

Correspondence Analysis of Genes and Tissue Types and Finding Genetic Links from Microarray Data

Hirohisa Kishino¹

kishino@wheat.ab.a.u-tokyo.ac.jp

Peter J. Waddell²

waddell@cimmed.com

¹ Graduate School of Agriculture and Life Sciences, University of Tokyo, 1-1-1 Yayoi Bunkyo-ku, Tokyo 113-8657, Japan

² Chugai Research Institute for Molecular Medicine, INC., 153-2 Nagai Niihari Ibaraki 300-4101, Japan

Abstract

In this paper, we propose and use two novel procedures for the analysis of microarray gene expression data. The first is correspondence analysis which visualizes the relationship between genes and tissues as two 2 dimensional graphs, oriented so that distances between genes are preserved, distances between tissues are preserved, and so that genes which primarily distinguish certain types of tissue are spatially close to those tissues. For the inference of genetic links, partial correlations rather than correlations are the key issue. A partial correlation between i and j is the relationship between i and j after the effect of surrounding genes has been subtracted out of their pairwise correlation. This leads to the area of graphical modeling. A limitation of the graphical modeling approach is that the correlation matrix of expression profiles between genes is degenerate whenever the number of genes to be analyzed exceeds the number of distinct expression measurements. This can cause considerable problems, as calculation of partial correlations typically uses the inverse of the correlation matrix. To avoid this limitation, we propose two practical multiple regression procedures with variable selection to measure the net, screened, relationship between pairs of genes. Possible biases arising from the analysis of a subset of genes from the genome are examined in the worked examples. It seems that both these approaches are more natural ways of analyzing gene expression data than the currently popular approach of two way clustering.

Keywords: expression profile, correspondence analysis, genetic link, microarray

1 Introduction

With the remarkable accumulation of genomic data, comparative bioinformatics has become a powerful alternative to bench science for gene function elucidation [8, 13]. Amongst the most powerful of recent methods is the analysis of gene expression profiles gathered with the aid of microarrays. This area, dealing as it does with 103 to 105 gene expression profiles in a single experiment, calls for a unity of experimental approaches in molecular biology and information sciences. To extract more useful information from microarray data, there is the urgent need to develop appropriate statistical procedures and to examine their properties.

Currently, clustering algorithms have been successfully applied to microarray data to classify genes and tissue types [7, 16]. On the other hand, there are few works on search for genetic links based on microarray data. In this paper, we first introduce correspondence analysis as a powerful candidate for gene search by associating specific genes with specific phenotypes. With the explicit cost function, the correspondence analysis relates genes in directly to tissues in which they over-express. As a numerical example, we use the expression profiles of colon-cancer tissues and normal tissues reported by Alon *et al.* [2].

Genetic links are not directly estimated by the correlated expression profiles between genes, because some genes in the third parties may have promoted or recessed a pair of genes concurrently. Partial

correlations measure the net relationships between pairs of genes. However, microarray examines thousands of genes, while the number of chips is largely limited by the budgetary constraint and manpower. This makes it impossible to calculate the full partial correlation matrix. Here, we propose two applications of regression analysis to obtain the net relations. The first regresses the expression profiles of a pair of genes to those of the other genes in the array, and the second regresses each of the pair to the other genes. Genes that have essential influence on the gene(s) of interest are extracted by variable selection. Both procedures are simple and practical for the analysis of a large number of genes. Numerical example shows that they approximate the net relationships between pairs of genes well.

As always, it is important to evaluate the information content of current microarray experiments, and hopefully identify ways of better matching data to model. A general consideration is that as methods of analyzing the data become more complex, there is always the danger of being over confident in erroneous associations, i.e. false positives. We consider how serious such problems are at present and what can be done to reduce them.

2 Methods

Correspondence analysis Correspondence analysis basically locates genes close to tissues in which they are over expressed. The display coordinates of genes $x_i^{(g)}$ ($i = 1, \dots, n_g$) and those of tissues $x_j^{(t)}$ ($j = 1, \dots, n_t$) are obtained by minimizing

$$L = \sum_{i=1}^{n_g} \sum_{j=1}^{n_t} f_{ij} (x_i^{(g)} - x_j^{(t)})^2 \quad (1)$$

under the constraints that the mean of all coordinates is zero with variance equals 1 [3, 11], and, where $f_{ij} \geq 0$ is the expression of the i th gene in the j th tissue. Since all terms must be nonnegative, the larger f_{ij} give the larger contribution. In contrast to two-way cluster analysis that classifies genes and tissues separately, the cost function in (1) relates genes to tissues in a more direct way.

Some general themes that occur with correspondence analysis are that genes expressed at moderate to high levels in most tissues, such as house keeping genes, are drawn toward the origin, while genes which have low average expression level except for the occasionally high expression, are located toward the periphery of the display. In this way, genes can be associated with tissues that are located close by.

Genetic link identification Genes and tissues are typically classified using correlations of gross expression level. For the inference of genetic links, however, negative relationships between genes are just as important as positive correlations. Furthermore, it is essential to subtract the contribution of surrounding genes before inferring any possible causal relationship between a pair of genes, which is what genetic links aim to present. This is because genes whose expressions are controlled by a common gene can have high apparent correlations, but are not directly linked in the pathway. The net relationship between a pair of genes may be measured by partial correlation. A complete set of partial correlations between genes may be obtained after inverting the correlation matrix, such that the partial correlation, (\tilde{r}_{ij}) between genes i and j is,

$$\tilde{r}_{ij} = -\frac{r^{ij}}{\sqrt{r^{ii}r^{jj}}}$$

where r^{ii} , r^{jj} and r^{ij} are elements of the inverse of the correlation matrix.

Graphical modeling is a large field and often implies assuming an explicit multivariate model. The usual steps are often calculation of partial correlation, selection of a set of partial correlations that indicate a non-cyclic graph, and inference of the properties of this graph and data, e.g. the likelihood

of the data, under an explicit model. Unfortunately, model selection can be problematic when the data do not fit the model. One way of sidestepping this issue and going more directly to an exploratory analysis is to visualize the “raw data” of partial correlations.

For exploratory analysis of partial correlations, we apply a particular type of multidimensional scaling, MDS, [4] to plot the genes from the matrix of absolute partial correlations, i.e. $|\tilde{r}_{ij}|, i = 1, \dots, n_g, j = 1, \dots, n_g$. The coordinates of genes $x_i^{(g)} (i = 1, \dots, n_g)$ are obtained by minimizing

$$L = \sum_{i=1}^{n_g} \sum_{j=1}^{n_g} n_g |\tilde{r}_{ij}| (x_i^{(g)} - x_j^{(g)})^2$$

under the constraint that the mean and the variance of the coordinates remain constant. This method is due to Hayashi (ref.) and basically places points in space paying most attention to their raw similarity, with greater similarities carrying greater weight. The more usual approaches to MDS would use a distance matrix (e.g. $1 - |\text{partial correlation}|$). In doing so, all distances have a similar impact, and the choice of a distance transformation (e.g. why not $1 - |\text{partial correlation}|^2$) is somewhat arbitrary.

Statistical procedures such as graphical modeling [12] and Bayesian networks [6, 14] can be used to select graphs that might best estimate the underlying genetic links. However, it is very important to first examine the information content of the whole data, particularly when the number of genes studied vastly exceeds the number of experiments conducted to relate them (as is likely to remain the case for some time). The correlation matrix is inevitably degenerate in this situation. Even if we extract a small subset of genes that have known functions related to the link of interest, it is difficult or impossible to include *a priori* all genes that strongly influence the expression of genes within the link. Unless the genes of the link are all identified, and unless all are independent from the remaining genes of the whole genome, the estimated model will be biased. Put another way, latent (hidden or excluded) variables, in this case genes, will distort the inferred genetic link.

Here, we propose two procedures, APCR 1 (Approximate Partial Correlation with Regression) and APCR 2, which approximate the partial correlation without the need to invert a correlation matrix. Both of these methods utilize regression analysis with variable selection. The essential idea is to sequentially subtract the effects of genes, which have significant correlations with a pair of genes, from the apparent correlation between of that pair. We adopt the Akaike Information Criterion, AIC [1] as the criterion for the variable selection, although other criteria such as BIC may also be used.

APCR 1: Regress the expression of the pair of genes $\mathbf{y} = (x_i, x_j)'$ to the expressions of all other genes:

$$\mathbf{y}_m = \mathbf{a} + \mathbf{b}_1 x_{h_1 m} + \dots + \mathbf{b}_k x_{h_k m} + \mathbf{e}_m$$

where x_{h_1}, \dots, x_{h_k} are the expressions of the all the other k genes, while the error term $\mathbf{e}_m, m = 1, \dots, n$ has mean $\mathbf{0}$ and the 2×2 variance-covariance matrix \mathbf{V} . From $\tilde{r}_{ij} = V_{ij}/(V_{ii}V_{jj})^{1/2}$, the net correlation is obtained. Note, this method requires order k^2 regressions due to regressing pairs of genes at a time. APCR 2: Use partial regression coefficients to estimate the partial correlation [17]. That is,

$$b_{ij}b_{ji} = \tilde{r}_{ij}^2$$

where b_{ij} and b_{ji} are the partial regression coefficients of the regression equations

$$\begin{aligned} x_{im} &= a_i + b_{ij}x_{jm} + c_1x_{h_1m} + \dots + c_kx_{h_km} + \epsilon_{im} \\ x_{jm} &= a_j + b_{ji}x_{im} + d_1x_{h_1m} + \dots + d_kx_{h_km} + \epsilon_{jm} \end{aligned}$$

To get a reliable estimate for the significant coefficients from insufficient number of experiments, we select variables using AIC. From the estimated regression coefficients, we obtain the approximate partial correlations as

$$\hat{r}_{ij} = \text{sign}(\hat{b}_{ij})\sqrt{\hat{b}_{ij}\hat{b}_{ji}} \quad (2)$$

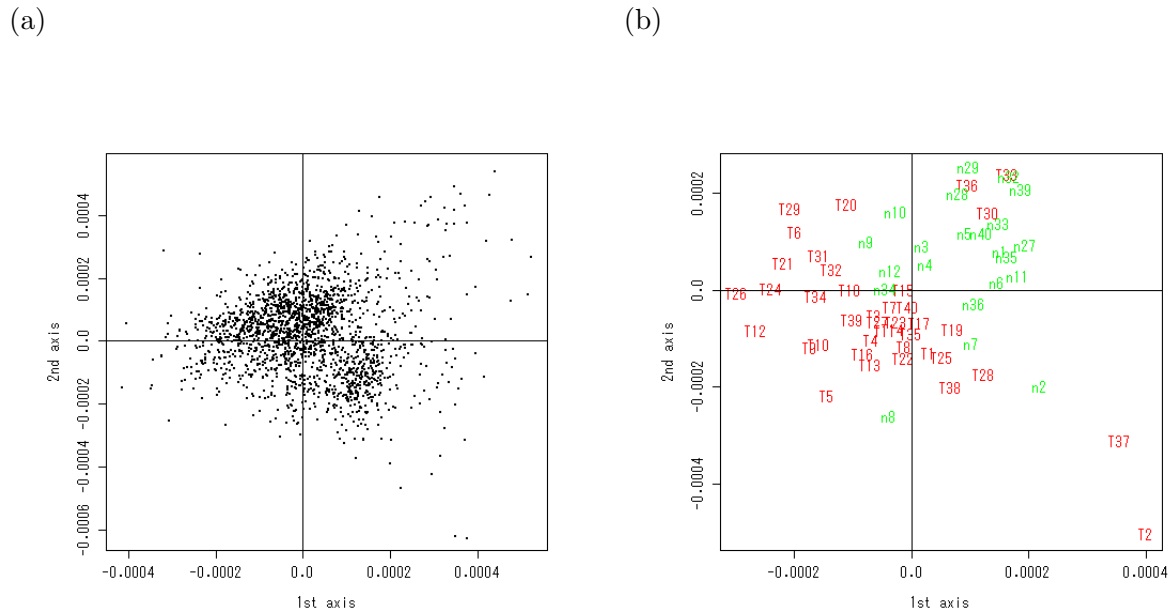


Figure 1: Correspondence analysis. (a) Scatter plot of genes, (b) Scatter plot of cells (“T##” represent tumor cells and “n##” represent normal cells)

here function $\text{sign}(x)$ takes value 1 if x is positive and -1 if it is negative. If each pair of genes in the experiment is influenced by a relatively a small number of genes, the above two estimates discount well the apparent correlation caused by other genes influencing both members of a pair. This approach is order k with respect to the number of regressions made.

3 Worked Example: Correspondence Analysis

3.1 Oriented Scatter Plots of Genes and Tissues

We now apply correspondence analysis to the data by Alon *et al.* (1999). These data are a series of 62 Affimetrix GeneChip experiments upon normal (designated N) and cancerous (T) colon tissue.

Figure 1 shows the scatter plots of genes (a) and cells (b) in the 1st and 2nd most significant dimensions (i.e. obtained from the eigenvalues of the data). The normal cells are mostly distributed in the upper-right region, whereas the tumor cells are distributed in the lower-left region, so the visual separation is moderately good, about as good as the clustering used in the original paper. Exceptions are T30, T33, T36 that appear in the upper-right region and n8 that appears in the lower-left region. This is not too surprising, as these samples had a high fraction of normal tissue mixed during the biopsy [2], and a similar pattern is observed even in cluster analyses that improve upon those in the original paper, as discussed in Waddell and Kishino (unpublished). Tumor tissues T2 and T37 are located in the peripheral area of the scatter plot, which implies that changes in the expression pattern of a small number of genes typifies them.

3.2 Correlated Genes and Genes with Strong Specificity

Four tissues T33, n39, T26 and T37, are examined in detail to illustrate how correspondence analysis associates tissues based on gene expression. Tissues T33 and n39 are located close together, although they are of different tissue types and from different individuals. Figure 2(a) shows the high correlation between these two tissues gene expression patterns. On the other hand, the distant pair of T33 and

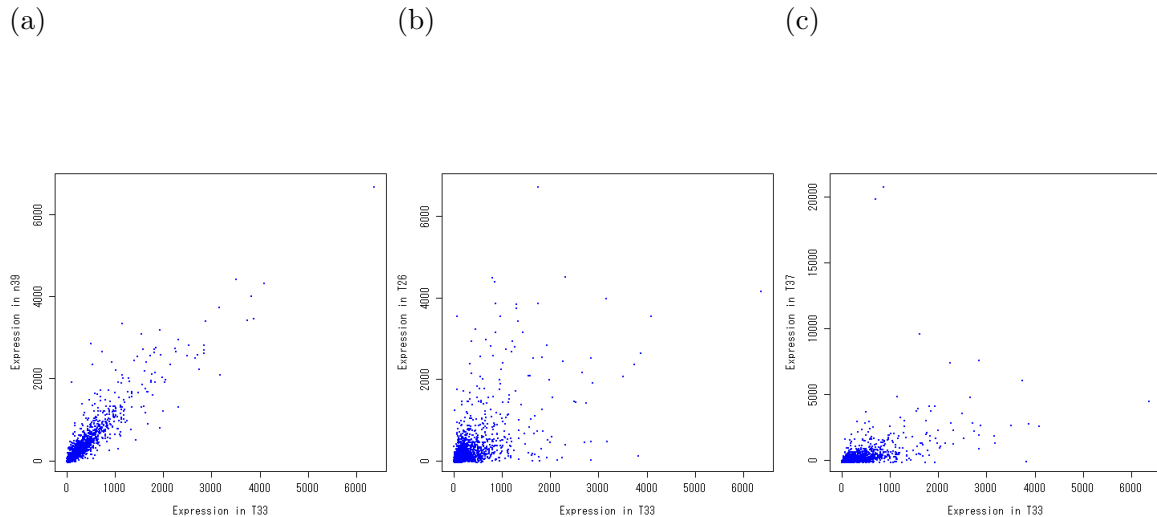


Figure 2: Correlations of T33 with n39, T26, and T37. (a) T33 vs. n39, (b) T33 vs. T26, (c) T33 vs. T37.

T26 shows a low correlation of gene expression, although they are of the same tissue type (Figure 2(b)). The low correlation between T33 and T37 (Figure 2(c)) is also consistent with the scatter plot of correspondence analysis (Figure 1(b)).

Reflecting the fact that T37 is located in the periphery region, a number of genes with typically low expression levels, are very high in this tissue and distinguish it and T2 from other tissues. Four genes, J00231, M27749, M87789, R62549, are located near T37 and T2 (Figure 3). They are all immunoglobulin related genes. As shown in Figure 4, these genes over-express in both T2 and T37 but have low expression in most of the other cells.

3.3 Distribution of Genes

Figure 5(a) shows the distribution of the z-scores (i.e. difference divided by s.d.) between the mean expression level of a gene in tumor cells and that in normal cells. Orange accession numbers are attached to the genes with z-scores larger than 3, and blue accession numbers are attached to the genes with z-scores smaller than -3 on the scatter plot of Figure 5(b). The result shows clearly that correspondence analysis is highlighting genes of considerable importance in separating tissue types, which is what is hoped.

The genes, which appear in the extreme lower left of Figure 5(b), generally exhibit both a high expression level and a marked increase in level from normal tissues. Included are genes, which code for ribosomal proteins, transcription factors, polymerases, etc., which may be associated with cells undergoing active growth. (For a list of these genes see the supplementary information (see Tables 1 and 2). Another is the SET protein gene, which is implicated in the modulation of chromatin structure, and has been reported over expressed in certain (but not all) types of tumor. Another was lysozyme, over expressed in certain tumors and associated with an inflammatory-type reaction. The S-100p gene is also in this list and it is known for high expression in certain tumors (e.g. glial), as is the monocyte-derived neutrophil-activating protein. An over expressed pancreatic stone protein is also found in pancreatic tumors.

In contrast, the genes showing high expression in normal tissues as opposed to tumors, and appearing in the upper left of Figure 5(b), include many microtubule, and muscle fiber related proteins (e.g. actin, desmin, caldesmon), consistent with the normal tissues containing much muscle. Other genes expressed at high levels, such as GLVR, a retrovirus receptor gene, do not have such obvious

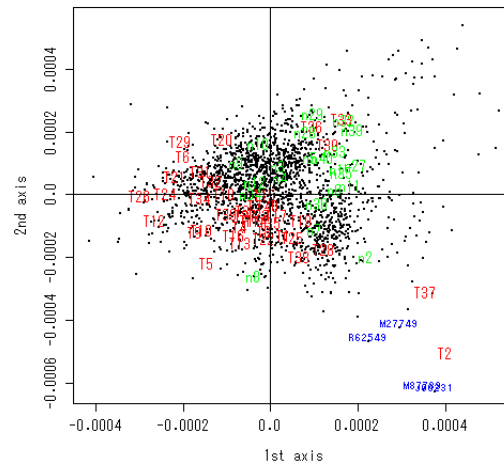
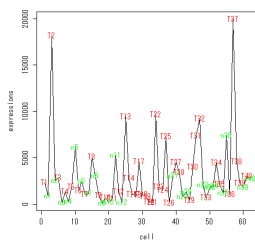
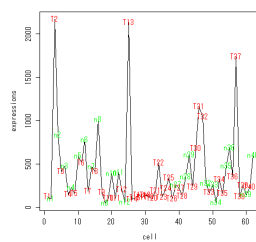


Figure 3: Overlaid scatter plot of genes and cells

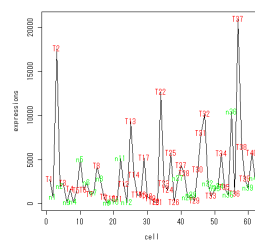
(a)



(b)



(c)



(d)

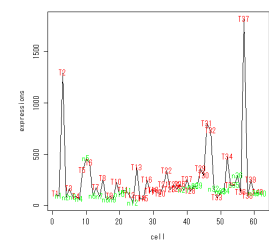


Figure 4: Expression profiles of the genes surrounding T37 and T2. (a) J00231, (b) M27749, (c) M87789, (d) R62549

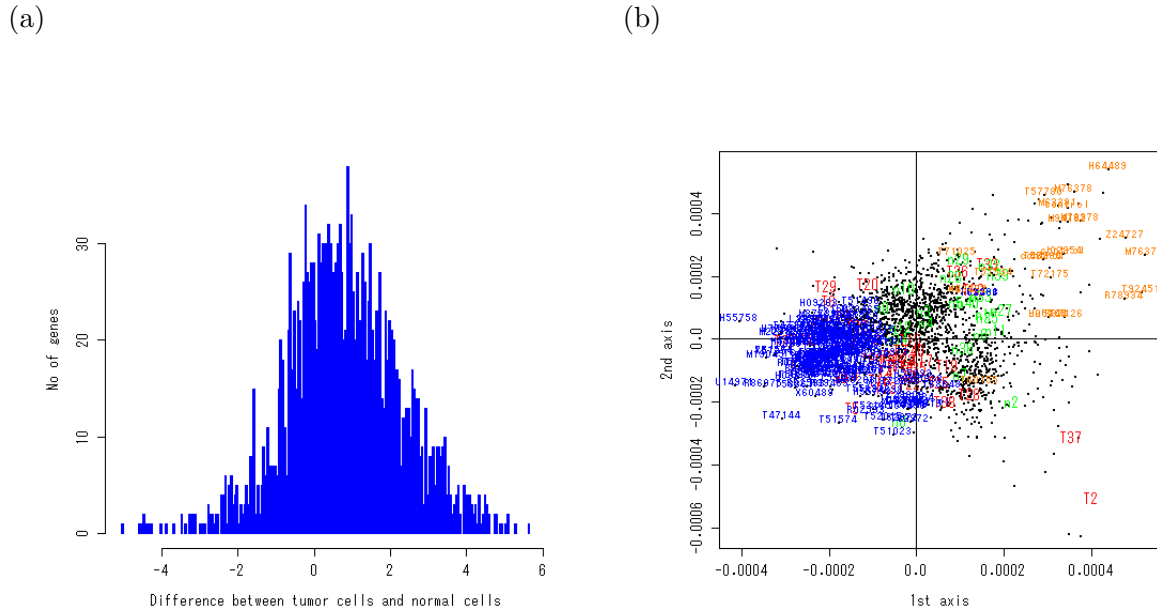


Figure 5: Distribution of genes with large z-scores between tumor cells and normal cells. (a) distribution of z-scores, (b) distribution of genes with z-scores larger than 3 (orange) or smaller than -3 (blue).

associations. Another high level protein was interferon gamma treatment inducible mRNA; whether this was induced by treatment, or stimulated by tumors in the same patient, is unclear. Overall then, the correspondence diagram seems to be giving a reasonable picture of highly expressed genes associated with particular tissue types.

4 Exploratory Analysis for Genetic Links

4.1 Genetic Links between Genes Correlated with Cancer

The net relationships between of a pair of genes can be measured by partial correlations, which can be obtained by the inverse of the correlation matrix. To illustrate ways of inferring the partial correlation when it is not possible to invert the correlation matrix, we chose 44 genes. These genes have correlations of either > 0.42 or < -0.42 with cancer tissues. To mimic the search for cancer genes and the analysis of genetic links, an imaginary gene expressing in a unit amount only in cancer tissues is added in with the real data. While these 45 genes do not compose a whole genome, the numerical methods for doing these calculations are the same as what would be used for a genome. The results would of course vary if all gene expressions could be regressed against the expressions of all others. (We assume that the partial correlations among the 45 genes are true, although they are probably inaccurate due to missing genes.). Accordingly, we do not look in detail at the graphical models, since the genes were not picked a priori for comprising a single pathway, and since a random sample of only 50 genes from the genome is likely to show many erroneous features in the graph. The purpose, rather, is to demonstrate general points.

Firstly, note that no correlation is observed between the partial correlations and correlations between pairs of genes. This implies that correlations can be used for clustering genes but we would expect these clusters may have little to do with particular genetic links assuming the partial correlations are accurate (which is not necessarily the case here due to latent variables etc.) Either correlations or partial correlations between genes can be used to link genes together graphically. The

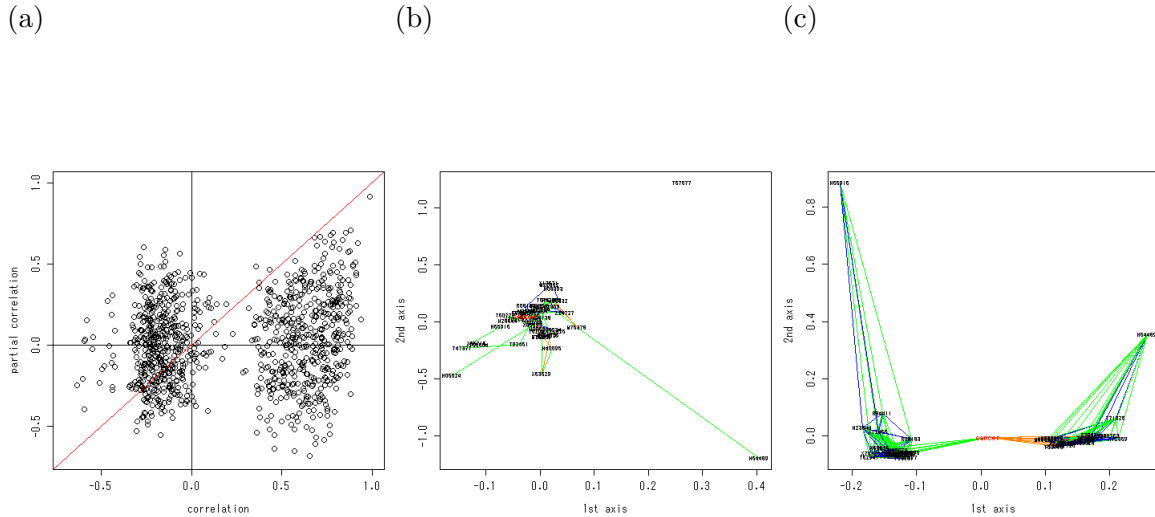


Figure 6: Graphical links between genes with large correlations with cancer. (a) A plot of partial correlations vs. correlations, (b) graphical (“genetic”) links estimated from partial correlations, (c) erroneous links estimated from correlations: “+” represents the cancer state. In both (b) and (c) pairs with partial correlation > 0.6 are connected by blue lines, pairs with partial correlation between 0.45 and 0.6 are connected by green lines and pairs with partial correlation < -0.45 are connected by orange lines.

genetic associations inferred from correlations (Figure 6(c)) are totally different from that based on the partial correlations (Figure 6(b)). Pairs with partial correlation > 0.6 are connected by blue lines, pairs with partial correlation between 0.45 and 0.6 are connected by green lines and pairs < -0.45 are connected by orange lines. Figure 6(b) has only a few pairs with large net relationship, whereas Figure 6(c) has many pairs of apparent relationship through the “third parties” of other genes.

4.2 Illustrating Bias When Modeling a Subset of Genes

To examine the possible bias when making estimates from a subset of genes, we randomly choose 21 genes including the imaginary “cancer” gene from the previous 45 genes. Figure 7(a) compares the partial correlations based on the correlations between the 21 genes with those based on the correlations between the full set of 45 genes. No correlation is observed between the two sets of partial correlations ($r = 0.137$). Additionally, the picture of the links from the 21 (Figure 7(c)) is very different from one from that on the 45 (Figure 7(b)). This is because pairs of genes in the subset of 21 are correlated with genes outside of the subset to differing extents. While our set of 44 genes may be more correlated than most, so the effect larger than expected, we do expect many examples where only 1/2 or less of the genes in a link are known with any confidence. Unless we are sure from supplementary information that there are no essential genes missing from the experimental set, genetic link analysis requires considerable caution.

4.3 Approximating \tilde{r} with Regression and Variable Selection

First, we estimate the net relationship between a pair of genes using APCR 2 and equation (2). Figure 8(a) compares the inferred net relationships with the partial correlations calculated directly from the correlation matrix of the full set of 45 genes. Of the 990 pairs in total, significant net relations were not detected in 700 pairs (70.7%). However, these partial correlations, which ranged between -0.473 and 0.443 , are relatively small. Excluding these pairs, the correlation coefficient between the estimated net

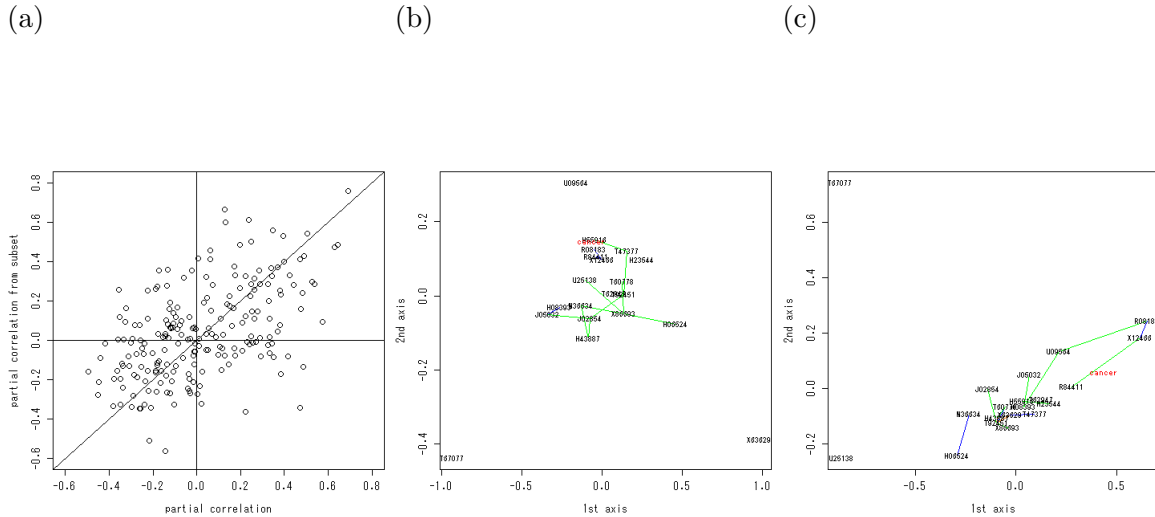


Figure 7: Genetic links in the subset of genes. (a) partial correlations among the subset of the genes compared with those from larger set, (b) genetic links based on the sub matrix of the partial correlations, (c) genetic links based on the subset of genes. Pairs with partial correlation > 0.6 are connected by blue lines, pairs with partial correlation between 0.45 and 0.6 are connected by green lines and pairs with partial correlation < -0.45 are connected by orange lines. Next, we illustrate how APCR 1 and APCR 2 can be used to identify genes that are essential to include in the estimation of a single link, even when the partial correlations are not available directly.

relations and the partial correlations is very high at 0.967. Figure 8(b) shows a picture of the genetic links estimated using the net relations from equation 2, in place of partial correlations. We also need to be careful about assigning too much confidence to the directly estimated partial correlations. The largest eigenvalue of the correlation matrix is 18.904 and the smallest is 0.001 (Figure 9). Even with just 45 genes, the condition number of the matrix is 21191.2. This means that the inverse of the matrix is strongly sensitive to random noise in the data. Since measurement errors with coefficient of variation between 20 and 40% are unavoidable in microarray data, a conservative estimate of net relations using APCR 1 or APCR 2 i.e. a regression procedure with variable selection, has an advantage with respect to not only computational ability but also robustness against measurement errors.

5 Discussion

Exploratory analysis with graphical representation is an important step in data mining, especially when analyzing large datasets that are not fully understood. Since microarray records the expression profiles of thousands of genes, the methods presented in this paper offer useful tools in searching for novel genes with especially important functions and study of genetic links.

The inference of genetic links in section 3 clearly illustrates some present limitations of microarray data. The exact partial correlations can be calculated, only when the number of genes to be compared is less than the dimension of the profiles. However, even if we restrict our analysis to a limited number of genes of major concern, there is no guarantee that they are uncorrelated with the other genes in the microarray data. If genes not sampled, selected, or just lost in the background noise of the experiment, happen to correlate strongly with genes of concern, the estimated picture of the link may often be seriously distorted. Regression analysis proposed in the paper approximates the net relations between pairs of genes well, and helps to identify critical genes to include if they are included on the array.

Although the depth, or number of distinct experiments, making up the expression profiles are

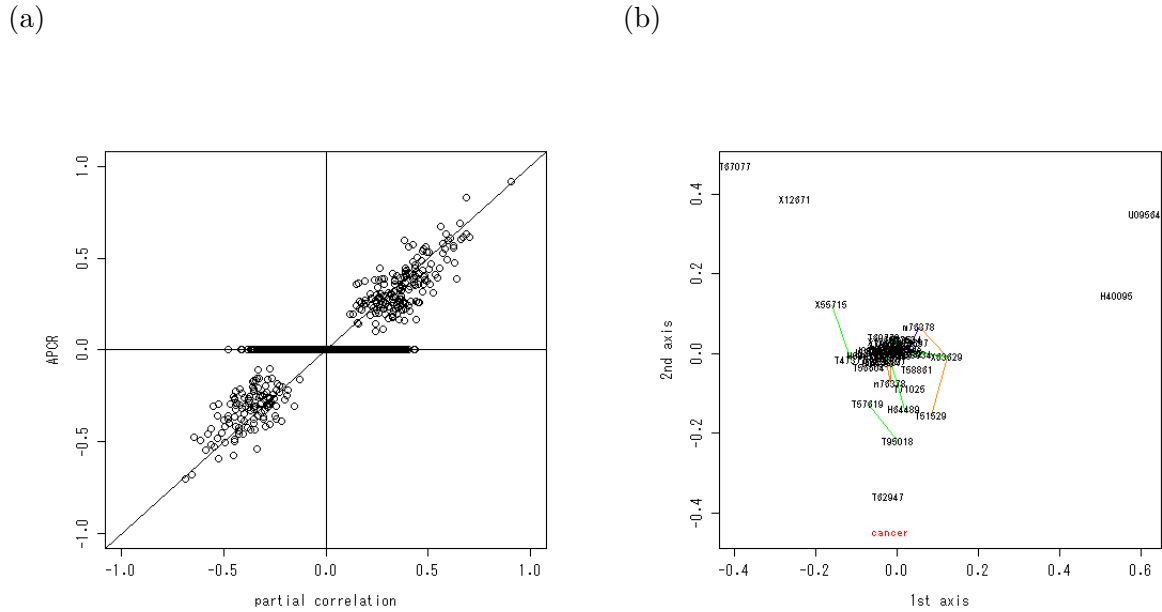


Figure 8: Approximate partial correlations, or net relations, inferred by regression analysis plus variable selection and the resulting genetic links. (a) net relations obtained by regression analysis with variable selection compared with the partial correlation, (b) genetic links estimated using the approximate partial correlations. Pairs with an approximate partial correlation of; > 0.6 are connected by blue lines; between 0.45 and 0.6 are connected by green lines; and < -0.45 are connected by orange lines.

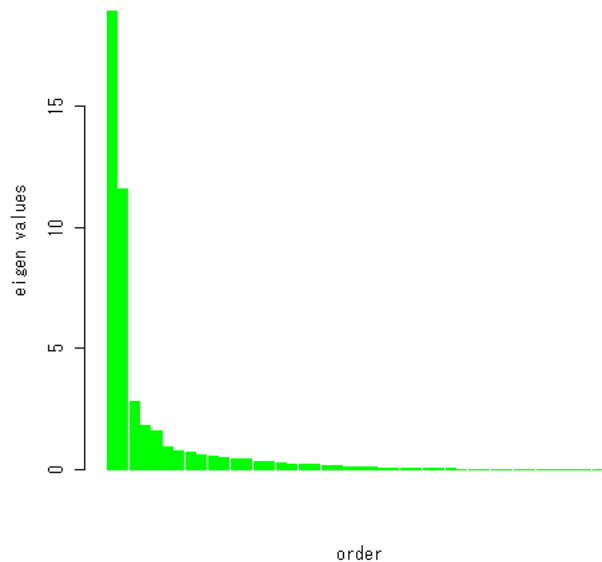


Figure 9: Eigenvalues of the correlation matrix of the 45 genes. The rapid and continuous decrease in size is not a good sign, and indicates the matrix inverse is likely to be poorly conditioned, hence making estimation of partial correlations hazardous.

Table 1: Genes overexpressed in tumors.

Sequence accession number		IMAGE clone number	gene description
M19045	gene		Human lysozyme mRNA, complete cds
T56940	3' UTR	68306	P24050 40S RIBOSOMAL PROTEIN
M22382	gene		MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN)
R36977	3' UTR	26045	P03001 TRANSCRIPTION FACTOR IIIA
T47377	3' UTR	71035	S-100P PROTEIN (HUMAN)
T84049	3' UTR	114175	SET PROTEIN (Homo sapiens)
M26383	gene		Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds
X54942	gene		H.sapiens ckshs2 mRNA for Cks1 protein homologue
U21090	gene		Human DNA polymerase delta small subunit mRNA, complete cds
M27190	gene		Homo sapiens secretory pancreatic stone protein (PSP-S) mRNA, complete cds
R67999	3' UTR	138262	PROBABLE ATP-DEPENDENT RNA
R62945	3' UTR	139080	HELICASE PRH1 (Schizosaccharomyces pombe)
D00760	gene		COMPLEMENT DECAY-ACCELERATING FACTOR 1 PRECURSOR (Homo sapiens) PROTEASOME COMPONENT C3 (HUMAN)

Table 2: Genes more highly expressed in normal.

Sequence accession number		IMAGE clone number	gene description
R78934	3' UTR	146232	ENDOTHELIAL ACTIN-BINDING PROTEIN (Homo sapiens)
Z24727	gene		H.sapiens tropomyosin isoform mRNA, complete CDS
M63391	gene		Human desmin gene, complete cds
T60155	3' UTR	81422	ACTIN, AORTIC SMOOTH MUSCLE (HUMAN)
R87126	3' UTR	197371	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
M76378	gene		Human cysteine-rich protein (CRP) gene, exons 5 and 6
X15882	gene		Human mRNA for collagen VI alpha-2 C-terminal globular domain
M83667	gene		Human NF-IL6-beta protein mRNA, complete cds
T92451	3' UTR	118219	TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE (HUMAN)
H43887	3' UTR	183264	COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)
X15880	gene		Human mRNA for collagen VI alpha-1 C-terminal globular domain
X74295	gene		H.sapiens mRNA for alpha 7B integrin
M92843	gene		TRISTETRAPROLINE (HUMAN)
M26683	gene		Human interferon gamma treatment inducible mRNA
L05144	gene		PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC (HUMAN); contains Alu repetitive element; contains element PTR5 repetitive element
L20859	gene		Human leukemia virus receptor 1 (GLVR1) mRNA, complete cds
R48303	3' UTR	153505	TYROSINE RICH ACIDIC MATRIX PROTEIN (Bos taurus)
J02854	gene		MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN); contains element TAR1 repetitive element
X86693	gene		H.sapiens mRNA for hevin like protein
T61333	3' UTR	78034	METALLOPROTEINASE INHIBITOR 3 PRECURSOR (Gallus gallus)
R60877	3' UTR	42396	DELTA-CRYSTALLIN ENHANCER BINDING FACTOR (Gallus gallus)
T67077	3' UTR	66563	SODIUM/POTASSIUM-TRANSPORTING ATPASE GAMMA CHAIN (Ovis aries)
H06524	3' UTR	44386	GELSOLIN PRECURSOR, PLASMA (HUMAN)
U25138	gene		Human MaxiK potassium channel beta subunit mRNA, complete cds
T60778	3' UTR	76539	MATRIX GLA-PROTEIN PRECURSOR (Rattus norvegicus)
M64110	gene		Human caldesmon mRNA, complete cds

largely limited by budgetary constraints and manpower, a much greater depth may be achieved with the aid of databasing the results of many different groups. Existence of a common standard with agreed controls common among institutions and evaluation of measurement errors are very important for successful meta-analysis [5, 15], because without them, the compounded errors would swamp the gains.

Comparison with other techniques

There are potentially many ways to use massive gene expression studies to classify tissues (e.g. [10]). An advantage of correspondence analysis is to simultaneously group tissues and genes in relation to one another. At present the method gives most weight to overexpressed genes, and these are best seen associating with the tissues that over express them. However, by transforming the original data, it is possible to see strongly under underexpressed genes associating with particular tissues also.

Graphical modeling is a general term that basically embodies the premise of measuring the partial correlation between all pairs and adding a link if there is significant evidence that the partial correlation is non-zero. There are many ways of doing this. Wright's path analysis is an early example, while work in the likelihood framework has lead to the development of what are hopefully more statistically efficient modeling techniques (e.g. [12]). In this framework, we have likelihood ratio fit statistics and various forwards and backwards measures of fit to allow edges to enter or exit the model ([12], for an application to microarray see [18]).

Apart from fully parameterized models, there is also the freedom to use the information in partial correlations in more "visual" ways, and as a type of distance. For example, figure 6 is such an example where the partial correlations are used to create a 2 dimensional representation of the distances between genes, and secondarily, large partial correlations linking pairs of genes are highlighted. Another approach is taken in [18] where partial correlations are used as the basis for cluster or tree building analysis. These methods are also of utility in defining compact groups of genes that may form especially "tight" pathways that can be further dissected with more exact methods.

Bayesian networks (e.g. [9]) are a type of graphical model that tend to use mutual information measures rather than correlations. They can also place some prior probability on edges or links, which can be a very attractive feature in helping to guide the construction of a model. Such information can include the results of more traditional and focused experiments. Other methods of graphical modeling can also use some prior information, but tend to need to force edges to be either in or out of the selected model.

At present, it is hard to compare the pros and cons of these various alternatives in a meaningful way. Perhaps a good analogy can be made with phylogenetic evolutionary tree inference (which is itself a special form of graphical modeling). While Bayesian and ML methods can claim asymptotic efficiency and a fair measure of robustness, all methods do fail. In such cases, comparison between techniques often helps to illuminate the probable cause. In other contexts, non-ML variants are much faster than ML and allow a better search of the model space with large amounts of data (taxa). We would anticipate each graphical modeling technique to have sufficient unique attributes to make no single method automatically better than others for the wide variety of genetic pathway problems to be resolved.

Acknowledgement

The authors appreciate Dr. Miyuki Shimane for her valuable comments on the manuscript. 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.

References

- [1] Akaike, H., A new look at the statistical model identification, *IEEE Trans. Autom. Contr.*, AC-19:716–723, 1974.
- [2] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, 96:6745–6750, 1999.
- [3] Benzecri, J.P., *Correspondence Analysis Handbook*, Dekker, New York, 1992.
- [4] Borg, I. and Groenen, P., *Modern Multidimensional Scaling: Theory and Applications*, Springer-Verlag, 1997.
- [5] Carrol, R.J., Ruppert, D., and Stefanski, L.A., *Measurement Error in Nonlinear Models*, Chapman & Hall, London, 1995.
- [6] Castillo, E., Gutierrez, J.M., and Hadi, A.S., Modeling probabilistic networks of discrete and continuous variables, *J. Multivariate Analysis*, 64:48–65, 1998.
- [7] 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, 1999.
- [8] Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O., Protein function in the post-genomic era, *Nature*, 405: 823-826, 2000.
- [9] Friedman, N., Linial, M., Nachman, I., and Pe'er, D., Using Bayesian networks to analyse expression data, *Journal of Computational Biology*, in press.
- [10] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander E.S., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286:531–537, 1999.
- [11] Hayashi, C., On the prediction of phenomena from mathematical statistic point of view, *Ann. Inst. Stat. Math.*, 3:69–98, 1950.
- [12] Lauritzen, S.L., *Graphical Models*, Oxford Statistical Science Series, 17, 1996.
- [13] Lockhart, D. and Winzeler, E.A., Genomics, gene expression and DNA arrays, *Nature*, 405:827–836, 2000.
- [14] Nilsson, D., An efficient algorithm for finding the M most probable configurations in probabilistic expert systems, *Statistics and Computing*, 8:159–173, 1998.
- [15] Olkin, I. and Sampson, A., Comparison of Meta-analysis Versus Analysis of Variance of Individual Patient Data, *Biometrics*, 54:317–322, 1998.
- [16] 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, *Nat. Genet.*, 24:227–235, 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] Waddell, P.J. and Kishino, H., *Cluster Inference Methods and Graphical Models evaluated on NCI60 Microarray Gene Expression Data*, Genome Informatics Series, Vol. 11.