Cluster Inference Methods and Graphical Models Evaluated on NCI60 Microarray Gene Expression Data

Peter J. Waddell[1] **Hirohisa Kishino**[2]
waddell@cimmed.com kishino@wheat.ab.a.u-tokyo.ac.jp

[1] Chugai Research Institute for Molecular Medicine, 153-2 Nagai Niihari Ibaraki 300-4101, Japan
[2] Graduate School of Agriculture and Life Sciences, University of Tokyo, 1-1-1 Yayoi Bunkyo-ku, Tokyo 113-8657, Japan

Abstract

At present, there is a lack of a sound methodology to infer causal gene expression relationships on a genome wide basis. We address this first by examining the behaviour of some of the latest and fastest algorithms for tree and cluster analysis, particularly hierarchical methods popular in phylogenetics. Combined with these are two novel distances based on partial, rather than full, correlations. Theoretically, partial correlations should provide better evidence for regulatory genetic links than standard correlations. To compare the clusters obtained by many alternative methods we use tree consensus methods. To compare methods of analysis we used tree partition metrics followed by another level of clustering. These, and a tree fit metric, all suggest that the new distances give quite different trees than those usually obtained. In the second part we consider graphical modeling of the interactions of important genes of the cell cycle. Despite the models seeming to fit well on occasions, and despite the experimental error structure seeming close to multivariate normal, there are considerable problems to overcome. Latent variables, in this case important genes missing from the analysis, are inferred to have a strong effect on the partial correlations. Also, the data show clear evidence of sampling distributions conditional on the status of important cancer related genes, including TP53. Without full information on which genes are wild type the appropriate models cannot be fitted. These findings point to the need to include and distinguish not only all relevant genes but also all splice variants in the design phase of a microarray analysis. Failure to do so will induce problems similar to both latent variables and conditional distributions.

Keywords: microarray, gene expression clustering, partial correlation distance, graphical modeling

1 Introduction

Microarray data monitors the expression level of a wide assortment of genes simultaneously. As a consequence, the amount of data generated is very large in comparison to other similar techniques used in molecular biology (e.g. Northern blots). Given such a fountain of data, ideally we would like analysis techniques that would infer direct causal relationships between genes. At present, the most widely used analyses are based on a correlation-derived distance and an ultrametric clustering method such as UPGMA [4].

One question that has yet to be explored in detail is how do UPGMA trees compare with trees derived by other methods? In particular, due to recent algorithmic improvements, least squares based methods of tree inference have been increased in speed by a factor of N or more [2]. With 1,000 genes this improves speed by $\sim$ 1,000 times, and makes such criteria based methods feasible for thousands of genes. Unlike UPGMA, these methods do not need to assume that expression derived distance data are ultrametric, and indeed there is no good reason to expect the distances between genes are ultrametric. Methods that assume ultrametric data often lead to errors when the data are not ultrametric [19].

To look for causal links between genes a crucial concept is that of partial correlation. That is, measuring the correlation between genes i and j after the common effects of genes $k, l, \ldots$, or all other genes in the genome, are removed. Like correlations, partial correlations range from -1 to 1. Partial correlations that remain significantly different from zero may be taken as indicators of a possible causal link. It would be interesting then to cluster genes based on their partial correlations with each other, rather than just their correlations. On trees that do not assume ultrametric distances, genes that are connected by a short path through the tree should be good candidates for further study of more direct interactions.

The area of graphical modeling has made considerable advances over the past two decades. If there are a small to moderate number of variables (genes), and assuming errors follow a member of the exponential multi-variate family, detailed predictions can be made from the models. For example, the likelihood of the data under the model can be calculated [3,9,20]. Theoretically, graphical models seem like a good possibility for elucidating genetic pathways from microarray data. The near multivariate normal nature of experimental errors in log transformed gene expression ratios is one such encouraging factor. Another useful feature of graphical models is that they can use directly the continuous data reported by microarray experiments, something Bayesian Networks [7] have yet to do.

As a first step, it is interesting to apply analysis methods based on partial correlations to the latest expression data coming from Stanford, NCI, etc. Due to strong interest from fellow researchers, analyses are concentrated on a set of 50 preselected cancer related genes from the data of Ross et al. and Scherf et al. (2000)[14,16] measured on 60 NCI cancer cell lines. These data are confirmed as having generally low experimental error rate, i.e. a coefficient of variation of $\sim 20\%$, and represent expression levels from 60 cell lines derived from cancers of many types [14,16]. Our findings suggest that to confidently infer genetic networks with microarray technology, experimental design needs to take into account a number of factors, including the need to *a priori* attribute splice variants the same status as genes.

What is partial correlation?

The partial correlation of variables x and y with respect to z may be considered to be the correlation of x and y after the effect of z is removed. For example, imagine that gene z influences the transcription of gene x to the extent that their correlation is 0.7, while the effect of z upon y results in their having a correlation of 0.8. Even if x and y have no direct relationship to each other, then due to their mutual correlation with z, they will have a correlation of 0.56. Partial correlation subtracts the 0.56 to reveal that, in this instance, x and y are conditionally independent. The partial correlation of x and y with respect to z may be expressed as $\mathrm{pr}_{xy.z} = (r_{xy} - r_{xz}r_{yz})/((1 - r_{xz}^2)(1 - r_{yz}^2))^{1/2}$. Like covariances and correlations, partial covariances and partial correlations go together. The partial covariance of genes x and y with the effect of z removed is, $\mathrm{pcov}_{xy.z} = \sigma_{xy}^2 - \sigma_{xz}^2\sigma_{yz}^2/\sigma_z^2$. When x and y are conditionally independent, both $\mathrm{pcov}_{xy.z}$ and $\mathrm{pr}_{xy.z}$ take value zero. With a whole genome of genes, we would really like to know the partial correlation of gene x and y after discounting the effects of all other genes, set g (i.e. $\mathrm{pc}_{xy.g}$). With more than three genes the calculations become more complicated, but can be expressed and solved using matrix notation. That is, $\mathrm{pr}_{xy.g} = -w_{xy}/(w_{xx}w_{yy})^{1/2}$, where $w_{ij}$is the ij-th element of $\mathbf{W} = \mathbf{V}^{-1}$ (the inverse of the covariance matrix of all genes).

Partial correlations also share a close relationship with multiple regression. If the expression of each gene x for all the m experiments is written as a vector, $\mathbf{x}$, it may be then be linearly regressed as the dependent variable against all other genes expression. For example, $\mathbf{x} = a_x + b_{xy.z}\mathbf{y} + b_{xz.y}\mathbf{z}$, where the first subscript indicates that x is the dependent variable, the second indicates the identity of the explanatory variable, and the last part after the period indicates the other explanatory variables in the regression. The relationship $(\mathrm{pr}_{xy.z})^2 = \beta_{xy.z}\beta_{yx.z}$ emerges, so $\mathrm{pr}_{xy.z} = \beta_{xy.z}/\mathrm{abs}(\beta_{xy.z})(\beta_{xy.z}\beta_{yx.z})^{1/2}$, where "abs" indicates absolute value (see chapter 27 of [17] for formal derivations). Such relationships are not used herein, but see [8] for examples of their use in gene expression analysis. When there

is insufficient data to solve the regression equations with all genes as explanatory variables, it is necessary to pick just those that seem to be having a detectable effect using something like the AIC model selection criterion [8]. We call such approaches Approximate Partial Correlations via Regression (APCR).

If the value of $\mathrm{pc}_{xy.g}$ is statistically indistinguishable from zero, then there is no evidence of a direct genetic link between x and y. If $\mathrm{pc}_{xy.g}$ is significant, then a direct link or edge may be added between them (which may be of either positive or negative sign). In this way, whole genetic pathways may be built up. The process of adding in links or edges is known as graphical modeling, and more derived forms of this type of procedure include Bayesian Networks [7]. Graphical modeling has a long history reaching back to path analysis pioneered by Sewall Wright in the 1920's and used for none-other than inferring the causal relationships between genes (e.g. Wright [21]).

2 Methods

The data set consists of 9,703 cDNA's whose expression levels were determined in 60 cancer cell lines and then transformed to log ratios [14,16]. From this data we are looking particularly at genes associated with cancer, a fair number of which are involved in control of the cell cycle. Common abbreviated names of these genes are shown in Figure 1 (for the full list see http://www.jsbi.org/). Also included were a few housekeeping genes plus 30 genes selected completely at random; these are used to help diagnose attributes of the data. Samples that showed very low expression levels and were assigned a ratio of minus infinity by Ross *et al.* (0 to 6 such cases per gene studied here) were set to a value of -7. Far fewer entries were listed as NA, where the experiment had failed completely at a spot, and these were set to value 0. The analyses of [14, 16], and our own analyses of similar data, suggest that the log-transformed data have multi-variate normal errors.

For calculating correlations and partial correlations, the programs SPlus and MIM [3] were used, and these were also used for graphical modeling. Based on correlations, r, and partial correlations, pr, we obtain 4 distances: $\delta_r = 1 - r$ and $\delta_{pr} = 1 - pr$, with range 0 to 2; $\delta_{|r|} = 1 - |r|, \delta_{|pr|} = 1 - |pr|$, with range 0 to 1 (see Eisen *et al.* [4] for implementations of the first and third distances).

For tree building PAUP*4.0b4 [18] was predominantly used. The methods used include UPGMA and NJ which are algorithmic hierarchical clustering methods, an advantage of which is that they run quickly, but cannot explore many alternative solutions. Ordinary least squares (OLS) and Fitch-Margoliash (FM) or inverse-distance weighted least squares, are criteria based. They evaluate one tree at a time and rely upon effective searching of the tree space to find the best solution(s) [19]. OLS assumes that expected errors on distances are constant, while FM assumes that errors on distances grow in proportion to the distance. Since correlations near 1 or -1 have the smallest errors, while correlations near 0 have the largest errors, this is an approximately reasonable assumption when using the distance $1-|r|$. (Note, the parameter setting 0 of the program FITCH [6] also gives OLS, while the setting 2 gives FM, c.f. [10]). The last method considered is minimum evolution, ME. This approach optimises edge (branch or internode) lengths on a given tree using OLS, but then discards the residual least squares and uses the sum of the edge lengths as the criterion of fit. Each criteria based method can either allow or not allow negative edge length estimates. To distinguish these, a "+" indicates when negative edge lengths are prohibited, e.g. OLS+. The algorithms in PAUP* for implementing ME, OLS, and FM use the time optimal methods described in Bryant and Waddell [2].

3 Results

3.1 Clustering of Genes

For the purpose of examining how hierarchical cluster, or tree analysis, describes the relationships between these genes, the best tree inference methods identified in a previous study were used (Waddell

and Kishino unpublished). They were NJ, OLS+, FM+, ME+ and ME− combined with the Pearson correlation, r, distances, δ_r and $\delta_{|r|}$. The second distance being important when considering the transcriptional interaction of genes that are antagonistic. None of these tree-building methods assume the distance data are ultrametric. Clustering methods, which had previously shown lesser performance such as UPGMA (used by [4,14]), OLS− and FM−, were also considered.

Reported here is a new type of clustering for gene expression data where distances are functions of the partial correlations (pr) of genes. Clusters based on such distances are particularly interesting since small, close, and tight clusters should, in expectation, often translate to prominent features of graphical models. Construction of such distance matrices is limited by the rank of the covariance matrix of the genes, i.e., whether the correlation matrix can be inverted. This presently limits us to clustering at any one time fewer genes than there are discrete experiments. The APCR method of Kishino and Waddell [8] for approximating partial correlations avoids this limitation and should allow genome wide assessment in future.

An improved understanding of the behaviour of tree algorithm/distance combinations for this type of data is important, as selection of one method over another can result in considerable differences. One way to assess tree inference methods is to consider the fit of tree to data. For this purpose, the residual sum of squares from OLS+ given the selected tree topology is used. To negate the effect of distance scaling the last value is reexpressed as a percentage distance error [6, 19].

There is a considerable improvement in fit of least squares estimators when the distance used the absolute value of the correlation measure (Table 1). This indicates that, as one might suspect, the absolute value based distances are more natural descriptions of transcriptional relations between genes. In contrast, the difference in fit with and without the non-negativity constraint in tree inference was generally minor. However, a difference is notable when using δ_{pr}, suggesting the need for extra caution whenever interpreting trees based on this distance. In column 3 of Table 1, the best trees for each method had their fit measured with an OLS+/δ_r combination, even though most were selected using other distances. As expected, the fit was best on the trees which were originally selected using δ_r and deteriorated consistently going to trees from other distance measures. However, even at worst, the fit did not approach anywhere near that of randomly generated trees evaluated on the same data. The same pattern holds for the OLS+/$\delta_{|pr|}$ fit. The OLS+ %s.d. fit is also acting like a tree comparison metric; column3 suggests that trees based on $\delta_{|r|}$ are more similar to trees based on δ_r than are trees based on partial correlation, while column 4 suggests that trees based on $\delta_{|pr|}$ are more similar to trees based on δ_{pr} than are trees based on correlation. The worse fit of trees based on partial correlation may reflect the increased sensitivity of partial correlations to stochastic error (since they are based on the inverse of the correlation matrix).

It is important to consider which cluster details are invariant with respect to the clustering method used, since tree building or clustering of gene expression data is exploratory and somewhat arbitrary. Accordingly, in Figure 1 strict consensus trees are shown for all trees from each distance measure. Generally, small clusters near the tips are most consistent between methods. A feature found consistently only in the trees based on correlation is a cluster of TP53, E2F4, and 14-3-3e. The trees that use partial correlations show quite different clusters from those seen with δ_r and $\delta_{|r|}$. None of the distance data observe the ultrametric condition, and this seemed particularly so with distances based on partial correlations. Thus, in general, it is advantageous to use methods such as NJ and OLS+ in preference to UPGMA.

Given so many different trees, each comprised of many clusters, it can be difficult to make general statements about them. One way to measure relationships between trees is using a tree-to-tree distance like the symmetric-difference metric [13], which counts groups in common between each pair of trees. To visualize the relationships within this matrix we applied OLS+ to obtain the tree in Figure 2. The results suggest:

1. Trees from a given distance tend to be much more similar to each other than trees from other distance measures.

Table 1: Fit of trees to the data.

Correlation Distances					Partial Correlation Distances				
Method	Native Fit[a]	%s.d.[b]	OLS+ %s.d./δ_r[c]	OLS+ %s.d./$\delta_{\|pr\|}$	Method	Native Fit*	%s.d.	OLS+ %s.d./δ_r	OLS+ %s.d./$\delta_{\|pr\|}$
UPδ_r	-	-	14.456	17.031	UPδ_{pr}	-	-	17.612	16.435
NJδ_r	-	-	15.007	17.001	NJδ_{pr}	-	-	17.437	16.451
OLS-δ_r	18.041	13.209	15.145	17.006	OLS-δ_{pr}	36.462	18.778	17.881	15.205
OLS+δ_r	21.238	14.331	14.331	17.000	OLS+δ_{pr}	184.932	42.291	17.408	16.499
FM-δ_r	19.364	13.685	15.142	16.932	FM-δ_{pr}	62.538	24.593	17.931	15.260
FM+δ_r	22.853	14.867	14.363	17.032	FM+δ_{pr}	165.514	40.009	17.640	12.679
ME-δ_r	16.322	-	14.528	17.027	ME-δ_{pr}	15.627	-	17.447	16.453
ME+δ_r	16.322	-	14.528	17.027	ME+δ_{pr}	15.894	-	17.440	16.458
UP$\delta_{\|r\|}$	-	-	16.768	16.926	UP$\delta_{\|pr\|}$	-	-	17.896	12.750
NJ$\delta_{\|r\|}$	-	-	16.619	16.951	NJ$\delta_{\|pr\|}$	-	-	17.783	13.948
OLS-$\delta_{\|r\|}$	7.524	8.530	16.809	17.011	OLS-$\delta_{\|pr\|}$	13.650	11.490	17.710	12.285
OLS+$\delta_{\|r\|}$	8.159	8.883	16.544	16.940	OLS+$\delta_{\|pr\|}$	15.612	12.288	17.746	12.904
FM-$\delta_{\|r\|}$	12.010	10.777	16.920	17.010	FM-$\delta_{\|pr\|}$	38.230	19.228	17.688	12.901
FM+$\delta_{\|r\|}$	13.399	11.384	16.502	16.963	FM+$\delta_{\|pr\|}$	44.795	20.814	17.655	12.432
ME-$\delta_{\|r\|}$	15.236	-	16.390	16.931	ME-$\delta_{\|pr\|}$	8.607	-	17.703	12.774
ME+$\delta_{\|r\|}$	15.236	-	16.390	16.931	ME+$\delta_{\|pr\|}$	8.646	-	17.705	12.702

[a] Native Fit is the score of the optimal tree on the data for the tree inference/data combination indicated in the left column, note these are in different units for different prefixes; [b] %s.d. (column 2) is given where calculable for the original tree and edge (branch) lengths on the original data; [c] the last two columns are refitting of the tree topology using OLS+ on the data indicated, with fit again reported by %s.d. of observed to tree distances.

2. That the new distances, $1-pr$, $1-|pr|$, produce quite distinct trees from those generated using previous methods, and these trees are more similar to each other than distances based on Pearson correlation, r.

3. That the trees previously identified with the best tree inference methods for this type of data (i.e. NJ, OLS+, FM+, ME, do typically tend to cluster together (with the exception of FMpr).

4. That trees of the most different methods are very close to the maximal possible tree partition distance apart, i.e. in topology very little in common.

5. The ME methods behave similarly, irrespective of whether they have a non-negativity constraints, and share a more distant relationship with NJ. In contrast, the least squares methods behave quite differently, to the extent that FM+ and OLS+ are distinct but more similar to one another than either is to least squares allowing negative edges.

6. The diversity of the different trees from the same distances can quickly be ascertained using NJ, OLS and OLS+.

As a technical note, using a tree to relate tree-building methods seems to be quite informative. An advantage of tree to express the relative properties of objects is that the final expression is insensitive to the scaling and rotation effects that occur with multi-dimensional scaling techniques.

Taken together, the large partition distance between trees but the generally reasonable OLS+ fit of all trees on any distance might seem contradictory. However, the interpretation that seems most reasonable is that all the trees contain sets of genes that wander about considerably (backed up by reference to the trees themselves), but some general structure is preserved in all the analyses. Additionally, edge lengths should also be taken as part of the reading of such cluster data, and is also a reason for favouring methods that allow all edges to have independent lengths, herein all methods

Figure 1: Consensus trees of trees based on different distances, (a) δ_r, (b) δ_{pr}, (c) $\delta_{|r|}$, (d) $\delta_{|pr|}$. The housekeeping genes are A2Ma, B2Mi, Ker19, and HLA1F.

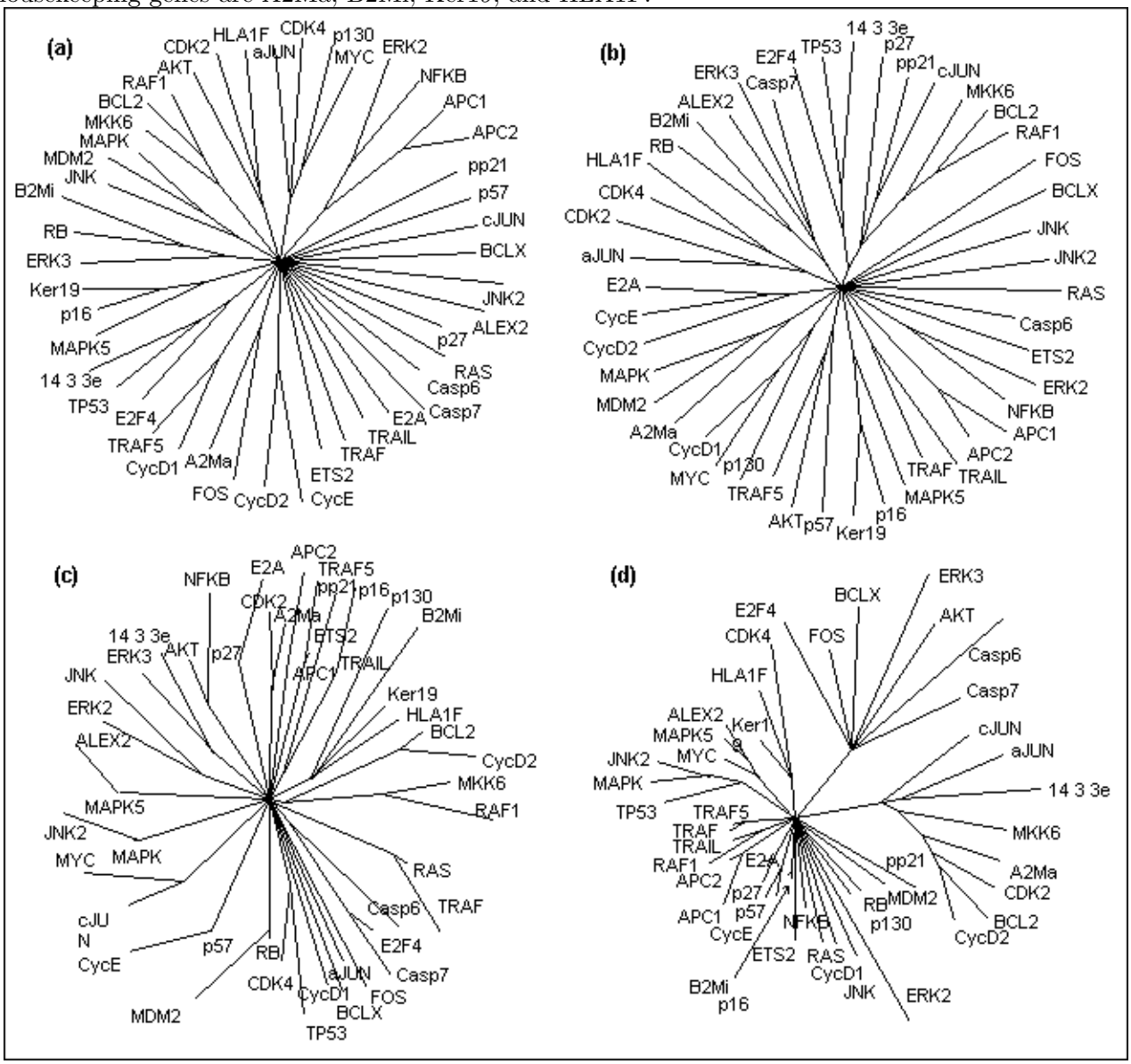
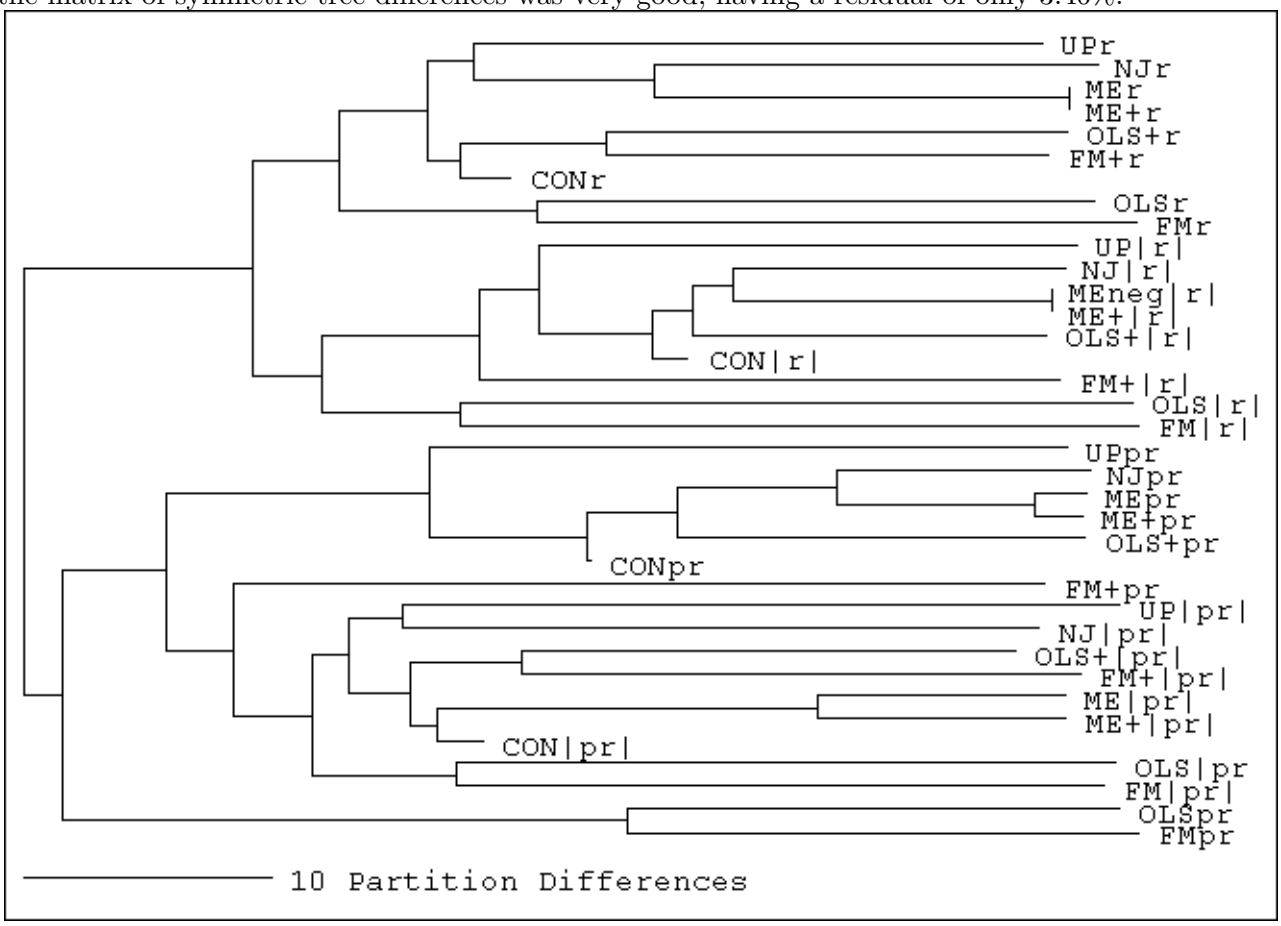


Figure 2: An FM+ tree relating the tree inference methods together based on the partition metric distance between trees. The root is at the mid-point. CON denotes strict consensus trees. The fit to the matrix of symmetric tree differences was very good, having a residual of only 3.46%.



except UPGMA. Thus, a gene at the end of a short terminal edge may not be directly clustered with another gene, but the path between them may be short, so it should be expected that they might have a similar pattern of transcriptional expression. Mapping distances back onto trees confirms the trees are a reasonable, but not perfect, representation of the raw distances.

3.2 Graphical models

The data exhibit a slight sparsity problem due to missing cells. While the matrix dimension is 60, the rank of the variance-covariance matrix is approximately 40 to 50, depending upon which genes are included. Graphical modeling of this many variables is a challenging exercise, so results should be regarded as exploratory until the underlying nature of the data is known to be well matched to the model being used [3, 9].

Application of nearly any of the refined graphical modeling techniques to this data leads to large numbers of significant links, e.g. > 25 per gene. Thus, with 50 or so genes, the resulting graph is highly linked and looks like a string ball. Clearly, the hope that picking a set of *a priori* genes and getting a clear and informative model needs to be tempered. A key feature is that the data show a poor fit to the model (e.g. deviance of 2000 with degrees of freedom 1000). While there is strong evidence of poor fit, it should be remembered that phylogenetic models, for example, fit most data very poorly, but are often of considerable accuracy and utility even on occasions. Unfortunately, none of the usual

variations on the model, e.g. Box-Cox transformations, AIC or BIC or F ratio edge selection criteria, forward or backwards edge selection, adjusting the models for sparsity, or making edge inclusion more stringent e.g. P=0.01 versus P=0.05, results in a graph that seems any more intuitively, or biologically, pleasing. For example, increasing stringency of edge selection to 0.001 would often see a huge drop in the number of edges in the graph, but different methods did not agree on the model then selected. Further, unexpected links between cancer and housekeeping still occurred at this higher stringency.

Latent variables are a likely problem since there are many important genes missing from the data, for example just one member of the E2F family of six was found on the array. To assess their probable effect, we looked at the partial correlations between 30 randomly selected genes. The distribution of partial correlations had more than 5% of values with absolute values of nearly 0.3, as opposed to 5% greater than an absolute value of 0.2 as expected under the model. This observation also concurs with the rapid decrease in the number of selected links as the stringency went from 0.05 towards 0.001.

3.3 Cell cycle and the status of TP53

A subset of the cancer related genes selected are close to the cell cycle, including TP53. Since the cell lines examined by Ross et al. were all derived from advanced cancers, then experienced further selection in culture, activity-altering mutations are a concern. For example, loss of function TP53 mutants are often over expressed compared to wild type (WT) TP53, and indeed only 15 of the cell lines considered have WT TP53 [12]. Unfortunately, such data is not available at present for other genes in all cell lines. Thus, there are good reasons to believe that parts of the data will not behave in a rational way with regard to inferences based on observed correlation and related measures.

To examine this factor further, we apply graphical modeling to a subset of 11 genes, which are close to cell cycle regulation, then look at the differences in expression pattern between mutant and WT TP53 cell lines. The small number of genes is due to there being only 15 WT lines, so that, as discussed previously, the partial correlation matrix is not well conditioned with more variables. Interestingly, mutant TP53 cell lines did not show over-expression of TP53 with respect to WT TP53 lines, as has been reported previously, e.g. [22]. Rather, the mutant lines show a greater dispersion of TP53 levels than WT lines, a factor partially explained by their three times greater number.

The graphical model from coherent backward selection using F ratio tests and P=0.05 is shown in Figure 3(a) and next to it in 3(b) is the model selected from the mutant TP53 cell lines. The estimated models are quite different, and neither shows features one might expect such as positive links between E2F4 and TP53, TP53 and WAF1, or TP53 and BAX [22]. The fit of the model is marginal for the wild type data, but acceptable for the mutant data. As Figure 4 shows, generally, the correlations amongst genes in the mutant cell lines are smaller than in the WT lines, but due to there being far fewer data points with the latter, more features are attributed significance in the mutant data. Consistent with the quite different models selected, there is no apparent relationship between pairwise correlations estimated with each class of cell line, and also no apparent relationship between the inferred partial correlations (Figure 4). Examination of pairwise plots of gene expression verifies that most correlations and pairwise relationships are not striking, and are often exaggerated considerably by a single outlying observation.

One relationship that has changed noticeably in the WT lines is the appearance of a negative correlation between TP53 and WAF1 (borderline significant at P=0.06 when each is regressed separately). In the mutant lines there is no significant pairwise relationship. WAF1 is a potent cancer suppressor gene activated by WT TP53. Thus, if TP53 is still active in a cancer cell it makes sense that its agent is not. The usual prediction is that WAF1 would be mutated, although a lingering mystery surrounds failure to find somatic mutations of this gene in most cancers (however, there is evidence that mutant population variants of WAF1 are more common when TP53 is wild type, [11]). A couple of other genes that show borderline significant changes in expression between WT and mutant cell lines are p16 (from no to a positive association with TP53) and CycD (from no to a negative

Figure 3: Graphical models of cell cycle gene interactions in cell lines containing either wild type or mutant TP53. (Note, dashed lines indicate a negative partial correlation).

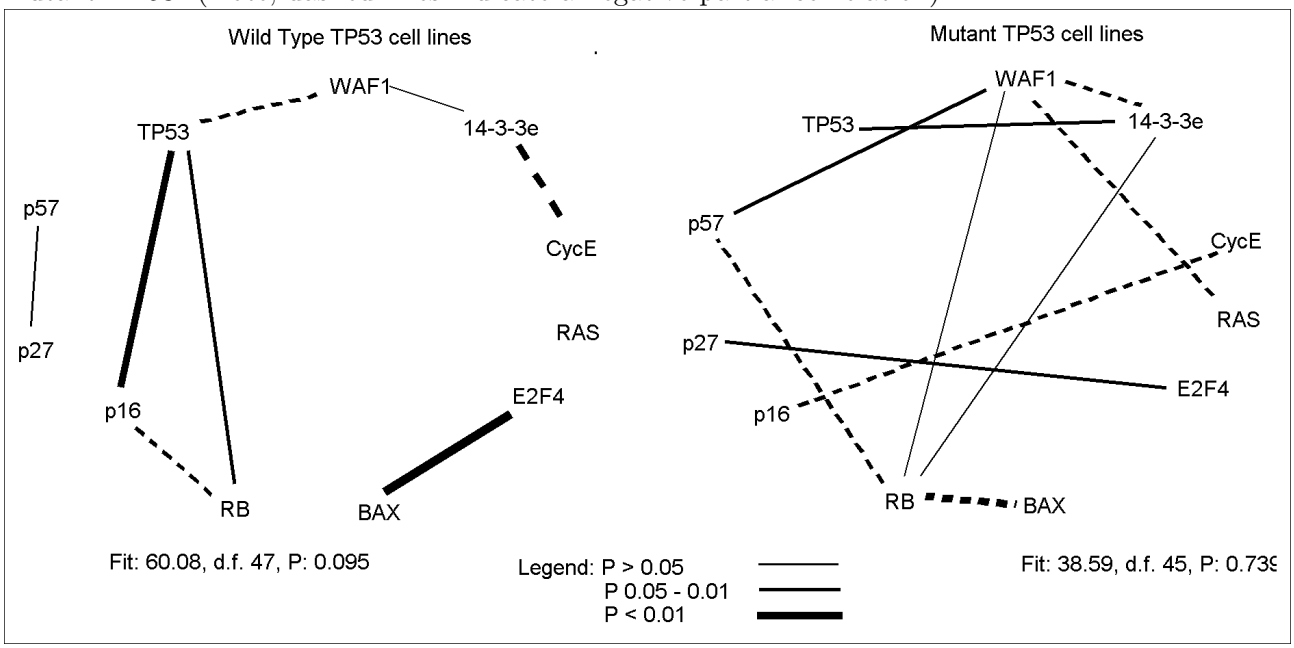
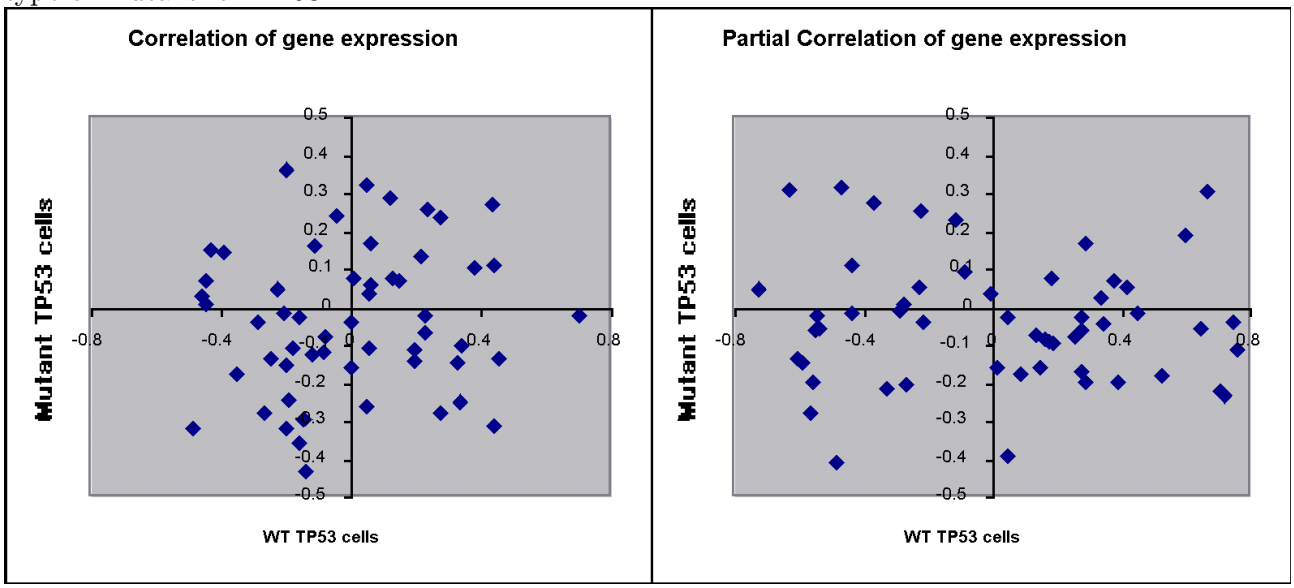


Figure 4: Estimates of correlations and partial correlations coming from cell lines that are either wild type or mutant for TP53.



association). An interesting feature lacking in the cell lines mutant for TP53 is the frequently reported inverse relationship of p16 and Rb expression levels [5]. This feature does show up in the graph of WT TP53 cells, due to a negative partial correlation, although the direct pairwise relationship is unconvincing (and the numerical significance being due largely to a single outlier). Based on predictions and observations taken together, there seems to be a good case that important genes in this data are showing conditional distributions which are dependant upon their mutation status, and modeling them in cancer cell lines would at least require mutations' status in all cell lines. It also seems likely that there is a fair bit of "washing over" caused by multiple complex conditional distributions, so that the unconditional distributions show little evidence of any particular trend.

4 Discussion

Ross *et al.* [14] and Scherf *et al.* [16] did not emphasise the potential impact of mutations in cell lines upon their own analyses. However, it was found above that the status of a gene (here TP53) in a particular cell line, lead to quite different observed correlations between genes. In the database of Ross et al., TP53 is associated with a list of mostly unnamed genes showing the best numerical correlation with it. However, many of these genes show only partly resolved expression patterns, that is many −inf and NA results in experiments. These must be tempting targets for further functional elucidation. Yet, given that TP53's expression level seems to be decoupled from its normal behaviour, it is uncertain what to make of such associations. Add to this uncertainty the fact that there are so many genes on the array, so that some by chance will show a correlation with TP53, plus partial data for some genes, it would not be surprising if most of this list were false positives in the sense they do not have any direct association with TP53.

What is the information content of the Ross et al. data with respect to modeling genetic links? While the data seems technically excellent, with a claimed coefficient of variation due to experimental errors of approximately 20–30%, a number of lines of evidence suggest the information content is low in its present form. These include very weak pairwise relationships, including cases of genes expected on prior data to have a clear relationship (e.g. TP53/Waf1, p16/Rb). Another is the lack of any consistency of relationship for pairs of genes between the TP53 wild type and the mutant cell lines. Expression profiles being conditional on the mutation status of a selection of key genes are expected to be a considerable part of the explanation. Expression profiles of a large number of genes may allow us to classify cancers successfully, even if the data is accompanied by large systematic errors. However, these biases seriously mislead us in the search for genetic links and functional relationships between important genes.

The results offer valuable angles on the design/analysis of microarray experiments. If we are to use microarray data to explore directly the causal links between human genes, and not just to offer suggestions for bench experiments to follow up, then experimental design needs to be considered. For example, if cancer cell lines are going to be used, then the mutation status of genes needs to be catalogued. Given the difficulty of doing this, it tends to suggest use of wild type cells and tissues in preference to cancer cells and cell lines in many experiments. Further, the impact of latent variables has to be minimised, so the array must be designed to include any gene suspected of being in a neighboring pathway or process of interest.

We predict that, due to a combination of the mutant conditional distribution effect and the latent variable effect, it will be especially important to have all splice variants of a pathway present and monitored. Splice variants are being recognised as increasingly important in many regulatory processes, e.g., [1, 15]. Splice variants can often act in antagonistic ways, e.g. a gene of exons A and B, has a splice variant A which lacks B, the catalytic domain, and effectively shuts down the action of AB due to competitive binding. Thus, for modeling regulatory pathways splice variants should, as far as possible, be accorded the status of the unit of interest. A challenge ahead is to catalogue such variants and implement them on microarrays. For this purpose, oligonucleotide arrays seem to have a natural

advantage over cDNA arrays. However, for many genes, carefully selected partial cDNA sequences will allow cDNA microarrays to detect and distinguish known splice variants.

Beyond mutant genes and splice variants, there are other factors know to make expression patterns highly conditional. An important consideration is how much protein has built up in the cell. This can be measured using protein arrays, and this is usually considered an aspect of proteomics. In addition, the behaviour of proteins is conditional upon factors such as phosphorylation status, cell localization, and the local frequency of binding partners. When these factors come to dominate the direct effects of transcription regulation, it is apparent that the input to the model of cellular pathways needs to grow. Clearly, we are on the road to full-scale models of a cell. Protein frequencies can be incorporated into the same types of framework discussed here, and in addition, there is the challenge to accommodate conditional expression levels and protein levels simultaneously. Accordingly, distinctions between proteomics and expression studies disappear at the analysis level, and each must be aware of the potential impact of the other when elucidating pathways.

In conclusion, the methods presented here offer useful ways to progress our understanding of microarray data. Clustering based on partial correlations, or APCR based distances [8], should be useful steps towards finding a sufficiently tight and closed set of genes to use for graphical modeling, or any other method aimed at uncovering causal relationships. The use of model selection criteria such as AIC to limit the estimation of partial correlations in APCR to just the envelope of genes with significant impact [8], should reduce stochastic noise in partial correlation distances and improve the robustness of clustering using partial correlation distances. Graphical modeling itself, seems a natural partner to detailed expression data, and has the advantage of being able to model conditional distributions of continuous variables. Extensions to the methods presented may also help to tackle the multi-layered problem of integrating gene expression plus both protein expression levels and their conditional states, e.g. phosphorylated or not. Together these define a much more complete picture of most regulatory pathways. An important step in moving to that level is to understand why data don't match the expectations of models, and considering to what extent the data gathering can be made more appropriate to the models at hand. It is critical to get the nature of the experiment correct, rather than hoping that large volumes of data and databases will make gene interactions apparent.

Acknowledgements

This work was supported by Chugai Research Institute for Molecular Medicine and H.K. was partly supported by grant 1255407 and BSAR-497 from the Japan Society for the Promotion of Science. The authors thank Drs Miyuki Shimane and Hitoshi Nomura for their comments on the manuscript.

References

[1] Atamas, S.P, Alternative splice variants of cytokines: Making a list, *Life Sciences*, 61:1105–1112, 1997.

[2] Bryant, D. and P.J. Waddell, Rapid evaluation of least squares and minimum evolution criteria on phylogenetic trees, *Molecular Biology and Evolution*, 15:1346–1359, 1998.

[3] Edwards, D., *Introduction to Graphical Modeling*, Springer-Verlag, New York, 1996.

[4] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[5] Fang, X., Jin, X., Xu, H.J., Liu, L., Peng, H.Q., Hogg, D., Roth, J.A., Yu, Y., Xu, F., Bast, R.C. Jr, and Mills, G.B, Expression of p16 induces transcriptional downregulation of the RB gene, *Oncogene*, 16:1–8, 1998.

[6] Felsenstein, J., *PHYLIP (Phylogeny Inference Package) and manual, version 3.5c*, Department of Genetics, University of Washington, Seattle, 1993.

[7] Friedman, N., Linial, M., Nachman, I., and Pe'er, D., Using Bayesian networks to analyse expression data, *RECOMB 2000*, Tokyo, Conference Proceedings, Academic Press, 2000.

[8] Kishino, H., and Waddell, P.J, Correspondence Analysis of Genes and Tissue Types and Finding Genetic Links from Microarray Data, *Genome Informatics*, 11, 2000.

[9] Lauritzen, S.L, *Graphical Models*, Oxford Statistical Science Series, 17, 1996.

[10] Michaels, G.S., Carr, D.B., Wen, X., Fuhrman, S., Askenazi, M., and Somogyi, R, Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data, *Pacific Symposium on Biocomputing*, 3:42–53, 1998.

[11] Mousses, S., Ozcelik, H., Lee, P.D., Malkin, D., Bull, S.B., and Andrulis, I.L, Two variants of the CIP1/WAF1 gene occur together and are associated with human cancer, *Hum. Mol. Genet.*, 4:1089–1092, 1995.

[12] O'Connor, P.M., Jackman, J., Bae, I., Myers, T.G., Fan, S., Mutoh, M., Scudiero, D.A., Monks, A., Sausville, E.A., Weinstein, J.N., Friend, S., Fornace, A.J. Jr, and Kohn, K.W, Characterization of the p53 tumor suppressor pathway in cell lines of the National Cancer Institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents, *Cancer Research*, 57:4285–4300, 1997.

[13] Robinson, D.F. and Folds, L.R., Comparison of phylogenetic trees, *Mathematical Biosciences*, 53:131–147, 1981.

[14] Ross, D.T, Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., and Brown, P.O, Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, 24:227–235, 2000.

[15] Saxon, A., Diaz-Sanchez, D., Zhang, K., Regulation of the expression of distinct human secreted IgE proteins produced by alternative RNA splicing, *Biochem. Soc. Trans.*, 25:383–387, 1997.

[16] Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O., and Weinstein, J.N, A gene expression database for the molecular pharmacology of cancer, *Nature Genetics*, 24:236–244, 2000.

[17] Stuart, A. and Ord, J.K., *Kendall's advanced theory of statistics, fifth edition, volume 2: Classical inference and relationship*, Edward Arnold, London, 1991.

[18] Swofford, D.L., *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0b4*, Sinauer Associates, Sunderland, Massachusetts, 2000.

[19] Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M., Phylogenetic Inference, Pp. 407-514 in Hillis, D.M., Moritz, C., and Mable, B.K., eds. *Molecular Systematics (second edition)*, Sinauer Associates, Sunderland, Massachusetts, 1996.

[20] Whittaker, J, *Graphical Models in Applied Multivariate Statistics*, Wiley, New York, 1990.

[21] Wright, S., The theory of path coefficients: A reply to Niles' criticism, *Genetics*, 8:239–249,

[22] Yu, J., Zhang, L., Hwang, P.M., Rago, C., Kinzler, K.W., and Vogelstein, B., Identification and classification of p53-regulated genes, *Proc. Natl. Acad. Sci. USA*, 96:14517–14522, 1999.