AIM: To link gene expression data with survival for lung cancer.

1. General Strategy

CLUSTER ANALYSIS
We apply a model-based clustering approach to classify tumor tissues on the basis of microarray gene expression.

SURVIVAL ANALYSIS
The association between the clusters formed and patient survival (recurrence) times is examined.

DISCRIMINANT ANALYSIS
We attempt to 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.

2. Datasets and Pre-processing

Ontario [1] We start with the reduced set of 2880 genes, and all 39 tumor samples. Stanford [1] We start with the reduced set of 918 genes. Initial cluster analysis of the 73 tissue samples showed that 6 of the 41 adenocarcinoma tumors cluster into other histological groups (large cell or squamous cell). These were excluded from the analysis, leaving just 35 adenocarcinoma tissues for the survival analysis.

Michigan [1] We start with the reduced dataset of 4965 genes and use only the 86 adenocarcinoma tumor samples. We imposed a floor of –100 and a ceiling of 26,000.

Harvard [1] We start with the reduced dataset of 3312 genes and use only the 127 adenocarcinoma tumors for the survival analysis. We further reduce the geneset to 3190, by imposing a floor of -1 and a ceiling of 3,000.

Pre-processing steps: The data were logged except in the Michigan set, where the generalised log transformation was used. We column normalised (except in the Michigan and Harvard datasets) and then row normalised the data. The datasets were then input into the EMMIX-GENE algorithm for cluster analysis.

3. The EMMIX-GENE algorithm

Many current methods for cluster analysis of microarray data lack a statistical model. Hence it is difficult to define what is meant by a ‘good’ clustering algorithm or the ‘right’ number of clusters. Mixture models provide a sound theoretical framework for approaching these questions. However, microarrays present a non-standard problem, in that the number of tissue samples is very small compared to the number of genes. The EMMIX-GENE algorithm overcomes this by fitting mixtures of factor analyzers to cluster the tissues [2].

EMMIX-GENE has three main steps:

select-genes Filter genes to select for most informative genes.
cluster-genes Group the retained genes into a user-specified number of clusters.
cluster-tissues Cluster the tissues on the basis of the group means of the gene clusters (METAGENES), into a user-specified number of clusters.

EMMIX-GENE is available at: http://www.maths.uq.edu.au/~gjm/EMMIX-GENE

4. CLUSTER ANALYSIS

The results of the cluster analysis are presented in the figures on the right, showing the heat maps for the gene clusters based on tissues, and the tissue clusters are represented on the survival curves.

5. SURVIVAL and DISCRIMINANT ANALYSES

The relevance of the clusters obtained to survival, is demonstrated by the Kaplan-Meier estimates of the survival functions for time to recurrence in the Ontario data and time to death in the other three datasets. The significant differences between these curves was obtained using the log-rank and Wilcoxon test statistics. Also significant differences were obtained when allowances were made for the difference in clinical covariates (stage, lymph node involvement), using Cox’s proportional hazards model and fully parametric mixture survival models. The potential for expression data for prognosis was illustrated by the use of support vector machines [3].

6. Conclusions and Discussion

We show that by reducing the number of genes, then establishing clusters in the tissues, we obtain significant correlation with patient survival. We can do this for data obtained from different platforms. For one dataset (Ontario) we relate different histological types of tumor (non-small cell carcinomas) to survival, while for the other three datasets we relate adenocarcinoma tumors to survival. A major limitation to the study is in the number and histological subtypes of tumors in each dataset. In turn, clinical data are often available for a small subset of these tumors, further limiting the survival analysis.

Interestingly, our cluster analysis of the full Stanford and Harvard datasets with all tissues, including nonadenocarcinomas, separates tissues according to histological subtypes. This was not the case for the Ontario dataset, as was also found in their study.

Further detailed analysis of the genes making up the top metagenes is of biological importance. While we find many of the genes mentioned in the papers, several interesting new genes are found to be important in distinguishing tumor classes and thus for patient prognosis.

References