

Use of Microarray Data via Model-Based Classification in the Study and Prediction of Survival from Lung Cancer

Liat Jones

Department of Mathematics
University of Queensland
Brisbane, Q4072, Australia
(617) 33462623

liatj@maths.uq.edu.au

Shu-Kay Ng

Department of Mathematics
University of Queensland
Brisbane, Q4072, Australia
(617) 33656139

skn@maths.uq.edu.au

Cristophe Ambroise

Laboratoire Heudiasyc
UMR CNRS 6599
Compiègne 60200, France

ambroise@utc.fr

Geoff McLachlan

Department of Mathematics
University of Queensland
Brisbane, Q4072, Australia
(617) 33652150

gjm@maths.uq.edu.au

ABSTRACT

We developed a model-based clustering approach to classify tumor tissues on the basis of microarray gene expression. The impact of this classification on cancer biology and clinical outcome was studied. The association between the clusters so formed and patient survival (recurrence) times was examined. The method is illustrated using the Stanford and Ontario lung cancer data. We show that the prognosis clustering is a more powerful predictor of the outcome of disease than current systems based on histopathology criteria and extent of disease at presentation.

Keywords

Mixture models, EMMIX-GENE algorithm, Selection Bias, Microarrays, Survival analysis, Cox proportional hazards, Kaplan-Meier survival curve

1. INTRODUCTION

Lung cancer patients with the same stage of disease can have markedly different treatment responses and clinical outcome. Recent studies have suggested that information from gene expression profiles could be used to develop molecular classifications of cancer [7,10]. Here, we present a novel model-based clustering of microarray data for prediction of clinical outcome from the Stanford [5] and the Ontario [10] Datasets. Both these datasets include tissues of various tumor types, as shown in Table 1. The major differences in samples are that the Stanford Dataset contains relatively more adenocarcinoma samples, and the Ontario Dataset contains only non-small cell carcinomas. In both studies cDNA microarrays were used to obtain gene expression profiles for the samples. However, the tissue clusters as obtained by hierarchical clustering are quite different in the two papers [5,10].

Hierarchical clustering methods lack a statistical model; hence it is difficult to determine what is meant by a ‘good’ clustering algorithm or the ‘right’ number of clusters. We use a mixture model-based approach, which has a sound theoretical framework

for approaching these questions. However, microarray data presents a non-standard problem, in that the number of tissue samples is very small compared to the number of genes. Recently, McLachlan *et al.* [8] developed a mixture-model based approach to allow clustering of tissue samples in the EMMIX-GENE procedure. Here we used the EMMIX-GENE algorithm to carry out unsupervised clustering of the tissue samples based on their gene expression. EMMIX-GENE initially reduces the number of genes to be used in the clustering process by selecting only the most relevant ones. The number of genes can be still much greater than the number of tissues. Such a dimensionality problem is handled with the EMMIX-GENE approach by fitting mixtures of factor analyzers to identify various subgroups among tissues [9]. The impact of this classification of tissues on cancer biology and clinical outcome was investigated. The correlation between the subgroup label phenotype and patient’s survival (recurrence) times as given in terms of the clinical data was examined. Based on the survival analysis, we show that the classification using microarrays can be used to predict clinical outcome for cancer prognosis. To identify reliably good and poor-prognosis subgroups, we also developed a supervised classification method to construct a classifier that has a small error rate after correction for the selection bias [1]. Important genes that play a key role in linking gene expression with clinical outcome can be identified.

Table 1. Comparison of Tumor Types for Stanford and Ontario Datasets

Tumor Type	Number of Samples	
	Stanford	Ontario
Adenocarcinoma	41	19
Squamous cell	16	14
Large cell	5	4
Adenosquamous	0	1
Carcinoid	0	1
Small Cell	5	0
TOTAL	67	39

2. MATERIALS AND METHODS

2.1 Data Selection

We start with the reduced datasets; the Stanford subset of 918 genes (with most similar expression within tumor pairs, but which differed among the other tumor samples) and the Ontario subset of 2880 genes (which contained data points in at least 80 percent of the samples and the transcripts had at least two samples with an absolute value of two in log₂ space). The Stanford Dataset had a total of 73 samples (including matched samples) and the Ontario Dataset had a total of 39 samples (see Table 1 for the tumor types).

In our pre-processing steps, we imputed missing values using the method of Dudoit *et al.* [4]. The datasets were then column normalized, followed by row normalization. The datasets were then input into the EMMIX-GENE algorithm.

2.2 Model-based Clustering Approach

The EMMIX-GENE algorithm has three main steps. In the first select-genes step, we consider each gene separately, and delete it if it is not useful in revealing group structure in the tissues. In the second cluster-genes step, we cluster the retained genes on the basis of Euclidean distance (essentially k-means) so that highly correlated genes are put in the same cluster. The user specifies the number of gene clusters; here we use 20 gene clusters. In the final cluster-tissues step, we cluster the tissues on the basis of the means of these clusters of genes. We call these means of the gene-clusters metagenes. We performed model-based clustering of the tissues by fitting mixtures of factor analyzers, q , to these metagenes. The number of tissue clusters was varied by using a different number of components, g . For the Stanford Dataset we used an increasing number of components from $g = 3$ to $g = 7$, and a mixture of factor analyzers with $q = 6$ factors. We also ran the dataset using just the 43 tissues classified in their study into one of the adenocarcinoma subgroups, with $g = 2, 3$, and 4 components. For the Ontario Dataset, we ran cluster-tissues with $g = 2, 3$ and 4 components, also with $q = 6$ factors.

2.3 Survival Analysis

In the analysis of the probability of overall survival with the Stanford Dataset, only tissues with known clinical characteristics and survival times are included. There were 28 tissues with 4 tumor pairs derived from the same patients. The Kaplan-Meier method was adopted to estimate the overall survival of patients

whose tumors were classified as good-prognosis and poor-prognosis subgroups. The Kaplan-Meier survival curves were compared with the use of the log-rank test. We determined the difference between the relative hazard ratios with respect to the prognosis subgroups using the Cox proportional hazards model [2]. Multivariate Cox proportional hazards regression were performed to adjust for other clinical characteristics. The tumor grade (3 vs. 1 or 2), tumor size, the number of tumors in lymph nodes, and the presence of metastases were used as variables. The significance of estimated hazard ratios were tested using the Wald test.

With the Ontario Dataset, we defined the outcome as the time between surgery and the recurrence. This is equivalent to the definition in [10] because patients free from recurrence are all still alive at the end of follow-up period. The exact recurrence and survival times of 2 patients were unknown and they were excluded from the survival analysis. There were 37 patients with 15 censored (recurrence-free and still alive at the end of the follow-up). Kaplan-Meier estimates of probabilities of recurrence-free for each prognosis subgroup were compared. The relative hazard ratio with respect to the two subgroups was determined based on the Cox proportional hazards model. The tumor stage was included in the multivariate proportional hazards regression analysis. All calculations in the survival analysis were performed with the S Plus statistical package.

2.4 Supervised Classification Method

Based on a supervised clustering approach, a prognosis classifier was developed to predict the class of origin of a tumor tissue with a small error rate after correction for the selection bias [1]. With the Stanford Dataset, we considered the prognostic category as metastasis vs. metastasis-free. With the Ontario Dataset, we considered recurrence vs. recurrence-free. A support vector machine (SVM) [3] was adopted to identify important genes that play a key role on predicting the clinical outcome, using (a) all the genes, and (b) the metagenes. A cross-validation (CV) procedure was then performed to calculate the test error, after corrected for the selection bias.

To illustrate how the performance of the SVM can be improved by eliminating genes that do appear to have little discriminatory power, we applied the method of Guyon *et al.* [6]. The external cross-validation procedure as advocated by Ambroise and McLachlan [1] was used to estimate the prediction error rate of the SVM based on the selected genes.

3. RESULTS

3.1 Clustering of Tumor Tissues

For the Stanford Dataset, we retained 451 genes out of the 918 genes after our select-genes step. The top five genes (as indicated by the highest log likelihood value) were aldo-keto reductase

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference '00, Month 1-2, 2000, City, State.

family 1, achaete-scute complex (*Drosophila*) homolog-like 1, fibrinogen, A alpha polypeptide, tumor protein p63 and XAGE-1 protein. We clustered the retained genes into 20 groups. The heat maps for these clusters are shown in Figure 1, where the tissues are arranged by their histological classification. The majority of these gene clusters clearly separate tissues into one or more histological categories. Garber *et al.* [5] could obtain the histological classification using hierarchical clustering, and identified genes differentially expressed in the SCC (squamous cell), LCLC (large cell) and SCLC (small cell) groups. Our 20 gene clusters represent all of their gene sets. Additionally, we find some interesting new gene groups; one up-regulated in LCLC and a group up-regulated in the normal tissues. Within the adenocarcinoma samples (plus two non-adenocarcinoma samples), the Stanford study identified three AC subtypes, and also genes that separate between them. Again all of their gene groups are found in our clusters. We further ran EMMIX-GENE on 43 AC classified tissues only, in order to find genes that could best differentiate AC subtypes. Using this reduced set of tissues, we could identify gene clusters corresponding to two of the Stanford marker genes groups [5, Figure 5]. These were the genes with high expression in AC group 3 but low in AC group 1 and AC group 2, and also genes with high expression in AC group 2 but low in AC group 3. Additionally, using the reduced set of tissues yields many new genes of interest.

On clustering the 73 tissues using three components, we reproduced exactly the Stanford groupings as follows, one cluster comprises their AC group 1, AC group 2 and Normal groups, the second comprises AC group 3, LCLC and SCLC groups and finally the third their SCC group. We do find one important difference in our classification. The primary tumor (PT) and the intrapulmonary metastasis (MT) samples from patient 313 are separated into different groups. This is discussed further in relation to the survival analysis.

For the Ontario Dataset, we retained 575 genes in the select-genes step, of the 2880. Here the top five genes included several genes that may be of interest in cancer: hypothetical protein (inhibitor of cancer growth), zinc finger protein 415, RPL13 ribosomal protein (decreased expression in cancerous cells compared to benign lesions in breast carcinoma) and CGI-96 hypothetical protein (highly conserved). None of our top five genes are referred to in Wigle *et al.* [10]. Of their mentioned genes of interest, we retain 7 of the 16 in the select-genes filtering step.



Figure 1. Heat Maps for the 20 Stanford Gene Clusters. Tissues are represented by columns and are ordered by their histological classification; Adenocarcinoma (1-41), Fetal Lung (42), Large cell (43-47), Normal (48-52), Squamous cell (53-68), Small cell (69-73). Genes are represented by rows.

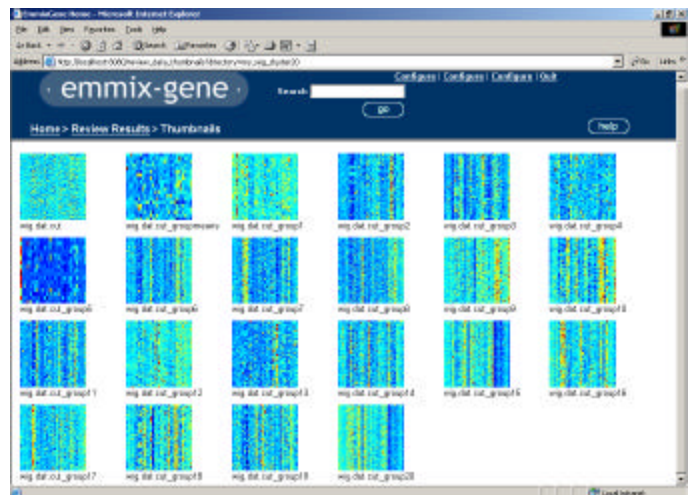


Figure 2. Heat Maps for the 20 Ontario Gene Clusters. Tissues are represented by columns and are ordered as Recurrence (1-24) and Non-Recurrence (25-39). Genes are represented by rows.

We next clustered the retained genes into 20 groups. The heat maps for these clusters are shown in Figure 2, where the tissues are arranged in order of recurrence against non-recurrence. From inspection, genes in groups 15 and 20 show decreased expression for the non-recurrence tissues, whereas those in groups 6 and 8 have higher expression in the non-recurrence tissues. Group 20 and group 15 include down-regulated genes found in the Ontario study: FLT1, PIK3R2 and ZNF136, but also FUS which, conversely, was found to be up-regulated in the non-recurrence group in the Ontario study. Group 8 contains 3 of the top ten genes: zinc-finger protein 415, CGI-96 and RPL13, and these are all down-regulated in the non-recurrence group. This suggests a role for RPL13 in distinguishing more aggressive lung tumors, analogous to its behaviour in breast tumors. Interestingly,

inspection of the heat maps when the tissues are ordered by histological classification (not shown here) does not indicate that our gene groups separate the tissues into clusters by histology. This fits in with the conclusion in the Ontario study, whereby they note "we did not observe a primary correlation of expression profiles with tumor histological type..."

3.2 Impact of Classification on Outcomes

With the Stanford Dataset, among the 28 tissues, 10 were classified as poor-prognosis and 18 were classified as good-prognosis. This classification matches the subdivision obtained by Garber *et al.* [5]. The poor-prognosis subgroup corresponds to the AC group 3 (worst clinical outcome) of [5], while the good-prognosis subgroup corresponds to AC groups 1 and 2. There is only one exception; in the Stanford study the PT and MT from patient 313 were both clustered into the AC group 1. With our clustering method, the PT was classified into the good-prognosis subgroup, while the MT was classified into the poor-prognosis subgroup. Our results also indicate that there is not significant difference between the genes expression profiles of AC groups 1 and 2.

The Kaplan-Meier survival curves (Figure 3) showed a significant difference in the probability of overall survival between the poor-prognosis and good-prognosis subgroups ($P < 0.001$). The mean (\pm SE) survival times were 7.4 ± 1.9 and 28.1 ± 3.1 months, respectively. The estimated hazard ratio for overall survival in the poor-prognosis subgroup as compared with the good-prognosis subgroup was 10.2 (95% confidence interval (CI): 3.1 to 33.1; $P < 0.001$). The prognosis clustering was significantly associated with the clinical outcome on the basis of survival. Table 2 shows the results of the multivariate Cox regression analysis. The prognosis clustering was the only significant factor correlating with overall survival ($P = 0.001$). The relative hazard ratio for overall survival was 11.2 (95% CI: 2.6 to 49.1).

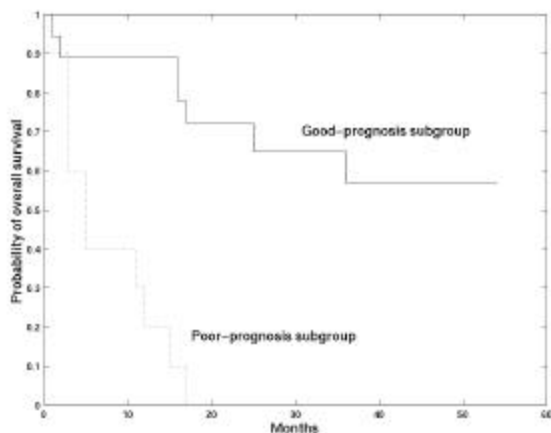


Figure 3. Kaplan-Meier survival curves for the two prognosis subgroups (Stanford Dataset)

Table 2. Multivariate Cox proportional hazards analysis of the risk of death (Stanford Dataset)

Variable	Hazard ratio (95% CI)	P-value
Poor-prognosis subgroup (vs. good-prognosis)	11.2 (2.6-49.1)	0.001
Tumor grade 3 (vs. grades 1 or 2)	0.54 (0.1-2.0)	0.35
Tumor size	1.31 (0.5-3.4)	0.58
No. of tumors in lymph nodes	1.96 (1.0-3.9)	0.051
Presence of metastases	1.71 (0.5-5.7)	0.39

With the Ontario Dataset, among the 37 patients, 19 were classified as poor-prognosis and 18 were classified as good-prognosis. In [10], they were clustered into two subgroups of high recurrence rate ($n=26$ patients) and of low recurrence rate ($n=11$), respectively. With our prognosis clustering, the good-prognosis subgroup involved all 15 patients who are recurrence-free plus 3 patients who had experienced relapse of their tumor. These 3 patients, however, were still alive at the end of the follow-up. Within the poor-prognosis, 68% of patients were either died or having smaller censoring time compared to these 3 patients. The Kaplan-Meier curves (Figure 4) showed a significant difference in the probability of recurrence-free between the poor-prognosis and good-prognosis subgroups ($P < 0.001$). The mean (\pm SE) times between surgery and recurrence were 491 ± 67 and 617 ± 21 days, respectively. The estimated hazard ratio for recurrence in the poor-prognosis subgroup as compared with the good-prognosis subgroup was 8.6 (95% CI: 2.6 to 29.3; $P < 0.001$). The prognosis clustering was significantly associated with the clinical outcome on the basis of recurrence-free. Table 3 shows the results of the multivariate Cox regression analysis. It is evidence that the two prognosis clusters were clinically different after the adjustment for the tumor stage ($P < 0.001$). The relative hazard ratio for recurrence was 8.9 (95% CI: 2.6 to 30.7) between the poor-prognosis subgroup against the good-prognosis subgroup.

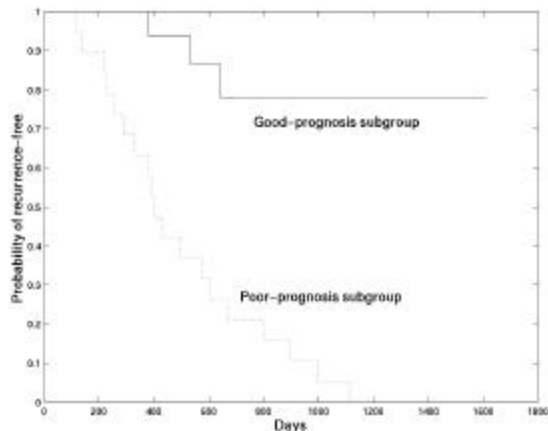


Figure 4. Kaplan-Meier curves of the probability of recurrence-free for the two subgroups (Ontario Dataset)

Table 3. Multivariate Cox proportional hazards analysis of the risk of recurrence (Ontario Dataset)

Variable	Hazard ratio (95% CI)	P-value
Poor-prognosis subgroup (vs. good-prognosis)	8.89 (2.6-30.7)	<0.001
Tumor stage	0.92 (0.5-1.7)	0.80

3.3 Cross-validation

The overall error rates are listed in Table 4, where the apparent error rate was calculated by applying the SVM to the training data and CV10E denotes the cross-validated (ten fold) error rate. It can be seen for these two data sets that there is not much difference between the cross-validated error rates for the SVM based on (a) all genes and (b) 20 metagenes. Concerning the size of the cross-validated error rate, it can be seen that the error in predicting the clinical outcome of a new lung cancer tumor is approximately 15% for the Stanford Dataset as compared to around 28% for the Ontario Dataset. The ten-fold cross-validated rates are reported in Table 5, where it can be seen that the prediction error is approximately 23% for 13 genes.

Table 4. Error rates (in percentage) with linear SVM

Error rate	Stanford Dataset	Ontario Dataset
(a) all genes		
Apparent error	0.00	0.00
CV10 error	15.07	28.21
(b) metagenes		
Apparent error	1.37	5.13
CV10 error	10.96	33.33

Table 5. Ten-fold CV prediction error rate

No. of genes	No. of variables	Estimated error
1	1	0.38
2	2	0.41
3	4	0.26
4	8	0.23
6	32	0.26
8	128	0.28
10	512	0.26
11	1024	0.28
12	2048	0.31
13	2880	0.23

4. CONCLUSIONS

We developed a model-based clustering approach to classify tumor tissues using microarrays genes expression profile. The clustering performed best as a predictor of the clinical outcome based on the overall survival or recurrence times. The results obtained from the analysis of both Stanford and Ontario Datasets indicate that classification of patients into good-prognosis and poor-prognosis subgroups on the basis of microarrays could be a useful tool for linking the impact on lung cancer biology and guiding treatment therapy and patient care to lung cancer patients.

Of particular biological interest, the Stanford analysis which had matched tissue samples, showed that gene expression could differentiate between the primary and metastatic tumors for a particular patient. The two metastases from patient 319 were clustered in AC group 3 (poor prognosis), while the primary tumor was in AC group 1 (good prognosis). Our cluster analysis reveals a similar result for patient 313. This further supports the hypothesis that all AC tumors derive from a common precursor, and become invasive as their gene expression switches to that of the AC group 3 type.

There are some obstacles to integrate information from the Stanford and Ontario Datasets. In the Stanford Dataset, we have clinical data only for the adenocarcinoma (plus two other) samples that are classified into their AC groupings. On the other hand, in the Ontario Dataset we have clinical data for various tumor types. The Ontario study attempts to relate non-adenocarcinoma samples, as well as adenocarcinoma samples, to clinical outcome. In order to do so, they simply cluster the tumors into two groups; recurrence vs non-recurrence, with no evidence for adenocarcinoma subclasses as found in the Stanford study. Additionally, within

our gene clusters, we could not find genes common to both datasets.

5. ACKNOWLEDGMENTS

We thank Nazim Khan for data pre-processing, Abdollah Khodkar for his cluster-tissues code and for running bootstrap analyses, and Katrina Monico for running some initial survival analyses.

6. REFERENCES

- [1] Ambrose, C. and McLachlan, G.J. Selection bias in gene extraction on basis of microarray gene expression data. *PNAS*, 99 (2002), 6562-6566.
- [2] Cox, D.R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B*, 34 (1972), 187-220.
- [3] Davison T.S., *et al.* Support Vector Machine classification of data quality in microarray experiments. In *Proceedings of CAMDA'01 (2001)*. *Methods of Microarray Data Analysis*, vol. 2. Kluwer, Boston, M.A.
- [4] Dudoit, S., Fridlyand, J., and Speed, T.P. Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. *Technical Report #576* (June 2002).
<http://www.stat.berkeley.edu/~sandrine/tecrep/576.pdf>.
- [5] Garber, M.E., *et al.* Diversity of gene expression in adenocarcinoma of the lung. *PNAS*, 98 (Nov 2001), 13784-13789.
- [6] Guyon, I., *et al.* Genes selection for cancer classification using Support Vector Machines. *Machine Learning*, 46 (2002), 389-422.
- [7] Mateos, A., *et al.* Supervised and hierarchical unsupervised neural networks for clustering both gene expression profiles and samples. In *Proceedings of CAMDA'01 (2001)*. *Methods of Microarray Data Analysis*, vol. 2. Kluwer, Boston, M.A.
- [8] McLachlan, G.J., Bean, R.W., and Peel, D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18 (2002), 413-422.
- [9] McLachlan, G.J. and Peel, D. *Finite Mixture Models*. Wiley, New York, 2000, Chapter 8.
- [10] Wigle, D.A., *et al.* Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research*, 62 (Jun 2002), 3005-3008.