Identifying Relevant microRNAs in Bladder Cancer using Multi-Task Learning

Adriana Birlutiu[(1),(2)], Paul Bulzu[(1)], Irina Iereminciuc[(1)], and Alexandru Floares[(1)]

(1) SAIA & OncoPredict, Cluj-Napoca, Romania

(2) "1 Decembrie 1918" University of Alba-Iulia, Romania

adriana.birlutiu@saia-institute.org, paul.bulzu@saia-institute.org, irina.iereminciuc@saia-institute.org, alexandru.floares@saia-institute.org

Abstract. We present a strategy based on multi-task learning to identify relevant cancer genes across three types of bladder cancer. Our objective in this study was to identify microRNAs which are relevant in discriminating between cancer and normal samples across three types of bladder cancer. We used machine learning techniques that learn the features which are relevant and in the same time build a classifier that can discriminate between cancer and normal samples. Experimental comparison showed that the performance of multi-task learning improves upon the performance of single-task learning. Applying the algorithms, we obtained a small set of microRNAs that are relevant in discriminating between cancer and normal samples and we further investigated their biological functions.

Keywords: multi-task learning, regularization, prediction, cancer

1 Scientific Background

Despite significant efforts, cancer is still a lethal disease with a high mortality rate. Cancer is known to be as a highly heterogeneous disease specific to cell type and tissue origin. Bladder cancer is the fourth most commonly diagnosed malignancy in men and it is a burdensome disease with significant costs and mortality. The biology of bladder cancer is incompletely understood, making the management of this disease difficult. The majority of tumors are urothelial cell carcinomas (UCC). Most UCCs belong to a low grade pathway, around one-third are high grade in differentiation and arise as lesions initially confined to the bladder tissue (non-muscle invasive). Progression to muscle invasion occurs in around 50% of high-grade lesions.

The development of novel biomarkers for bladder cancer could significantly improve clinical outcomes and decrease health related costs. Recent evidence suggests a regulatory role for microRNA in bladder cancer [1]. MicroRNAs are small, non-coding RNAs, mainly involved in the negative regulation of gene expression at the post-transcriptional and translational levels.

Our objective in this study was to identify which microRNAs are relevant in discriminating between UCCs and normal samples across three types of bladder cancer: low-grade, high-grade and muscle invasive. We used machine learning techniques that learn from data which features are relevant and in the same time build a classifier that can discriminate bewteen UCCs and normal samples. Computational intelligence and machine learning techniques are tools that are starting to be used in the emerging field of cancer systems biology. Multi-task learning is a machine learning technique that has been previously used in learning contexts where data is available from multiple scenarios [6, 8, 4]. Using the staging of bladder cancer, we consider three learning scenarios for which we apply the multi-task learning formalism. We compared the performance obtained with multi-task learning versus single-task learning and observed that indeed multi-task learning improves the performance upon single-task learning.

A recent work, related to ours, is [4], which finds core cancer genes across multiple cancers. In contrast to this work, we use other machine learning and multi-task learning techniques for finding the relevant features.

2 Materials and Methods

2.1 Data Set

We used the bladder cancer data published in [2] and available on the Internet. It consists of expression values of 333 microRNAs from 78 samples, where 72 samples were taken from patients and 6 belong to cell lines. Given the differences in cell lines, we further considered only the 72 samples from patients. The normal group contains a total of 20 samples and were obtained from 10 subjects with normal urothelium (taken distant to any tumor) and from 10 disease-free controls. The cancer group contains the rest of 52 samples and were sub-classified into three subgroups: *(i)* low-grade non-muscle invasive– 22 samples, *(ii)* high-grade non-muscle invasive– 12 samples and *(iii)* muscle invasive– 18 samples.

2.2 Data Representation and Preprocessing

Let T represent the number of tasks and d the number of features. The input data is represented by the matrices X^t, $t = 1, \ldots, T$, where each X^t is an $n_t \times d$ matrix. The output data is represented by Y^t, $t = 1, \ldots, T$, where each Y^t is a binary vector of dimension n_t, one class representing normal samples and the other cancer.

Based on bladder cancer staging, we defined 3 learning tasks as follows (each task is framed as a binary classification problem that discriminates between cancer versus non-cancer): *1)* Task 1: low-grade non-muscle invasive cancer patients versus normal patients. *2)* Task 2: high-grade non-muscle invasive cancer patients versus normal patients. *3)* Task 3: muscle invasive cancer patients versus normal patients.

From the group of normal patients we further considered only the disease-free controls, and not the subjects with normal urothelium taken distant to any tumor. We did this since we obtained better results this way.

Approximately 20% from the total of microRNA values are missing. We used the Matlab *knnimpute* function to impute these missing values. *knnimpute* replaces missing values in data with the corresponding value from the nearest-neighbor column using Euclidean distance.

2.3 Single-Task Learning

The problem we want to solve can be reduced to a binary classification problem. We want to build a classifier that can discriminate between normal and cancer samples and in the same time identify a subset of relevant features (microRNAs).

We use Lasso which is a linear model that estimates sparse coefficients. It is useful in contexts in which the dimension d of data is large. Lasso's tendency is to prefer solutions with fewer parameter values, effectively reducing the number of variables upon which the given solution is dependent. Mathematically, it consists of a linear model trained with l_1 norm as regularizer [7]. This regularization acts as a prior information which favors sparse solutions.

$$\min_{\boldsymbol{w},c} \sum_{j=1}^{n} \log(1 + \exp(-Y_j(\boldsymbol{w}\boldsymbol{X}(:,j) + c))) + \rho_1 \sum_{i=1}^{d} |w_i|$$

$\boldsymbol{w}$ is a parameter vector of the same dimension as the number of features, the regularization parameter ρ_1 controls sparsity.

In the equation from above instead of $\sum_{i=1}^{d} |w_i| = ||\boldsymbol{w}||_1$, other norms can be used, for example the $l_{2,1}$ which is the column grouped l_1, i.e., the group sparsity learning [6].

2.4 Multi-Task Learning

We extend the single-task learning scenario to a multi-task learning by considering the three classification tasks for bladder cancer staging defined above.

The basic idea in multi-task learning is that models learned on different scenarios have parts in common. In multi-task learning, the tasks are learned simultaneously by extracting and utilizing appropriate shared information across tasks. In a Bayesian framework this often boils down to the sharing of a hierarchical prior [8], while in other learning settings a regularizer term across tasks is being used [6]. In this study we consider the latter situation and we used the l_1-norm which can be extended to the multi-task learning scenario as follows:

$$\min_{\boldsymbol{W},c} \sum_{t=1}^{T} \sum_{j=1}^{n} \log(1 + \exp(-Y_j(\boldsymbol{W}(:,t)\boldsymbol{X}(:,j) + c_t))) + \rho_1 ||\boldsymbol{W}||_1$$

Here $\boldsymbol{W}$ is a matrix, each column of $\boldsymbol{W}$ represents the model of a single task.

Other norms can be used instead of l_1-norm, for example the $l_{2,1}$-norm. In this case the way to capture the task relatedness from multiple related tasks is to constrain all models to share a common set of features [6].

2.5 Experimental Protocol

We divided the data into training and testing using stratified 5-fold cross-validation. Each of the 5 subsamples has roughly equal size and roughly the same class proportions as in the entire data set. We further used again 5-fold cross-validation but this time on the training set to find the optimal parameter settings. We evaluated the performance on the testing set using as evaluation measure: AUC (area under the receiver operator curve) and accuracy (percentage of correct prediction made of the model over the data set).

We performed the experimental evaluation in Matlab, using a toolbox called Malsar [3] which implements several algorithm for multi-task learning.

3 Results

3.1 Single vs multi-task learning

The results in Table 1 compare single-task versus multi-task learning using two regularizations (l_1 norm and $l_{2,1}$ norm) and two evaluation measures (AUC and accuracy). The results show the mean and standard deviation over 10 runs of the entire experimental protocol.

The multi-task scenario takes into account the 3 tasks defined above. In the multi-task learning scenario, the models of each of the tasks are learned simultaneously, with data from the other task influencing the learned model of a certain task. In single-task learning the models of different tasks are learned independently, and there is no transfer of information between tasks.

The results show the mean values. The best results have an AUC and accuracy of 1 and are consistent with our previous study [5]. In [5] we optimized a single-task learning using decision trees, here we show that using multi-task learning we can improve the performance upon single-task learning.

3.2 Feature selection

Figure 1 shows the number of feature selected in the models, as a function of the regularizer parameter: left: l_1 norm, right: $l_{2,1}$ norm.

We determined a small list of microRNAs that have been repeatedly selected by the algorithms as relevant for discriminating between normal and cancer samples. These microRNAs are: *hsa-miR-133b, hsa-miR-135b, hsa-miR-204, hsa-miR-30a-3p, hsa-miR-411, hsa-miR-564*. We noticed that l_1 norm classifier gives better performance with a small subset of classification features compared to $l_{2,1}$ norm classifier.

3.3 Biological Interpretation

We further applied several functional analysis methods on the list of six relevant microRNAs in order to identify their potential roles in carcinogenesis. Known functional roles and disease involvement of each microRNA were investigated by

Table 1. Comparison between multi-task learning using two regularizations (l_1 norm and $l_{2,1}$ norm) and two evaluation measures (area under curve and accuracy). The results show the mean and standard deviation over 10 runs of the entire experimental protocol.

	Area under curve		
$l_{2,1}$ norm	Task 1	Task 2	Task 3
Single-task	0.961±0.036	0.952±0.050	0.916± 0.036
Multi-task	0.980±0.014	0.990±0.011	0.954±0.010
l_1 norm			
Single-task	0.956±0.032	0.898±0.065	0.918±0.024
Multi-task	0.983±0.014	0.973±0.047	0.944±0.017

	Accuracy		
$l_{2,1}$ norm	Task 1	Task 2	Task 3
Single-task	0.927± 0.060	0.868± 0.062	0.858±0.043
Multi-task	0.928±0.017	0.914±0.037	0.902±0.033
l_1 norm			
Single-task	0.911±0.046	0.858± 0.033	0.847±0.040
Multi-task	0.904±0.040	0.896±0.033	0.886± 0.043

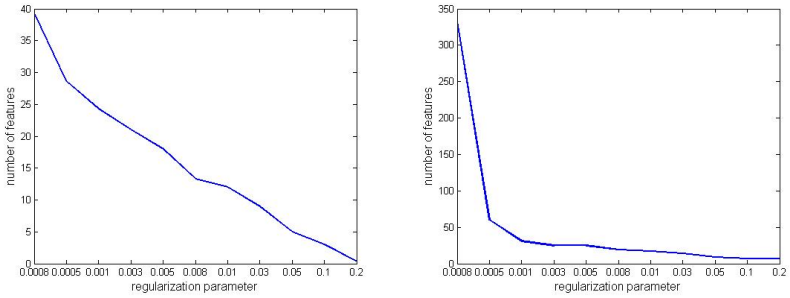

Fig. 1. Number of feature selected in the models, as a function of the regularizer parameter. Left: l_1 norm, right: $l_{2,1}$ norm.

consulting the miRcancer database [9]. Afterwards, we conducted an integrated analysis of all relevant microRNAs by using two web-based ontology enrichment analysis applications: MetaCore^{TM} and ToppGene [10].

The miRcancer database shows that *hsa-miR-133b* has been commonly identified as being down-regulated in bladder cancer. This can lead to increased cell proliferation, migration and invasion by lowering the inhibitory effect that *hsa-miR-133b* normally has on the epidermal growth factor receptor and its downstream proteins. In the case of *hsa-miR-135b*, no information related to bladder cancer was available. However, this microRNA is frequently found overexpressed in colorectal, gastric and lung cancers and it is known to favor metastasis. *Hsa-mir-204*, *hsa-mir-30a-3p*, *hsa-miR-564* and *hsa-mir-411* are known to be down-regulated in most malignancies in which they have been investigated thus preventing their normal tumor suppressor activity [9].

Integrated functional analysis for the six microRNAs was first performed in the MetaCoree^{TM} application. The Expand-by-one algorithm was first used to obtain a molecular interaction network by automatically adding known targets for each microRNA. We then exported the list of molecules from the network and analyzed it using the function enrichment analysis tools in MetaCore^{TM} and ToppGene. All results indicated strong implications in cancer for the investigated microRNAs and their targets (see Figure2). The most relevant signaling pathways identified were associated with apoptosis regulation, the epithelial-to-mesenchymal transition and cytoskeleton remodeling.

4 Conclusion

We presented an approach based on multi-task learning to identify relevant cancer cellular microRNA across three types of bladder cancer (low-grade non-muscle invasive, high-grade non-muscle invasive and muscle invasive). We used neural networks and the Lasso and group regularization techniques that learn the relevant features and build a classifier that can discriminate between three types

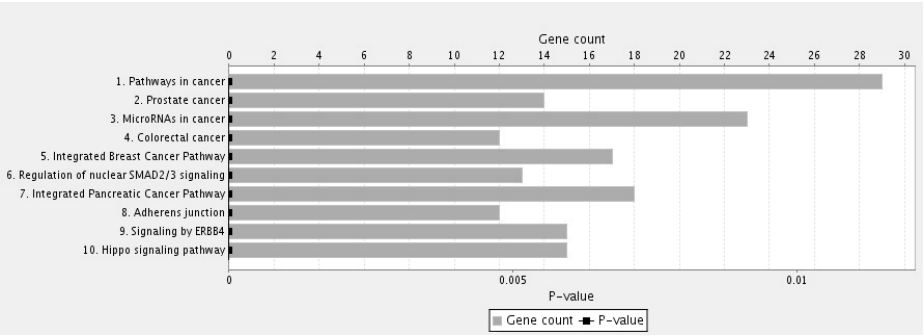

Fig. 2. The highest ranking pathways associated with the informative microRNA list and their targets identied using ToppGene. The microRNA list consists of: *hsa-miR-133b, hsa-miR-135b, hsa-miR-204, hsa-miR-30a-3p, hsa-miR-411, hsa-miR-564*.

of cancer and normal samples. The experimental evaluation showed that the performance of multi-task learning improves upon the performance of single-task learning. The results of the functional analysis performed on the six informative microRNAs have revealed strong relations to molecules and pathways already known to be involved in cancer.

Acknowledgments

We acknowledge funding through the program PN II, developed with the support of ANCS, CNDI - UEFISCDI, Romania, project no. PN-II-PT-CACM-2011-3.1-1221

References

1. L. Adam, M.F. Wszolek, C.G. Liu, W. Jing, L. Diao, A. Zien, J.D. Zhang, D. Jackson, C.P. Dinney. "Plasma microRNA profiles for bladder cancer detection" *Urol Oncol*, Nov;31(8):1701-8, 2013.
2. W.F. Catto, S. Miah and H.C. Owen and et. al "Distinct microRNA alterations characterize high and low grade bladder cancer" *Cancer Research*, vol. 69, no. 21, pp. 84772-81, 2008.
3. J. Zhou, J. Chen and J. Ye. "MALSAR: Multi-tAsk Learning via StructurAl Regularization" Arizona State University, *http://www.public.asu.edu/ jye02/Software/MALSAR*, 2012.
4. S. Gao, S. Xu, Y. Fang, J. Fang. "Prediction of core cancer genes using multi-task classification framework" *J Theor Biol.*, Jan 21;317:62-70, 2013.
5. A. Floares, A. Birlutiu. "Decision tree models for developing molecular classifiers for cancer diagnosis" *International Joint Conference on Neural Networks (IJCNN)*, Brisbane, Australia, June 10-15, 2012.
6. A. Argyriou, T. Evgeniou, M. Pontil. "Convex multi-task feature learning" *Machine Learning*, 73, 243272, 2008.

7. R. Tibshirani. “Regression shrinkage and selection via the Lasso” *Journal of the Royal Statistical Society*. Series B (Methodological), 267288, 1996.
8. A. Birlutiu, P. Groot, T. Heskes “Multi-task preference learning with an application to hearing aid personalization” *Neurocomputing*, 73 (7), 1177-1185, 2010.
9. B. Xie, Q. Ding, H. Han, D. Wu “miRCancer: a microRNA-cancer association database constructed by text mining on literature” *Bioinformatics*, 1;29(5):638-44, 2013.
10. J. Chen, E. Bardes, B. Aronow, A. Jegga “Toppgene suite for gene list enrichment analysis and candidate gene prioritization.“ *Nucleic Acids Research*, 37(Web Server issue):W305-11, 2009.