true discriminative features in the third domain, and the features between 301 and 450 are the true discriminative features in the fourth domain. The features between 551 and 600 are the common discriminative features in all the domains. NetML was applied to both datasets with $\alpha = 0.01$, $\gamma_c = 1\mathrm{e}{-3}$ and $\gamma_d = 1\mathrm{e}{-3}$, and the learned coefficients f_1, f_2 and f_c were plotted in Figure 3. From the result we can see that, NetML can recover the hidden structures in the datasets and detect the common- and domain-specific features.

3.2 Experiments on cancer datasets

3.2.1 Analysis across breast cancer datasets

NetML and NetSML were first tested on two breast cancer microarray gene expression datasets: GSE6532 (Loi *et al.*, 2007) and GSE7390 (Desmedt *et al.*, 2007). The raw CEL files were downloaded from the GEO website and normalized by RMA (Irizarry *et al.*, 2003). After merging probes by gene symbols and removing probes with no gene symbol, a total of 13 261 unique genes derived from the 22 283 probes were included in this study. The patients were classified as cases and controls in the two datasets based on the time of developing distant metastasis. The patients who were free of metastasis for longer than 8 years of survival and follow-up time were classified as metastasis-free and the patients who developed metastases within 5 years were classified as metastasis cases. There are 96 metastasis-free patients and 51 metastasis samples in GSE6532 and 136 metastasis-free and 35 metastasis patients in GSE7390.

We applied NetML, *t*-test and Wilcoxon rank sum test to identify potential biomarkers from the two breast cancer datasets. To evaluate the prediction power of the marker genes we performed a 5-fold cross-validation on each of the two datasets with 3-folds for training, 1-fold for validation and 1-fold for test. We evaluated the classification performance using a support vector machine (SVM) with an RBF kernel. Specifically, we selected the markers genes by NetML based on non-zero entries in $f_1 + f_c$ and $f_2 + f_c$ on the training data in each dataset, then the parameters in the proposed models

Table 3. The mean AUROC and AUPRC scores of classifying patients in breast cancer datasets

Feature selection method	Classifier	GSE6532		GSE7390	
		AUROC	AUPRC	AUROC	AUPRC
NetML	SVM	0.707	0.555	0.671	0.592
t-test	SVM	0.623*	0.443*	0.546*	0.244*
Wilcoxon rank sum	SVM	0.650*	0.490*	0.567*	0.288*
NetSML		0.641	0.615	0.661	0.571
All the gene	DNN	0.610	0.450*	0.592*	0.228*

Note: *The difference between baseline and proposed methods is statistically significant (P-value < 0.01).

Table 4. The mean AUROC and AUPRC scores of classifying patients in lung cancer subtypes

Feature selection method	Classifier	LUAD		LUSC	
		AUROC	AUPRC	AUROC	AUPRC
NetML	SVM	1	1	1	1
t-test	SVM	0.990	0.963*	0.994	0.960*
Wilcoxon rank sum	SVM	0.990	0.963*	0.994	0.960*
NetSML		1	1	1	1
All the gene	DNN	0.991	0.942*	0.990	0.964*

Note: *The difference between baseline and proposed methods is statistically significant (P-value < 0.01).

and SVM were tuned based on the classification performance on validation data. In both NetML and NetSML, γ_c and γ_d were chosen from {1e−5, 1e−4, 1e−3}, and α was fixed to 0.01. The classification performance was evaluated on the test data in each dataset separately. We repeated the 5-fold cross-validation 50 times and report the mean of the AUROC and AUPRC scores in Table 3. Similarly, we selected the same numbers of features as in f_1+f_c and f_2+f_c based on P-values for both t-test and Wilcoxon rank sum test. The same 5-fold cross-validation setup was applied for the selected features in the baseline methods. Since NetSML is a semi-supervised learning model, and it can simultaneously select features and classify test samples, SVM was not applied to the identified features by NetSML. Instead, the classification performance of NetSML and a four layers fully connected feed-forward neural network model were compared and tested on the two datasets by using all the genes with the same 5-fold cross-validation. The mean of the AUROC and AUPRC scores in 50 repeats by applying NetSML and deep neural network (DNN) are also reported in Table 3. The NetML model outperformed CC in both datasets. NetML is clearly more capable of selecting more predictive marker genes in the experiments by learning the information from network structures in different cancer datasets for the same study purpose. In addition, our NetSML model outperformed DNN, since (i) NetSML learned more information from gene–sample bipartite structures in both test samples and the samples in the other dataset for the same study purpose; (ii) a general fully connected DNN may not work well on the datasets with high dimension but low-sample-size. It is also interesting to observe that without a feature selection step, the classification performance of NetSML was not as good as NetML + SVM in most cases.

3.2.2 Analysis across lung cancer subtypes

NetML and NetSML were also tested on two lung cancer subtypes. Both The Cancer Genome Atlas (TCGA; Weinstein *et al.*, 2013) LUAD and LUSC RNA-Seq gene expression datasets were downloaded from UCSC Xena Hub (Goldman *et al.*, 2018). The log2(x + 1) transformed RSEM normalized count was used in the analyses and 20 531 genes were included in this study. The patients in LUAD and LUSC have shorter survival time compared with breast cancer patients. So the patients in the two datasets were classified into cases and controls based on the samples were primary tumors or normal tissues instead of survival times. There are 517 primary tumors and 59 normal tissues in data LUAD, and 502 primary tumors and 51 normal tissues in data LUSC. The feature selection and cross-validation setups in this experiment were the same as the setups in Section 3.2.1. The mean AUROC and AUPRC scores are reported in Table 4. NetML outperformed t-test and Wilcoxon rank sum test, and NetSML also outperformed DNN. Since classifying samples into the tumor and normal tissue is a relatively easy task compared with predicting survival outcomes, the classification results in this experiment are much higher than the results in Section 3.2.1.

3.2.3 Analysis across cancer domains

We applied NetML and NetSML methods on three related cancer types, breast cancer, ovarian cancer, and prostate cancer, to detect common and cancer-specific differentially expressed genes and classify the patient samples into correct categories at the same time. All the TCGA breast cancer (BRCA), ovarian cancer (OV) and prostate cancer (PRAD) RNA-Seq gene expression datasets were also downloaded from UCSC Xena Hub with the same pre-processing step as in Section 3.2.2. There are 115 metastasis-free patients and 137 metastasis patients in breast cancer dataset, 92 metastasis-free samples and 106 metastasis patients in ovarian cancer dataset and 79 metastasis-free samples and 81 metastasis patients in prostate cancer dataset. The setups for feature selection and classification in this experiment were the same as the setups in Section 3.2.1. The mean AUROC and AUPRC scores are reported in Table 5. In breast cancer and prostate cancer datasets, NetML + SVM outperformed all the baseline methods and NetSML also beat DNN. In the ovarian dataset, NetSML outperformed DNN and the performance of NetML + SVM is similar to baseline methods. Overall, NetML+SVM and NetSML consistently achieved better or similar classification results compared with the baselines in all the experiments.

We performed a literature survey of each individual marker gene identified by NetML. Thirteen of them are supported by literature to be related to breast cancer, ovarian cancer, prostate cancer or all three as reported in Table 6. Gene CTBS, AKIRIN2 and FEXO4 identified in f_c were reported to play important roles in breast cancer, ovarian cancer, and prostate cancer. For example, AKIRIN2 codes a protein that targets BCAM, a suppressive oncogene in multiple cancers, including the three cancer types in this study (Akiyama *et al.*, 2013). FBXO4 mutation can affect Cyclin D1, a commonly dysregulated cyclin in prostate, breast and ovarian cancers (Qie and Diehl, 2016). It is clear that NetML identified signature genes are functionally coherent.

The top 100 common signature genes identified by NetML in three cancer types based on the values in f_c enriched in 130 GO (gene ontology) terms and three KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways (P-value < 0.05). The most significantly enriched GO functions are listed in Table 7. The similar enrichment analysis on the top 100 common signature genes identified by NetSML was also performed and the gene list enriched in 91 GP terms and two KEGG pathways, and the most significant ones are reported in Table 8. It is clear that the NetML models identified signature genes that are functionally coherent.

3.3 Running time

To measure the scalability of NetML and NetSML, we tested the algorithms on the two breast cancer datasets in Section 3.2.1. The two datasets contain 147 and 171 breast cancer samples, respectively. NetML took 320 CPU seconds to run the experiment with one pair of parameters to identify the signature genes. NetSML took 256 CPU seconds to run the experiment with one pair of parameters to select the signature genes and classify the unlabeled test samples. The CPU time was measured on Intel Core i7-7700CPU with 3.60 GHz.

4 Discussion and conclusion

Application of multi-task learning on gene expression analysis across multiple cancer types/subtypes is promising since the

Table 5. The mean AUROC and AUPRC scores of classifying patients in BRCA, OV and PRAD

Feature selection method	Classifier	BRCA		OV		PRAD	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
NetML	SVM	0.649	0.622	0.571	0.619	0.708	0.640
t-test	SVM	0.602	0.568*	0.581	0.571	0.613*	0.581*
Wilcoxon rank sum	SVM	0.645	0.610	0.598	0.602	0.635*	0.603
NetSML		0.636	0.658	0.623	0.664	0.686	0.625
All the gene	DNN	0.573*	0.522*	0.585	0.574*	0.605*	0.561*

Note: *The difference between baseline and proposed methods is statistically significant (*P*-value < 0.01).

Table 6. Literature review of the candidate cancer genes

Cancer type	Gene names	References	Description
BRCA	SAV1	Chen *et al.* (2014)	SAV1 is part of the Hippo pathway responsible for mammary gland regulation and is knocked out in breast cancer samples.
	IRF2	Doherty *et al.* (2001)	IRF2 is identified as being over-expressed in human breast cancer tissue compared with normal adjacent tissue.
	TANK	Wei *et al.* (2014)	TANK is the target of TBK-1, which is over-expressed in breast cancer tissue.
	MKRN1	Wang *et al.* (2013)	MKRN1 codes for an E3 ligase that regulates p21 protein level, where p21 controls breast cancer cell proliferation.
OV	CYP2R1	Downie *et al.* (2005)	CYP2R1 is expressed at a significantly higher level in primary ovarian cancer compared with a normal ovary.
	ADIPOR2	Tiwari *et al.* (2015)	ADIPOR2 is expressed at a significantly lower rate in cancerous ovaries.
PRAD	WDR5	Kim *et al.* (2014)	WDR4 is identified as being over-expressed in prostate cancer cells and significant for cancer cell proliferation.
	NASP	Amundson *et al.* (2008)	NASP is a significant prognostic marker in prostate cancer cells.
	MTA1	Kai *et al.* (2010)	MTA1 is over-expressed in prostate cancer and is associated with aggressive metastasis.
All three	CTBS	Varley *et al.* (2014)	CTBS-GNG5 fusion is prevalent in breast cancer cells.
		Plebani *et al.* (2012)	CTBS-GNG5 fusion was found in ovarian and prostate cells lines.
	AKIRIN2	Akiyama *et al.* (2013)	AKIRIN2 codes a protein that targets BCAM, a suppressive oncogene in multiple cancers.
	FBXO4	Qie and Diehl (2016)	FBXO4 mutation can affect Cyclin D1, a commonly dysregulated cyclin in prostate, breast and ovarian cancers.

Table 7. Enriched GO terms by the common signature genes in NetML

GO:0043170:macromolecule metabolic process
GO:0042058:regulation of epidermal growth factor receptor signaling pathway
GO:1901184:regulation of ERBB signaling pathway
GO:0006139:nucleobase-containing compound metabolic process
GO:0008152:metabolic process
GO:0090304:nucleic acid metabolic process
GO:0042059:negative regulation of epidermal growth factor receptor signaling pathway
GO:1901185:negative regulation of ERBB signaling pathway
GO:0046483:heterocycle metabolic process
GO:0006725:cellular aromatic compound metabolic process
GO:0006351:transcription, DNA-templated
GO:0080134:regulation of response to stress
GO:0006807:nitrogen compound metabolic process
GO:0015031:protein transport
GO:0051276:chromosome organization
GO:0071704:organic substance metabolic process
GO:0007173:epidermal growth factor receptor signaling pathway
GO:0034641:cellular nitrogen compound metabolic process
GO:1901360:organic cyclic compound metabolic process
GO:0033554:cellular response to stress

Table 8. Enriched GO terms by the common signature genes in NetSML

GO:0032561:guanyl ribonucleotide binding
GO:0019001:guanyl nucleotide binding
GO:0005525:GTP binding
GO:0071496:cellular response to external stimulus
GO:0043281:regulation of cysteine-type endopeptidase activity involved in apoptotic process
GO:0052548:regulation of endopeptidase activity
GO:0005737:cytoplasm
GO:0010950:positive regulation of endopeptidase activity
GO:0044444:cytoplasmic part
GO:0052547:regulation of peptidase activity
GO:2000116:regulation of cysteine-type endopeptidase activity
GO:0006919:activation of cysteine-type endopeptidase activity involved in apoptotic process
GO:0010952:positive regulation of peptidase activity
GO:0046033:AMP metabolic process
GO:1902494:catalytic complex
GO:0043280:positive regulation of cysteine-type endopeptidase activity
GO:0043231:intracellular membrane-bounded organelle
GO:0044424:intracellular part
GO:0043227:membrane-bounded organelle
GO:2001056:positive regulation of cysteine-type endopeptidase activity

deregulation of gene expression is a hallmark of human tumor cells. NetML and NetSML utilize multiple cancer domains/studies for detecting common and domain-specific differentially expressed genes. The comparison to *t*-test, Wilcoxon rank sum test and DNN, which essentially represents single-task learning methods, suggests that multi-task learning enables sharing information in domains of different cancer studies to discover hidden structures from the biological networks that can explain common and domain-specific cancer characteristics and better classify patient samples as shown in the experiments.

It is interesting that several deep learning approaches (Danaee *et al.*, 2017; Khoshghalbvash and Gao, 2019; Lyu and Haque, 2018; Way and Greene, 2018) have been proposed for cancer detection and biomarker identification. It was shown that the deep learning models can improve cancer outcome predictions. Our experiments have shown that the performance of our multi-task semi-supervised learning models is comparable to the deep learning approach by utilizing network information and training the data from different domains. The observation suggests that multi-task learning might be a more effective framework for mRNA expression analysis. With the recent TCGA, TARGET and LINCS research programs, more and more large-scale mRNA expression datasets are becoming available for different cancer types. It is expected that multi-task learning with biological networks as prior knowledge will play an important role in cancer transcriptome analysis with large patient cohorts to improve cancer biomarker detection and cancer phenotype predictions.

Acknowledgements

We thank Dr. Rui Kuang at University of Minnesota Twin Cities, for his valuable support and advice. The results are based upon data generated by The Cancer Genome Atlas (TCGA) established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov. The dbGaP accession number to the specific version of the TCGA dataset is phs000178.v8.p7.

Funding

This work was supported by grant from the National Science Foundation [IIS 1755761 and FET 1908495].

Conflict of Interest: none declared.

References

Akiyama,H. *et al.* (2013) The FBI1/Akirin2 target gene, BCAM, acts as a suppressive oncogene. *PLoS One*, **8**, e78716.

Amundson,S.A. *et al.* (2008) Integrating global gene expression and radiation survival parameters across the 60 cell lines of the National Cancer Institute Anticancer Drug Screen. *Cancer Res.*, **68**, 415–424.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Breitling,R. *et al.* (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.

Cao,Z. and Zhang,S. (2016) An integrative and comparative study of pan-cancer transcriptomes reveals distinct cancer common and specific signatures. *Sci. Rep.*, **6**, 33398.

Chen,Q. *et al.* (2014) A temporal requirement for Hippo signaling in mammary gland differentiation, growth, and tumorigenesis. *Genes Dev.*, **28**, 432–437.

Chung,F.R. and Graham,F.C. (1997). *Spectral Graph Theory, Number 92.* Regional Conference Series in Math. *CBMS, American Mathematical Soc.*

Danaee,P. *et al.* (2017). A deep learning approach for cancer detection and relevant gene identification. In: *Pacific Symposium on Biocomputing 2017*, vol. **22**, pp. 219–229. World Scientific.

Desmedt,C. *et al.* (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.*, **13**, 3207–3214.

Díaz-Uriarte,R. and Alvarez de Andrés,S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Ding,C. and Peng,H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.*, **3**, 185–205.

Doherty,G.M. *et al.* (2001) Interferon regulatory factor expression in human breast cancer. *Ann. Surg.*, **233**, 623.

Downie,D. *et al.* (2005) Profiling cytochrome P450 expression in ovarian cancer: identification of prognostic markers. *Clin. Cancer Res.*, **11**, 7369–7375.

Goldman,M. *et al.* (2018) The UCSC Xena Platform for cancer genomics data visualization and interpretation. *BioRxiv*, 326470.

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Jafari,P. and Azuaje,F. (2006) An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Med. Inform. Decis. Mak.*, **6**, 27.

Kai,L. *et al.* (2010) Resveratrol enhances p53 acetylation and apoptosis in prostate cancer by inhibiting MTA1/NuRD complex. *Int. J. Cancer*, **126**, 1538–1548.

Khoshghalbvash,F. and Gao,J.X. (2019). Integrative feature ranking by applying deep learning on multi source genomic data. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 207–216. ACM.

Kim,J.-Y. *et al.* (2014) A role for WDR5 in integrating threonine 11 phosphorylation to lysine 4 methylation on histone H3 during androgen signaling and in prostate cancer. *Mol. Cell*, **54**, 613–625.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Loi,S. *et al.* (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J. Clin. Oncol.*, **25**, 1239.

Lyu,B. and Haque,A. (2018). Deep learning based tumor type classification using gene expression data. In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 89–96. ACM.

Makhijani,R.K. *et al.* (2018) Identification of common key genes in breast, lung and prostate cancer and exploration of their heterogeneous expression. *Oncol. Lett.*, **15**, 1680–1690.

Petegrosso,R. *et al.* (2016) Transfer learning across ontologies for phenome–genome association prediction. *Bioinformatics*, **33**, 529–536.

Plebani,R. *et al.* (2012) Long-range transcriptome sequencing reveals cancer cell growth regulatory chimeric mRNA. *Neoplasia*, **14**, 1087.

Qie,S. and Diehl,J.A. (2016) Cyclin D1, cancer progression, and opportunities in cancer treatment. *J. Mol. Med.*, **94**, 1313–1326.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Thomas,J.G. *et al.* (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Tiwari,A. *et al.* (2015) Expression of adiponectin and its receptors is altered in epithelial ovarian tumors and ascites-derived ovarian cancer cell lines. *Int. J. Gynecol. Cancer*, **25**, 399–406.

Van't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530.

Varley,K.E. *et al.* (2014) Recurrent read-through fusion transcripts in breast cancer. *Breast Cancer Res. Treat.*, **146**, 287–297.

Wang,Z. *et al.* (2013) RNF115/BCA2 E3 ubiquitin ligase promotes breast cancer cell proliferation through targeting p21Waf1/Cip1 for ubiquitin-mediated degradation. *Neoplasia*, **15**, 1028.

Way,G.P. and Greene,C.S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, Vol. **23**, p. 80. NIH Public Access.

Wei,C. *et al.* (2014) Elevated expression of TANK-binding kinase 1 enhances tamoxifen resistance in breast cancer. *Proc. Natl. Acad. Sci. USA*, **111**, E601–E610.

Weinstein,J.N. *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113.

Winter,C. *et al.* (2012) Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput. Biol.*, **8**, e1002511.

Xing,E.P. *et al.* (2001). Feature selection for high-dimensional genomic microarray data. In *Icml*, Vol. **1**, pp. 601–608. Citeseer.

Zhang,H. *et al*. (2013). Transfer learning across cancers on DNA copy number variation analysis. In: *2013 IEEE 13th International Conference on Data Mining*, pp. 1283–1288. IEEE.

Zhang,W. *et al*. (2010). Network propagation models for gene selection. In: *Genomic Signal Processing and Statistics (GENSIPS), 2010 IEEE International Workshop on*, pp. 1–4. IEEE.

Zhang,W. *et al*. (2012). Signed network propagation for detecting differential gene expressions and DNA copy number variations. In: *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 337–344. ACM.

Zhang,W. *et al*. (2017) Network-based machine learning and graph theory algorithms for precision oncology. *NPJ Precis. Oncol*., **1**, 25.