BREAST CANCER SURVIVAL PREDICTION USING MERGED GENE EXPRESSION DATA SETS

Haleh Yasrebi, Philipp Bucher
Bioinformatics group, Swiss Institute for Experimental Cancer Research, Swiss Institute of Bioinformatics
155 Chemin Boveresses, 1066 Epalinges, Switzerland
Haleh.Yasrebi@isb-sib.ch, Philipp.bucher@isb-sib.ch

1 Abstract

The performance accuracy of survival prediction depends on the gene signature and potentially, the sample size of data sets. Overfitting could occur by the small size of samples as the learning ability of machine learning methods depends on the size of training set and validation on the size of testing set. The purpose of this work was to improve survival prediction accuracy by analyzing jointly the breast cancer gene expression data sets from different technologies. To this end, eight public data sets generated from different platforms such as cDNA, Affymetrix and Agilent were merged together with respect to their clinical endpoints. The gene signatures derived from the single and merged data sets were evaluated based on their prediction accuracy and hazard ratio to determine if merging data sets would improve the performance accuracy. Five patients' information and clinical parameters were also analyzed to assess if their prediction accuracy and hazard ratio could be enhanced by increasing the sample size.

2 Methods

Distance Weighted Discrimination (DWD) was applied to adjust the data sets prior to their fusion. Feature selection was done based on Cox p-value ranking and 100 top-ranked genes were selected.

3 Results

The prediction accuracy generated from the merged data sets are similar or inferior (max 6%) to the results obtained from the single data sets. The prediction accuracy is not improved in overall but it shows the reproducibility and consistency across microarray platforms and independent patients' cohorts

A B

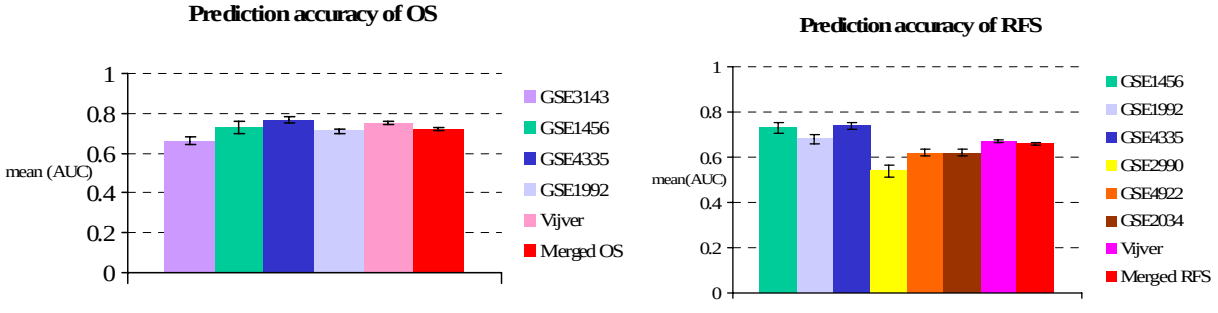


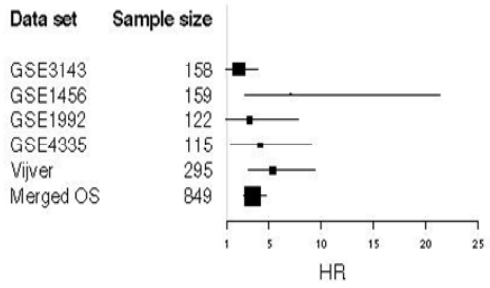
Figure 1. The prediction accuracy of the merged and single data sets with Overall Survival (OS) and Relapse Free

Survival (RFS) endpoints. The results present the mean of the average of the Area Under the Curve (AUC) over 10-fold cross validation. The error bars represent the standard deviation of AUC over 10 iterations.

The Hazard Ratio (HR) of the gene signatures derived from the single and merged data sets are shown in the following figure. Increasing sample size generates a shorter confidence interval, hindering a more robust and confident HR.

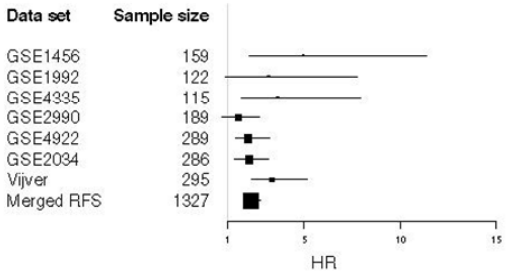
A

Overall survival



B

Relapse free survival



In an assay, GSE1992 was not merged with other data sets in order to be used as validation set. The results of the prediction accuracy of GSE1992 are very close to the results derived from the merged data sets. The HR of the gene signature validated on GSE1992 is higher compared to the HR derived from the merged data sets but they remain in the confidence interval of the HR generated from the merged data sets. These results illustrate a good generalization of prediction and hazard ratio of the survival predictors derived from the merged data sets.

Survival endpoint	Prediction accuracy	HR	CI	p-value
OS	0.76	5.5	1.88-16.1	0.001
RFS	0.64	4.35	1.63-11.6	0.003

4 Conclusion and discussion

Merging microarray gene expression data sets didn't improve the survival prediction but remained reproducible and consistent across independent data sets generated from different microarray platforms and laboratories. Increasing sample size helped to generate the gene signatures with a confident hazard ratio. More data sets should be merged together in order to determine if the survival prediction generated from the merged data sets could outperform the prediction derived from the single data sets.