Computational methods for cancer survival classification using intermediate information

Shinuk Kim[1,2,3*], Taesung Park[2], Mark Kon[1,3,*]

[1]Bioinformatics program, Boston University, Boston, MA 02215, USA

{kshinuk,mkon}@bu.edu

[2]Department of Statistics, Seoul National University, Seoul 151-747 Republic of Korea

{taesungpark}@statis.snu.ac.kr

[3] Department of Mathematics and Statistics, Boston University, Boston, MA 02215 USA

{mkon}@bu.edu

Abstract. We study a potentially useful methodology based on machine learning (ML) involving integration of separate biomarker classes, to improve prediction and separation of ovarian cancer survival times. We also imported intermediate survival information for separating extreme two groups. For prediction of survival phenotypes, we use four classifiers, first two existing machine learning methods (support vector machine, SVM; random forest, RF), the second a new regression-based method (REG) feature selection together with Cox proportional hazards model (FSCR), FSCR_REG, the third SVM-based classifier using FSCR data sets (FSCR_SVM). We compared these four methods using three types of cancer tissue features: i) miRNA expression, ii) mRNA expression, and iii) integrated miRNA and mRNA expression information, the latter with features selected separately from miRNAs and mRNAs profiles. The accuracies of survival classification using the combined miRNA/mRNA profiles are higher than those using miRNA or mRNA alone . The latter differences indicate sometimes strong interactions between miRNA and mRNA features which are not visible in individual analyses.

1 Introduction

Ovarian cancer is the fifth leading cause of death from gynecological malignancy in the United States and Western Europe [1], [2], [3]. The typically advanced stages of ovarian cancer at initial diagnosis have been a large contributing factor to the high mortality rate of this disease [2], [3]. According to cancer statistics, 75% of patients with ovarian cancer are commonly diagnosed at an advanced stage, for which the 5 year survival rate is only 5% to 30%, with an average survival time of 21 months [1], [4]. On the other hand, among patients diagnosed early, the 5 year survival rate exceeds 90% [1].

Many studies have been proposed suggesting better-performing molecular data for cancer-related classification to stratify patients for treatments [5], [6], [7] but there still remain a number of unresolved questions about their mutual relationship. On the other hands, the accuracy of classification based only on miRNA expression, which Lu [5] presented, was better than that based on mRNA. In contrast, using the same dataset, Peng [6] suggested that the result using mRNA is superior to that using miRNA for the same cancer classification problem. Regarding to this, we study three cancer data sets for prediction of survival time with: (1) use of mRNA expression profiles only, (2) use of miRNA expression profiles only and (3) use of both mRNA and miRNA gene expression profiles. In all three cases we assessed the quality of these features as predictors of phenotypes, in this case patient survival times. We have implemented two different methods to integrate information for predicting cancer survival times, for the above three data types. The first is a well-known classification algorithm, the support vector machine (SVM) [8], and random forest (RF) [9] based on a discretization of survival times (into two classes), while the second is a regression-based algorithm using feature selection with Cox proportional hazards model [10] denoted FSCR, and a SVM-based algorithm with FSCR based on continuous survival information. Our approaches for predicting survival times use machine learning protocols, which allow transparent combinations of the information types, miRNA and mRNA profiles. In principle, the integration of molecular information types in ML can be done using the standard machine learning method of kernel addition, with no limitation of number of data types. This means taking kernel matrices representing multiple data sources (e.g. mRNA and miRNA profiles), and adding their kernels to represent combined information. For biomarker-based prediction, kernel addition is a simple modular method for integrating different information sources for predicting cancer phenotypes.

2 Materials and Methods

2.1 Materials

All data were obtained from The Cancer Genome Atlas (TCGA available at http://cancergenome.nih.gov/), a source of standardized and comprehensive cancer data sets. We downloaded second updated gene expression data (AgilentG4502A) and miRNA expression data (Agilent miRNA_8x15K) provided by the University of North Carolina. We obtained data from 147 ovarian cancer samples, including 22 long (greater than 5 years) and 22 short (less than 1 year) survival time samples. Total feature numbers of mRNAs and miRNAs used were 17,814 and 799 respectively. Up to date, the data sets have been aggregated since then, however we presented the implementation results from our originally downloaded data sets.

2.2 Methods for classification

To investigate prediction of cancer survival subtypes, we analyzed three different ovarian cancer datasets, involving: (1) miRNA expression profiles, (2) mRNA expression profiles and (3) combinations of miRNA/mRNA expression profile data sets. In the case (3), we selected features for the 22 long and 22 short survival samples: based on individual feature selection from miRNA and mRNA expression profiles. The initial implementation was based on feature selection using the Fisher criterion score i.e. $|\mu_a - \mu_b|^2 / (\sigma_a^2 + \sigma_b^2)$, where μ and σ represent the mean and standard deviation of a given mRNA (miRNA) in one class, while and represent the same in a second class.

For each of the above datasets (1)-(3), we used four different classification methods: (A) RF and SVM-based machine learning algorithm, and (B) a modified regression analysis method involving Cox regression, based on initial feature selection (FSCR_REG) (C) a SVM-based machine learning algorithm using gene expression levels multiplied by Cox coefficient (FSCR_SVM). All implementations were performed using leave-one-out cross-validation, but RF was performed 5 fold-cross-validation. We will first describe the SVM protocol (A) above. For the two classes of 22 longest and 22 shortest surviving patients, we selected gene sets consisting of the *n* most significantly differentially expressed genes using Fisher criteria scores. For feature set sizes *n* = 1 to 100, we tested the accuracy of differentiation among the two classes, to obtain optimal numbers of features for classification. Under the leave-one-out protocol, we predicted the class (either long or short survival time) of each test sample using SVM based on the selected miRNA and/or mRNA expression features as predictors. In the case of FSCR_REG method (B), we again selected features using Fisher criterion scores differentiating the 22 short and 22 long survival samples (excluding left out test samples), but computed Cox proportional hazard regression coefficients using survival data for *all* patients. We computed *Cox risk scores* as linear combinations of the selected feature expression levels, weighted by multivariate Cox proportional hazard regression coefficients. We then computed survival predictors using the above Cox coefficients of test sample expression levels using as a threshold the overall mean of the 43 training data for each sample left out. The formulation of Cox risk score of *i* th sample is given as:

$$Cox\ risk\ score = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} \qquad \textbf{(1)}$$

where the $\beta's$ are Cox coefficients, x_{ki} is k th gene expression of i th sample.

In the case of FSCR_SVM method (C), we performed same procedural of (B) and then used SVM classifier for new matrix data sets generated by multiplying gene expression level and Cox coefficients using all data sets including intermediated data sets. The matrix form is followed.

$$\left[M_{ij}\right] = \left[\beta_j x_{ij}\right] \qquad for\ i = 1, \cdots, n \quad j = 1, \cdots, m \qquad \textbf{(2)}$$

where β's are Cox coefficients, and x_{ij} are j the gene expression of i th sample and M_{ij} is n by k matrix. Figure 1 shows a schematic diagram of FSCR methods.

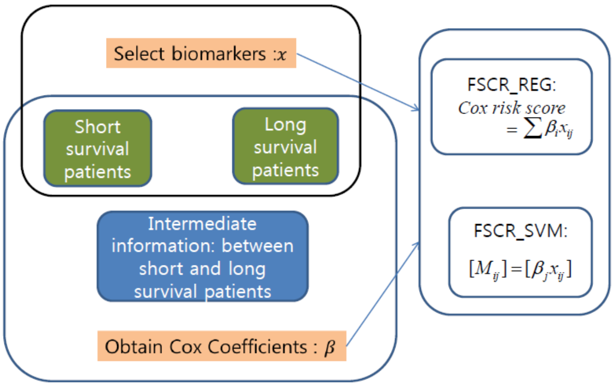
Select biomarkers : x
Short survival patients
Long survival patients
Intermediate information: between short and long survival patients
Obtain Cox Coefficients : β
FSCR_REG:
Cox risk score $= \sum \beta_i x_{ij}$
FSCR_SVM:
$[M_{ij}] = [\beta_j x_{ij}]$

Fig. 1. Schematic diagram of FSCR_REG, and FSCR_SVM methods. Gene were selected using short and long survival data sets. Cox coefficients were computed using all data sets including intermediate data sets with long and short survival data sets.

3 Results/Discussion

3.1 Comparing three different data types using SVM with feature selection.

In order to find optimal numbers of features for discriminating long and short survival phenotypes, we used Fisher criterion scores (Materials and Methods). To classify patients as long- or short-term survival, we implemented leave-one-out cross validation using SVM, with Fisher feature selection, using three different data types: (1) miRNA expression profile data, (2) mRNA expression profile and (3) a dataset generated by the combination of two data sets. The algorithm performed with 75% accuracy in balanced datasets using 60 features only from miRNA expression profiles, and 63.64% using five features only from mRNA. We note that feature selection was done entirely independently of the test samples here, including in the leave-one-out cross-validation (LOOCV) tests.

To understand better the interaction of the mRNA and miRNA biomarkers, we selected a fixed number *n* of mRNA markers and *m* of miRNA markers yielding a pair (*n*, *m*), with *n*, *m* ranging from 1 to 40 in all combinations. For each choice (*n*, *m*), we determined the LOOCV accuracy using the top *n* mRNA and the top *m* miRNA, based on Fisher feature selection repeated on each training set cycle for each of the 44 left out subjects. This gave a classification accuracy for each pair, yielding a function of (*n*, *m*), with contours in Figure 2. The best accuracy achieved was 86.36%, with a selection of 2 miRNAs and 8 mRNAs. This should be judged relative to the fact that 1,600 combinations of features are used; however, there are clear trends in the graph in Figure 2, which vary systematically with the variables *n* and *m* in the diagram, so it is unlikely that the strong discriminations are based on low-probability outcomes.

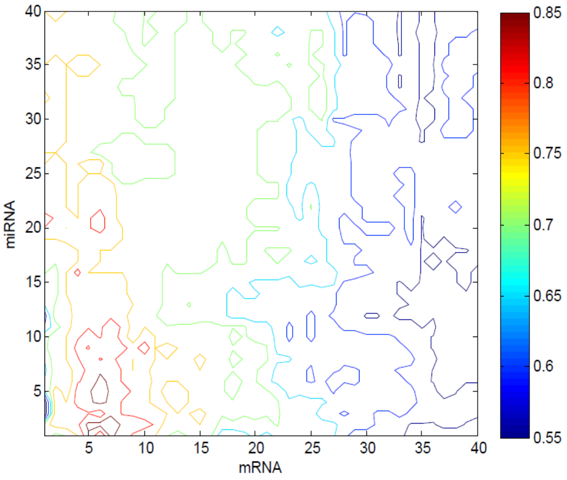

Fig. 2. Contour of accuracies for miRNA and mRNA pair sizes using SVM based on individual feature selection. Y-axis presents the number of miRNAs and x-axis presents the number of mRNAs.

3.2 Integrating intermediate survival time for FSCR

Here we introduce a modified approach for integrating intermediate information. We first created training and test subsets from the 44 samples including the 22 short and 22 long patients, for a leave one out procedure. After leaving out one of the 44 samples, we selected features using Fisher criterion scores based on feature vectors consisting of miRNA and/or mRNA expression profiles. We expanded the training set for computing Cox regression coefficients from the 44 to include all 146 patients (excluding the one left out from the original 44). We calculated the Cox regression-weighted linear combination of the expression level of these genes to generate Cox risk-scores. In test data sets we applied these Cox coefficients to expression levels generating a predictor whose prediction threshold was the median of the training Cox risk scores for classification- this method is denoted as FSCR_REG. On the other hands, we used FSCR_SVM method, which used SVM classifier for a matrix data sets, gene level multiplied by Cox coefficients (FSCR_SVM). We tested the three information types on survival prediction: (1) miRNA only, (2) mRNA only and (3) combined miRNA and mRNA expression profiles, using leave-one-out cross-validation. The highest classification accuracy is 88.64%, obtained by selecting 2 miRNAs and 5 mRNAs using FSCR_SVM (Figure 3), followed by performance of 75% and 70.45% using individual miRNA and mRNA data sets respectively (see Figure 4).

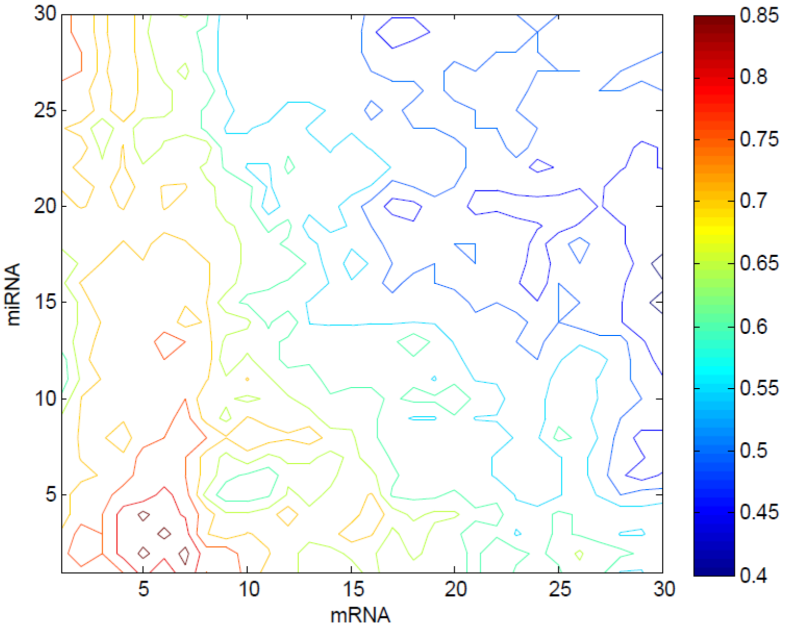

Fig. 3. Contour of accuracies for miRNA and mRNA pair sizes using FSCR based on individual feature selection. Y-axis presents the number of miRNAs and x-axis presents the number of mRNAs.

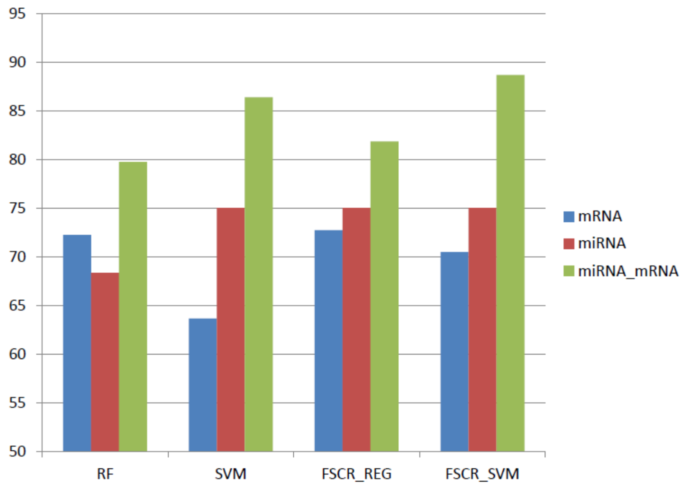

Fig. 4. Comparison of the performance of four different methods with three different type of data sets. RF denotes random forest; SVM, support vector machine; FSCR_REG, regression method using intermediate information; FSCR_SVM, svm classifier using intermediate information.

4. Conclusions

We have studied predictive classifiers for survival phenotypes using the combination of heterogeneous classes of ovarian cancer biomarkers (miRNA and mRNA), especially adopting both discrete (dichotomous) and continuous survival time information. It has been shown that while each biomarker class can individually predict survival times, there is an interactive improvement when the classes are combined in a single machine learning dataset. This combination of different data types in a single classifier is useful and can also be extended to simple algorithms in more general contexts of data integration. More generally in machine learning methods, kernel matrices have the same structure for all data types (if the number of samples is constant), and can be combined by simple kernel addition, thus extending our procedure to one for standardized data integration. The classification methodology is based on feature selection followed by integration of molecular information in miRNA and mRNA for classifying ovarian cancer survival times, with accuracies from the combined information improving that from individual mRNA or miRNA data alone. Among the approaches tested, the combined miRNA and mRNA data information gives better results than individual data set does. The best performance is achieved using the FSCR_SVM method, which uses intermediate information between short and long survival time and integrating two different data sets. The results of different methods and data types are in Figure 4.

Our discussion shows that while many survival-related cancer mechanisms are difficult to identify on an individual (e.g. single gene) level, it in some cases nevertheless possible to predict survival phenotypes from diffuse and noisy data sets such as large scale gene and miRNA expression data. In particular, the strong prediction using miRNAs of survival phenotypes indicates that such diffuse signature mechanisms may be quite pervasive in the progression of ovarian cancer. In contrast to integrating over large numbers of features including miRNA, it is likely also that there are detailed mechanisms to be found that will produce biologically significant information.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2012R1A3A2026438, 2012-0000644).

References

1. Hudson ME, Pozdnyakova I, Haines K, Mor G, Snyder M. Identification of differentially expressed proteins in ovarian cancer using high-density protein microarrays. Proceedings of the National Academy of Sciences of the United States of America (2007) 104,17494-17499.

2. The cancer genome research networks. Integrated genomic analyses of ovarian carcinoma (2011) Nature 474 609-615.

3. Crijns AP, Fehrmann RS, de Jong S, Gerbens F, Meersma GJ, Klip HG, Hollema H, Hofstra RM, te Meerman GJ, de Vries EG, et al. Survival-related profile, pathways, and transcription factors in ovarian cancer. (2009) PLoS medicine 6,e24.

4. Dressman HK, Berchuck A, Chan G, Zhai J, Bild A, Sayer R, Cragun J, Clarke J, Whitaker RS, Li L, et al. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. (2007) Journal of clinical oncology : official journal of the American Society of Clinical Oncology 25,517-25.

5. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, et al. MicroRNA expression profiles classify human cancers. (2005) Nature 435,834-8.

6. Peng S, Zeng X, Li X, Peng X, Chen L. Multi-class cancer classification through gene expression profiles: microRNA versus mRNA. Journal of genetics and genomics = Yi chuan xue bao (2009) 36,409-16.

7. Guo Y, Chen Z, Zhang L, Zhou F, Shi S, Feng X, Li B, Meng X, Ma X, Luo M, et al. Distinctive microRNA profiles relating to patient survival in esophageal squamous cell carcinoma. (2008) Cancer research 68,26-33.

8. Cortes C, Vapnik V. Support-vector networks. (1995) Matchine learning 20, 273-297.

9. Ho TK. A data complexity analysis of comparative advantages of decision forest constructors (2002) Pattern analysis and applications 5, 102-112.

10. Cox D.R. Regression models and life-tables. (1972) JSTOR 34 187-220.