Hierarchical integration of multi-layered data for classification and biomarker discovery in the presence of sample heterogeneity.

Ellis Patrick[1], Sarah-Jane Schramm[2], John T. Ormerod[1], Graham J. Mann[2], Samuel Müller[1] and Jean Y. H. Yang[1].

[1]School of Mathematics and Statistics, The University of Sydney and [2]Sydney Medical School, The University of Sydney.

Try this for yourself!

This poster was generated in knitr. As a result of this, it is reproducible. To access the code and files needed to generate this poster visit `www.ellispatrick.com/presentations`.

If you like any of the ideas in this poster please cite them. As the corresponding manuscript for this poster is currently in preparation please ask me for further details on how you can do this.

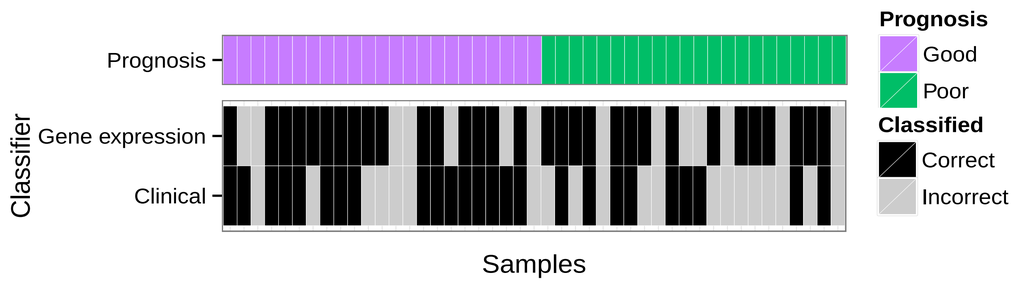
Summary

- Integrating multi-layered high-throughput data for biomarker discovery or classification is challenging from many aspects.
- Modelling the sample heterogeneity observed in the gene expression data with clinical data identifies treatment variables as informative about the predictive value of the gene expression data.
- Our proposed two-stage classification scheme performs well in comparison to single platform and additive approaches of multiple platforms on the two presented datasets. On both datasets it provides

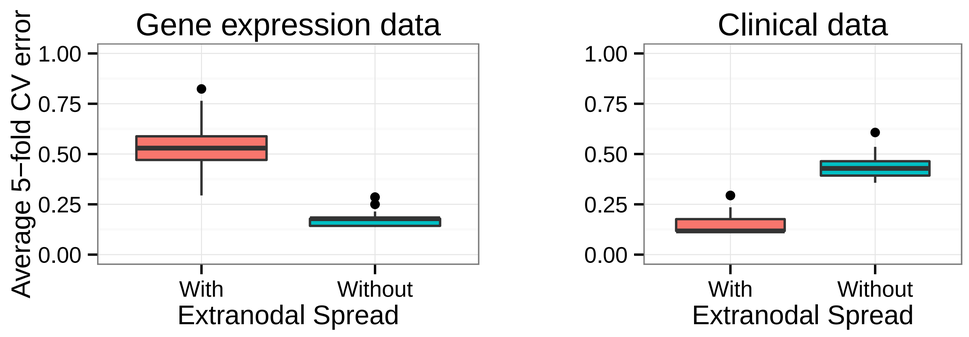
  1) improved predictive performance, and,
  2) models which make biological sense.

Background

The accurate determination of patient prognosis is an important unmet need in the management of many cancers. When analysing a melanoma dataset with multiple high-throughput data sources, it was noted that different data sources appear to correctly classify different patients into their correct prognosis groups [2]. This is illustrated below for some matched gene expression and clinical data.



After further investigation it was noted that a clinical variable (extranodal spread in the tumour) was reasonably adept at explaining which samples the gene expression could and could not classify correctly. We illustrate the significance of this in the boxplots below. The boxplots were created by performing average 5-fold cross-validation within those samples with extranodal spread and within those without it. We observe that there appears to be no genomic signal that differentiates prognosis in the patients with extranodal spread and strong signal in those without it. Furthermore, the reverse can be observed in the clinical data.



These observations motivated the creation of a hierarchical multi-stage classification procedure for using gene expression and clinical data to predict prognosis in cancer patients.

Datasets

We consider two cancer datasets. The datasets were chosen as they contained gene expression data and relatively detailed clinical data.

Melanoma
The melanoma data contains mRNA expression profiles of metastatic melanomas [2] generated on an Illumina HumanWG-6 v3.0 expression beadchip (GEO Accession Number: GSE54467). We restricted the analysis to stage III patients and split the patients into two survival classes: leaving 23 patients surviving more than 4 years after resection of lymph node metastatic disease with no sign of relapse and 22 patients who died of melanoma within 12 months of resection.

Breast cancer
The breast cancer data contains mRNA expression profiles of human breast cancer [1] generated on an Affymetrix HT-U133AA of Av2 GeneChip (ArrayExpress experiment number: E-TABM-158). We restricted the analysis to stage II patients and split the patients into two recurrence classes: consisting of 18 patients with recurrence less than 5 years and 31 with recurrence greater than 5 years.

A multi-stage classifier

Determine which samples the gene expression data can reliably classify.
Perform leave-one-out cross-validation of a classification scheme on the gene expression data. From this prediction errors can be calculated, giving an indication of which samples were able to be correctly classified by the gene expression data.

Determine if there are any clinical variables that explain for which samples the gene expression is informative.
Use logistic regression to identify clinical variables which can predict which samples the gene expression data correctly and incorrectly classified. To avoid over-fitting we perform a variable selection step, selecting the variable with smallest AIC.

Train two classifiers.
Split the data into those that are predicted to be classifiable by the gene expression data and those that are not. For those samples for which the gene expression data is predicted to be informative, train a new classifier using the gene expression data. For those samples for which the gene expression data is not predicted to be informative, train a classifier using the clinical variables.

Use classifiers to predict any new samples.
Given some new samples, predict if the gene expression would be informative for these samples. Given these predictions then classify the samples using the appropriate classifier.

Results

Here we compare the leave-one-out cross-validation balanced error rates of four approaches to classifying the melanoma and breast cancer data. These include performing DLDA on the gene expression data (Gene expression), logistic regression on the clinical data (Clinical variables), performing logistic regression on the CPM variables with a prevalidated decision vector from the gene expression data (Prevalidation) [3] and our proposed multi-stage approach. These are given in the following table

	Melanoma	Breast cancer
Gene expression	0.31	0.61
Clinical variables	0.45	0.50
Prevalidation	0.27	0.58
Multi-stage	0.20	0.24

We next consider performing our multi-stage approach outside of cross-validation. Here it selects extranodal spread and hormonal therapy as the variables that predict the informativeness of the gene expression data in the melanoma and breast cancer data respectively. While there are no treatment variables in the melanoma data, it is likely that patients with extranodal growth would have received additional radiation treatment and so this might be considered as a proxy for differential treatment. For each dataset we created images of the relative gene expression for the top 100 genes with largest fold change between good and poor prognosis below. In both datasets, for those patients that received treatment the gene expression is much less informative than for those that were not treated. This behaviour is most pronounced in the breast cancer data.

References

[1] Chin, K. *et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**(6), 529–541.

[2] Jayawardana, K. *et al.* (2014). Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information. *Int J Cancer*.

[3] Tibshirani, R. J. and Efron, B. (2002). Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol*.

Supported in part by the NHMRC (PG633004, GM and SJS), Cancer Institute New South Wales (10TPG/1/02 GM and SJS) and the ARC (FT0991918, YY, DP130100488, SM and YY, and DP110100061 JO).

Website: `www.ellispatrick.com` Email: `ellis.patrick@sydney.edu.au`