Binary tree-structured vector quantization approach to clustering and visualizing microarray data

M. Sultan¹, D. A. Wigle², C. A. Cumba ¹, M. Maziarz³, J. Glasgow⁴, M.S. Tsao⁵ and I. Jurisica¹, ³

¹Division of Cancer Informatics, Ontario Cancer Institute, 610 University Avenue, Toronto, Ontario, M5G 2M9, Canada, ²Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario, M5G 1X5, Canada, ³Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario, M5S 3H8, Canada, ⁴Department of Computing and Information Science, Queen's University, Kingston, Ontario, K7L 3N6, Canada and ⁵Division of Cellular and Molecular Biology, Ontario Cancer Institute, 610 University Avenue, Toronto, Ontario, M5G 2M9, Canada

Received on January 24, 2002; revised and accepted on April 1, 2002

ABSTRACT

Motivation: With the increasing number of gene expression databases, the need for more powerful analysis and visualization tools is growing. Many techniques have successfully been applied to unravel latent similarities among genes and/or experiments. Most of the current systems for microarray data analysis use statistical methods, hierarchical clustering, self-organizing maps, support vector machines, or k-means clustering to organize genes or experiments into ‘meaningful’ groups. Without prior explicit bias almost all of these clustering methods applied to gene expression data not only produce different results, but may also produce clusters with little or no biological relevance. Of these methods, agglomerative hierarchical clustering has been the most widely applied, although many limitations have been identified.

Results: Starting with a systematic comparison of the underlying theories behind clustering approaches, we have devised a technique that combines tree-structured vector quantization and partitive k-means clustering (BTSVQ). This hybrid technique has revealed clinically relevant clusters in three large publicly available data sets. In contrast to existing systems, our approach is less sensitive to data preprocessing and data normalization. In addition, the clustering results produced by the technique have strong similarities to those of self-organizing maps (SOMs). We discuss the advantages and the mathematical reasoning behind our approach.

Availability: The BTSVQ system is implemented in Matlab R12 using the SOM toolbox for the visualization and preprocessing of the data (http://www.cis.hut.fi/projects/somtoolbox/). BTSVQ is available for non-commercial use (http://www.uhnres.utoronto.ca/ta3/BTSVQ).

Contact: ij@uhnres.utoronto.ca

Keywords: microarray data clustering and visulization; self-organizing maps, partitive k-means clustering; lung cancer.

INTRODUCTION

Advances in microarray technology have made it possible to simultaneously monitor the expression of thousands of genes in genomes (Schena, 1999). The challenge is to effectively analyse and interpret this large volume of information. Diverse statistical and machine-learning techniques have been applied to unravel latent similarities among genes and experiments. The common goal is to find groups of genes that are coordinately or differentially expressed, and to find groups of samples that are biologically relevant. Several clustering approaches have been proposed and implemented in this domain (Eisen et al., 1998; Lukashin and Fuchs, 2001; Oja et al., 2002). Despite the general success of these approaches, some deficiencies remain.

Without an explicit bias, most of the current clustering methods applied to gene expression data often produce conflicting results and may produce clusters with little or no biological relevance. It is becoming increasingly clear that none of the approaches alone are sufficient and that the application of various techniques will allow different aspects of the data to be explored. Therefore, without a complete understanding of the underlying biology of gene regulation, there might never be a single technique that will allow us to find all the relationships in the data (Quackenbush, 2001). A suitable technique
would be one that produces biologically relevant clusters. In addition, if different clustering algorithms produce partially overlapping groups of genes or samples, one has increased confidence in the results.

This paper compares similarities and differences among diverse clustering approaches. We subsequently propose a novel technique that suits the characteristics of microarray gene expression data.

**GENE EXPRESSION DATA CHARACTERISTICS**

Gene expression data distributions can best be described by multivariate Gaussians (Hunter et al., 2001). Exploration of gene expression data sets is always problematic due to its inherent dispersion and missing values (Quackenbush, 2001). In order to apply a cluster analysis technique, it is important to explore the characteristics of gene expression data. The selection of a clustering technique should be based on how the technique relates to the characteristics of the data, and whether this technique is able to find any biologically significant clusters (Quackenbush, 2001).

In experiments involving a test versus reference design, representing differentially expressed genes with a log ratio provides an equal spread between up- and down-regulated genes, with a mean of zero for equally expressed genes. Although there are clear advantages to using log ratios to represent the normalized data, this introduces the problem of loss of information about the absolute gene expression level in the sample. This is a serious statistical shortcoming, since equal weights could be given to ratios that are based on widely different absolute levels of expression. The background intensity poses another set of issues to be addressed. First, there is no single established method for deciding what is the most appropriate area around the spot for measuring background. In addition, subtracting locally measured background from the spot intensity value can result in a background. In addition, subtracting locally measured background for measuring background.

Preprocessing of the original data using an appropriate normalization technique can minimize various types of errors. Normalization should account for the dye bias, both in labelling and detection efficiencies, and for the non-linear relationship between dye intensity and expression level (Quackenbush, 2001). It should also handle negative intensity values, which may result from background subtraction, as well as variance within and among experiments. In the worst case, normalization may introduce more error into the data than it removes (as is often the case with background subtraction), thus further decreasing the quality and reliability of the results.

There are many additional issues related to microarray data generation, with numerous potential sources of error from the biological and technical side. In addition, the concentration of cDNA used in hybridization is not linearly related to the signal intensity (Quackenbush, 2001), which complicates any experimental conclusions. However, these issues are beyond the scope of this paper.

**CLUSTERING METHODS**

Clustering can be defined as the process of partitioning a data set into subsets $D_i, i = 1, \ldots, N$, called clusters, such that some distance measure is minimized within clusters and maximized between clusters (Kohonen, 2001). One can differentiate between ‘crisp’ clustering approaches, where individual clusters are non-overlapping, and ‘fuzzy’ clustering approaches, where clusters may overlap. Distance measures determine the distance between items in a multidimensional space by considering dimensions as components of the vector representing each item.

In general, the cluster definition problem is NP-complete; thus, no efficient and optimal algorithm exists to solve this problem (Mangiameli et al., 1996). A significant reduction in the computational load is achieved if the subsets $D_i$ are partitioned in several sequential steps and the distance measure is applied to these subsets. This divides clustering methods into two major categories, simple clustering (without hierarchies) and hierarchical clustering. Hierarchical clustering is able to find generic relationships that exist between the resulting clusters (Kohonen, 2001). Figure 1 presents examples of approaches to clustering and distance/similarity measures.

**Hierarchical clustering: agglomerative and partitive**

Depending on how the comparisons are made, hierarchical clustering is divided into splitting methods, also known as partitive or divisive methods, and merging methods, also known as agglomerative methods. Splitting methods start with the whole data set as a single cluster that is partitioned into disjoint subsets $D_1$ and $D_2$, where the inter-cluster distance is maximized. The subsets $D_1$ and $D_2$ are further subdivided into $D_{11}$ and $D_{12}$ and $D_{21}$ and $D_{22}$, and so on. This iterative process of sample subdividing generates a binary tree. In contrast, the merging methods take individual data points as a singleton cluster $D_j^{(0)}$, and at every successive step $t$, the $k$ most similar subsets $D_{j1}^{(t-1)}, \ldots, D_{jk}^{(t-1)}$ are merged into a larger subset $D_j^{(t)}$. In general, $k = 2$ is adopted to build cluster hierarchies.

**Nature of distance measures**

Almost all of the distance measures assess similarity between the components of a vector, which in general can be formalized as $d_{ij} = \alpha \sum_{l=1}^{n} f(i_l, j_l)$, where $\alpha$ is a coefficient depending on vectors $i$ and $j$, and varying across similarity measures. The function $f$ may represent
the sum, difference, probability, or some other function applied to its arguments.

The selection of a distance measure plays an important role in partitioning the data set. The distance measures are divided into metric and semi-metric measures. A semi-metric measure is one where the distance between two vectors satisfies the following three rules: (i) The distance between two vectors is positive ($d_{ij} \geq 0$). (ii) The distance is symmetric ($d_{ij} = d_{ji}$). (iii) A vector is zero distance from itself, ($d_{ii} = 0$). A measure is metric if the distance $d_{ij}$ between two vectors $i$ and $j$ in a data set obeys the triangle rule in addition to above three rules, i.e. for any vectors $i$, $j$ and $k$, $d_{ik} \leq d_{ij} + d_{jk}$.

Most of the agglomerative clustering algorithms are based on the nearest-neighbour approach and utilize metric or semi-metric distance measures. Metric distance measures are sensitive to noise and outliers. Even the addition of a single data point to a cluster can radically change the resulting distances (Bezdek and Paul, 1998). Probabilistic distance measures are also affected by noise. However, the most common divisive clustering approaches use mean or median values to partition the data set; these minimize the effect of noise and outliers.

### CLUSTERING MICROARRAY DATA

Current state-of-the-art analysis of microarray data involves applying statistical methods, hierarchical clustering, self-organizing maps, support vector machines, or $k$-means clustering to organize genes into meaningful groups. Of these methods, hierarchical clustering has been the most widely applied, and has demonstrated considerable power in delineating clinically relevant patterns within databases of gene expression profiles from patient samples (Alizadeh et al., 2000; Garber et al., 2001; Sorlie et al., 2001; Takahashi et al., 2001). Other approaches, such as Bayesian clustering, require the availability of prior distributions on the data, while $k$-means clustering requires one to specify $k$.

The hierarchical clustering technique most widely used by the microarray community is an agglomerative (bottom-up) approach. It begins with individual data points being singleton clusters, and then successively merges them to generate a tree structure called a dendrogram. The dendrogram does not provide a unique clustering; rather, partitioning can be achieved by cutting the dendrogram at a certain level. Usually the dendrogram is cut at some level where there is a large distance between two merged clusters, but this does not guarantee that within-cluster distance is minimized. Each cluster may consist of several other very distant clusters. To obtain optimal clustering the dendrogram must be cut at several places. Thus, hierarchical clustering loses some of the information present in global patterns by concentrating on local patterns first, so it may not fully reflect the multiple ways in which expression patterns are similar.

In contrast, partitive clustering divides the data set into a predefined number of compact clusters by minimizing some criterion or error function. At each level of the partition the intra-cluster distance is minimized while inter-cluster distance is maximized. Two different ways to interpret the same agglomerative clustering result using a dendrogram are shown in Figure 2.

Tibshirani et al. (1999) have explored diverse clustering methods, and illustrated how they can be used to arrange both genes and a set of DNA microarray experiments. They found that both Two-way tree-Structured Vector Quantization (TSVQ) (Gersho and Gray, 1992) and hierarchical clustering successfully organized the genes and experiments, and produced some visible structure.

---

#### Table: Distance Measures

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$</td>
</tr>
<tr>
<td>Manhattan</td>
<td>$\sum_{i=1}^{n}</td>
</tr>
<tr>
<td>Minkowski</td>
<td>$\sqrt{\sum_{i=1}^{n}</td>
</tr>
</tbody>
</table>

- **Pearson’s correlation**: $\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\text{std}(x)\text{std}(y)}$
- **Cosine correlation**: $\frac{\sum_{i=1}^{n} x_i y_i}{P(\mathbb{D})}$
- **Probabilistic measures**: e.g. $P(i,j)$

---

#### Fig. 1. Different clustering algorithms and distance/similarity measures.
VISUALIZING GENE EXPRESSION PROFILES

The challenges of microarray data visualization limit effective data exploration by domain experts. Several techniques have been used to visualize this highly multi-dimensional data. The self-organizing map (SOM) algorithm (Kohonen, 2001) is an efficient tool for the visualization of multidimensional data (Vesanto, 1999). SOMs have previously been shown to be effective for the exploratory analysis of gene expression data (Tamayo et al., 1999). It has also been shown that SOMs can create clusters similar to ones provided by hierarchical clustering (Oja et al., 2002). In addition, SOMs provide an intuitive visualization of similarity relationships and cluster structures by placing similar clusters in the same area of the two-dimensional map.

SOMs are neural network algorithms widely used in data analysis and vector quantization. The algorithm is similar to k-means clustering, with the additional constraint that cluster centres are restricted to lie in a two-dimensional manifold. SOMs have two main characteristics: (1) quantization of a high-dimensional space, which is similar to other vector quantization techniques such as LBG (Gersho and Gray, 1992) and k-means, and (2) a topological property, which enables the ordering of centroids.

Component planes of SOMs are the planes of Voronoi Tessellations (Vesanto, 1999), each representing a sample in a microarray experiment (i.e. gene expression profile of a patient). Figure 3 presents quantized gene expression visualizations by component planes of a SOM. A hexagonal map unit in a component plane represents a gene selected by the SOM algorithm, which is a centre of a cluster in the data. The colour of the map unit represents the expression value of the gene associated with that unit. Thus, similar gene expression profiles correspond to similar colour patterns of the component plane.

Fig. 2. Partitive cluster tree and dendrogram. (a) Binary tree generated by partitive clustering, representing the structure of five arbitrary data points. (b) and (c) Two different representations of the same agglomerative clustering result using dendrograms.

Fig. 3. SOM component plane visualizations. Component planes are the planes of Voronoi Tessellations representing gene expression profile of a patient. The task is to find groups of patients with similar component planes (patients with similar gene expression profiles). These can then be located manually. (a) A component plane of SOM. Hexagonal units of a component plane, listed with labels of genes selected by SOM algorithm. The colour of a map unit represents the expression value of the genes associated with that unit. (b) Component planes of a group of early-recurring and non-recurring lung cancer patients. (c) Similar component planes are rearranged manually. The grouped patients have similar gene expression profiles.
BINNARY TREE-STRUCTURED VECTOR QUANTIZATION (BTSVQ)

Given the techniques available, the question remains, which clustering method is most useful for interpreting gene expression data? This decision can best be made by analysing mathematical properties of the data and biological relevance of the results. We experimented with several representative clustering approaches for microarray data analysis, including $k$-nearest neighbour, agglomerative hierarchical clustering, and SOMs. We also heuristically tested metric and semi-metric distance measures using these clustering techniques to find biologically relevant clusters, the partial evaluation of which we present in the next section.

Considering the inherent dispersion and skewness of microarray data, and its closeness to multivariate Gaussians (Hunter et al., 2001), we adopted an iterative partitive clustering approach which extends the TSVQ approach (Gersho and Gray, 1992). The algorithm partitions data using the standard $k$-means algorithm in sample space, where $k$ is kept constant at 2. Iteratively applying the algorithm and using evaluation of variance as a stopping criterion generates a binary tree. The SOM algorithm is then used to cluster the gene space. The cluster structure in gene space is visualized using component planes of the already computed SOM. An outline of the binary tree-structured vector quantization algorithm (BTSVQ) is given below.

1. B-Tree generation:
   (a) Start with the whole data set in a single cluster (parent).
   (b) Partition the data set (experiment space) into two subsets using standard $k$-means (simple Euclidean distance measure is used)
   (c) Compute variance of both subsets (children).
      • If variance of parent from step 1a is less than the variance of child from step 1b, stop further partitioning of that child.
      • Repeat step 1b for all remaining children, taking each child as a parent.

2. Visualization:
   (a) Generate component planes of SOM for all samples.
   (b) Arrange the component planes at nodes of the B-tree generated in step 1.

The algorithm uses self-organizing maps and partitive $k$-means clustering in a complimentary fashion. It applies the vector quantization and self-organization capabilities of SOMs to find significant gene centres in gene space, which is characterized by high dimensionality and a large number of clusters. BTSVQ uses the effectiveness of $k$-means in sample space, which is characterized by medium dimensionality and a low number of clusters. SOM algorithm finds gene centres by vector quantization and places genes with similar expression in neighbouring units in the two dimensional map grid. Thus, SOM preserves the similarity relationship in gene space. This quantization and re-arrangement of genes ensures that quantized genes would have similar expressions for similar samples. Thus, the component planes of similar samples would look similar. BTSVQ superimposes results from the two complimentary clustering approaches—the highest confidence in the clustering result is achieved when samples with visually similar component planes are placed in the same child of the cluster tree.

EXPERIMENTAL EVALUATION

We applied BTSVQ to the analysis of three large non-small cell lung cancer (NSCLC) sets:

1. Ontario Cancer Institute (OCI) data set of 40 patients with known clinical outcome (generated by us), processed using cDNA microarrays from the Toronto Microarray Centre (http://www.uhnres.utoronto.ca/services/microarray/) with 18 117 genes and ESTs,
2. Stanford data set of 24 patients with available clinical followup, using Stanford cDNA microarray technology with 22 696 genes and ESTs (Garber et al., 2001), and
3. MIT data set of 125 patients with known clinical followup, using 12 600 genes and ESTs on the U95A Affymetrix microarray (Bhattacharjee et al., 2001).

We hypothesized that information regarding clinically relevant molecular sub-classifications of NSCLC could be obtained from gene expression data. This approach revealed a number of novel observations on the pathology of lung cancer, most importantly, the presence of molecular subtypes of NSCLC correlating with differences in rates of tumor recurrence and disease-free survival. BTSVQ clustering results on the OCI data set are presented in Figure 4. As shown in the right-hand circled child in the cluster tree, the BTSVQ algorithm delineated a subgroup of patients not evident from agglomerative clustering of the same data (not shown) using all available gene expression data. A significant proportion of these selected patients is characterized by an early recurrence of disease, indicative of a more aggressive tumor biology. As highlighted in Figure 4c, 14/17 of these patients had evidence of early disease recurrence on clinical follow-up, compared to 2/7 in an analogous subgroup on the opposite arm of the cluster tree (Figure 4b).
Fig. 4. BTSVQ clustering of 40 gene expression profiles from OCI NSCLC patient data set on the basis of 18,117 genes. Clinical outcome data revealed 24 patients with evidence of tumor recurrence versus 16 patients who were disease-free. Both groups had a minimum 12 months follow-up with a mean of 24 months each. (a) Binary tree generated by BTSVQ algorithm with SOM component planes at the nodes. (b) Sub-group of patients enriched for those with no evidence of recurrence on clinical follow-up. (c) Sub-group of patients enriched for clinical evidence of early disease recurrence.
Fig. 5. BTSVQ clustering results of (Garber et al., 2001) data set; 24 patients were analysed on the basis of 22,696 genes. Clinical followup showed that 14 patients died due to a tumor-related cause (category 1), versus 10 patients who were either alive or dead due to a non-cancer related cause (category 0). (a) Binary tree generated by BTSVQ algorithm with SOM component planes at the nodes. (b) Circled cluster level 1, child 2 depicts 6/7 patients in category 1. (c) Circled cluster level 2, child 2 shows 7/11 patients of category 0.

Fig. 6. Evaluation of BTSVQ and Gene Cluster software (Eisen et al., 1998) sensitivity to log normalization, using a randomly selected subset from (Garber et al., 2001). (a) Expression profiles of original data set. (b) An agglomerative clustering of the original data groups genes 101,778 and 101,769 as similar. (c) The agglomerative clustering of the log-normalized data places genes 101,778 and 101,769 far apart. (d) Partitive k-means clustering detects the similarities in both situations. Here we show patient component planes of SOMs for genes. The full tree is for log normalized data, and although the two genes in question are visibly different, they are still placed in the same cluster. The small fragment of the tree in the lower-right corner shows results for the original data.
To further evaluate the applicability of BTSVQ for microarray data analysis, and to test the strength of clinical findings emerging from the OCI NSCLC data set, we have analysed the two other lung cancer data sets of Garber et al. (2001) and Bhattacharjee et al. (2001). Using all available genes we have found that in both data sets patient samples are split into groups enriched for differences in disease-free survival, further suggesting that gene expression profiles can be used to differentiate patients with more aggressive tumor biology.

Figure 5 shows the results for the Garber et al. (2001) data set. Clinical followup showed that 14 patients died due to a tumor-related cause (category 1), versus 10 patients who were either alive or dead due to a non-cancer related cause (category 0). Circled clusters, depicted in detail in Figure 5b and c indicate a segregation of gene expression profiles based on aggressiveness of tumor biology. The first cluster (level 1, child 2) represents 6/7 patients in category 1, while the second cluster (level 2, child 2) includes 7/11 patients in category 0. Similar results were also obtained from the MIT dataset. Clusters were evident with 7/12 patients dead due to recurrent disease with a mean survival of 34 months, compared with an analogous cluster on the same level containing 5 patients alive with greater than 60 months mean survival alongside 3 normal lung specimens.

Since clustering results can be affected by preprocessing, we tested how sensitive to data normalization agglomerative and partitive clustering methods are. During normalization, the data is transformed to remove various types of noise, biases and outliers. This often results in a new range of the data that is easier to work with in further analysis. The transformation may introduce several distortions and biases, some of which improve the information content, while others may eliminate existing valid patterns. Microarray data is generally log-normalized to provide an equal spread between up- and down-regulated genes. To evaluate the sensitivity of BTSVQ to normalization, we selected a subset of genes from the Stanford data set (Garber et al., 2001) at random, with the additional requirement that it contains at least a pair of genes with similar expression profiles. This subset was analysed with and without log normalization using both partitive (BTSVQ) and agglomerative approaches (Gene Cluster by Eisen et al. (1998)). BTSVQ was found to be less sensitive to data normalization as compared to agglomerative clustering, since it properly identified similarity between genes in data sets processed with and without log normalization (see Figure 6).

**DISCUSSION**

To uncover underlying similarities in microarray data, we require a robust analysis technique that will give biologically relevant data summarizations. In addition, we need an intuitive visualization of the results. We have experimented with a number of different approaches to improve the processes of data analysis and visualization.

We designed a hybrid algorithm BTSVQ that combines partitive k-means clustering with self-organizing maps. We demonstrated that BTSVQ is useful for analysing, visualizing and interpreting gene expression profiles. The main advantage of this approach is that it enables visualization of similarities and differences among individual samples and patient groups, using all available genes. We have also shown that BTSVQ is less sensitive to data normalization compared to agglomerative clustering. As a result, BTSVQ is applicable to a wide range of microarray data sets.

Further evidence of BTSVQ’s usefulness comes from an empirical evaluation of clustering techniques on 252 imperfect data sets (Mangiameli et al., 1996). The data sets included various levels of defects, such as data dispersion, outliers, irrelevant variables, and non-uniform cluster densities. Results of this evaluation show that k-means clustering works best when the number of clusters is medium and the dimensionality is low, while the SOM approach produces best results in high-dimensional problem domains with a large number of clusters and high dispersion. Furthermore, the k-means algorithm gives better clusters when the data distributions resemble multivariate Gaussians and the value of k is determined beforehand (Hunter et al., 2001).

Several other approaches are based on similar principles as BTSVQ. Cho et al. (1998) introduced a technique for manual visual inspection of clusters of gene expression patterns. This system has been used to cluster genes whose expression correlated with particular phases of the cell cycle, a process resembling that shown in Figure 3. The method is best suited for problems where the patterns are clear and the data set is relatively small. However, this approach is not practical for large data sets. In addition, unknown clusters cannot be found by visual inspection.

Oja et al. (2002) introduced a SOM approach for clustering and visualizing similarity relationships among yeast genes. They characterized relationships among clusters by calculating and comparing average gene expression profiles at each map area. The authors show that similar clusters are in neighbouring units of the two-dimensional map.

The BTSVQ algorithm differs from the above mentioned approaches by its ability to scale up to large problems. More importantly, it automates the process of cluster identification, and combines two algorithms into one. The BTSVQ system uses SOMs for visualization and the partitive k-means clustering for grouping similar component planes of SOMs and organizing them into binary tree structure. This serves as a solution to the manual clustering problem. It also enables one to find unknown clusters that are visually hard to detect (see...
Figure 4). Moreover, if the visually similar component planes are placed in the same child of the cluster tree, one gets an instant verification of results by two clustering approaches. We have observed a significant overlap between clusters identified by SOMs and the partitive k-means algorithm in the presented lung cancer data sets, as well as on other microarray data sets (unpublished observations).

Further improvements to the user interface and performance of the BTSVQ algorithm are underway to enable broader use of the proposed approach to cluster analysis by the microarray community.

ACKNOWLEDGEMENTS

This research was supported in part by the National Science and Engineering Research Council of Canada #203833–98, IBM Shared University Research grant, and IBM Faculty Partnership Award (IJ), Canadian Cancer Society and the National Cancer Institute of Canada #012150, the Physicians’ Services Incorporated Foundation R00-10, the PMH microarray Clinical Research Program, and CIHR post-doctoral fellowship (DW). The authors are indebted to Francis Shepherd from the PMH thoracic oncology group and Andrea Jurisicova for helpful comments on the earlier draft of this manuscript.

REFERENCES


