Modeling Microarray Data: Interpreting and communicating the biological results.

Y.E. Pittelkow and S.R. Wilson

Centre for BioInformation Science, Mathematical Sciences Institute, Australian National University, Building 27, Canberra, ACT, Australia, 0200. `http://dayhoff.anu.edu.au`

Summary

Various statistical models have been proposed for detecting differential gene expression in data from microarray experiments. Given such detection, we are usually interested in describing the differential expression patterns. Due to the large number of genes that are typically analysed in microarray experiments, possibly more than ten thousand, the tasks of interpretation and communication of all the corresponding statistical models pose a considerable challenge, except perhaps in the simplest experiment involving only two groups. A further challenge is to find methods to summarize the resulting models. These challenges increase with experimental complexity.

Biologists often wish to sort genes into 'classes' with similar response profiles/patterns. So, in this paper we describe a likelihood approach for assigning genes to these different class patterns for data from a replicated experimental design.

The number of potential patterns increases very quickly as the number of combinations in the experimental design increases. In a two group experimental design there are only three patterns required to describe the mean response: up, down and no difference. For a factorial design with three treatments there are 13 different patterns, and with four levels there are 75 potential patterns to be considered, and so on.

The approach is applied to the identification of differential response patterns in gene expression from a microarray experiment using RNA extracted from the leaves of *Arabidopsis thaliana* plants.

We compare patterns of response found using additive and multiplicative models. A multiplicative model is more commonly used in the statistical analysis of microarray data because of the variance stabilizing properties of the logarithmic function. Then the error structure of the model is taken to be log-Normal. On the other hand, for the additive model the gene expression value is modeled directly as being from a gamma distribution which successfully accounts for the constant coefficient of variation often observed.

Appropriate visualization displays for microarray data are important as a way of communicating the patterns of response amongst the genes. Here we use graphical 'icons' to represent the patterns of up/down and no response and two alternative displays, the *Gene*-plot and a grid layout to provide rapid overall summaries of the gene expression patterns.

1 Modeling Gene Expression Data

1.1 Background

A microarray experiment differs from more classical experiments described in the biometrics literature by the very large number of parallel measurements (genes' expression values) on

typically only a few cases (slides or chips). The small number of cases is, in part, a function of the high costs of the experiment and the limitation of available sample.

The problem of assigning genes into 'classes' with similar response profiles/patterns is an important one, which is usually tackled in the microarray literature with methods such as clustering, Self-Organizing Maps etc or by using fold change or $t-$ like statistics to compare all the different pairs. These methods do not all take into account the information about the experimental design, nor do they incorporate statistical modeling. Rather than considering the problem as one of finding which genes have similar patterns, we can ask the related question: given the experimental design, what patterns are possible, and which genes belong to which patterns. Thus, given the experimental design, the patterns are treated as 'known' and we have a type of 'supervised clustering'.

In this paper we describe a likelihood approach for assigning genes to different class patterns for data from a replicated experimental design.

Various statistical models have been proposed for detecting differential gene expression in microarray experiments, but usually they have been used for two-group comparisons. Often models are estimated on a gene by gene basis although recently there have been attempts to 'share' information (see for example Lönnstedt and Speed, 2002 [8]). Whatever model is used, when there are more than two groups, biologists are often interested in first the detection of differential expression followed by description/s of the pattern/s of differential expression. The large number of genes which are typically modeled - possibly tens of thousands - makes this aspect of modeling a considerable challenge in microarray data analysis.

When there are only two groups in a microarray experiment, such as two genotypes, two treatment groups, samples from normal and diseased tissues and so on, interpretation of the model is usually relatively straightforward. Then interest usually focuses on allocating genes to classes of no response, up or down regulation, say. Such allocation is the most common case covered in the statistical literature on microarray data analysis.

For more complex models, such as to analyze experimental data involving more than two groups, multiple factors, interactions or continuous covariates, even at the single gene level, interpretation can be quite challenging. The biological problem of describing the patterns of response can be seen as one of model interpretation, model simplification or model selection. In the classical Analysis of Variance (ANOVA) model the problem of detecting which means are different and which ones are probably not different following a rejection of a hypothesis of equal treatment effects is known as the problem of 'multiple comparisons'. As discussed later, it is useful to distinguish this problem from the problem of 'multiple hypothesis testing' for the thousands of genes in microarray experiments.

Although the 'multiple comparison' problem is not new and it is unlikely given the potential range of models that a 'one approach fits all' is possible, the importance of using well-established statistical methods rather than *ad hoc* computational methods to address the problem cannot be understated.

In this paper we compare the patterns of response found using additive and multiplicative models. We focus on models with independent and identically distributed (*iid*) errors and qualitative factors. Moving to consideration of three treatment levels (groups) in a single factorial model (one way ANOVA) there are twelve potential 'patterns' of interest, given by not only the relative up/down patterns, but also accommodating no difference between pairs of levels. When

there are more than three levels the number of combinations of up/down/no difference increases rapidly. For example, with four levels, as for the data described in Section 2, there are 75 possible combinations.

A multiplicative model is more commonly used in the statistical analysis of microarray data because of the variance stabilizing properties of the logarithmic transformation. Then the error structure of the model is taken to be log-Normal. In the additive model, the gene expression value is modeled as being from a gamma distribution. So the data is not transformed, and the constant coefficient of variation that is often observed for fluorescent measurements obtained from hybridization (microarray) experiments is accommodated.

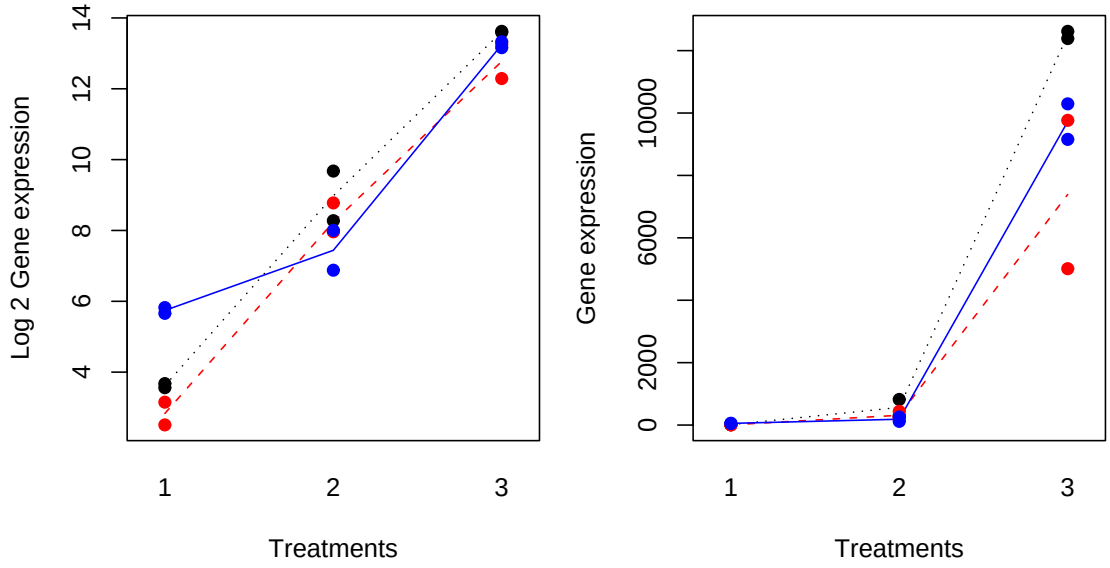

Figure 1: Plots of a gene's mean expression values from a stimulation experiment with three treatments (horizontal axis) and three genotypes (different lines). The expression values on the right are on the original scale, while those on the left are logarithm (base 2) transformed.

Both types of models are found in the literature but can give rise to very different 'patterns' or interpretations; see Figure 1. The data in Figure 1 (not published) are for a gene from a stimulation experiment wherein there are three genotypes, shown by different lines in the plot, and three stimulation treatments. The interpretation of the pattern on the left of the plot of the logarithm (base 2) of the gene expression values is quite different to that on the right that plots the original (untransformed) values. In the case of multiplicative models, the patterns of gene expression values are being modeled in terms of fold change while in an additive model, the observed or a scaled version of the measurement of gene expression is being modeled. Note that we assume appropriate pre-processing steps, such as normalization, have been already been undertaken.

1.2 Modeling

In this section we describe the likelihood approach to modeling (see for example Aitkin [3]). We consider both a multiplicative model (log-Normal) and an additive model (gamma).

The model for a single gene's expression values, Y, is written as

$$\mu = E[f(Y)|\mathbf{X}] = \beta'\mathbf{X} + \epsilon.$$

For the multiplicative model the function $f = log_2$ and the errors, ϵ, are assumed *iid* and to follow a Normal distribution. For the additive model the function f is the identity function, and the distribution of $Y|X$ is assumed to be a gamma with constant shape index.

The design matrix, $\mathbf{X}$, in a typical one-way ANOVA experiment would be a set of dummy variables indexing the different treatment conditions (see for example McCullagh and Nelder [9]).

The null model (of no difference between the groups) provides a natural basis for evaluation of differential gene expression. But following detection of differential expression the question of describing and interpreting the patterns of response remains. Tukey [15] p100, wrote:

> What we should be answering first is 'Can we tell the direction in which the effects of A differ from the effects of B?' In other words, can we be confident about the direction from A to B? Is it 'up', 'down' or 'uncertain'?

and in this spirit we approach the multiple comparison problem for gene expression data. We call each different combination a pattern or response profile. There are several methods available for evaluating which pattern is the best description, and we describe two popular ways. One approach might be to use the method of contrasts or 'Wald' tests which use the asymptotic result that for large Fisher information, $i(\beta)$, the distribution of β is often adequately approximated by $(\hat{\beta} - \beta) \sim N(0, i(\beta)^{-1})$ (McCullagh and Nelder [9], 471). The approximation is much closer for Normal, independently distributed errors than for gamma distributed variables, and also for large sample sizes. Small samples however are common in microarray data and the probability of model mis-specification is high. This latter fact is a consequence of the technology, the underlying biological processes, the preprocessing and the large numbers of genes involved. For some data, use of pairwise comparisons can lead to conflicting results. A simple example is where application of pairwise tests, in a 3 group design, lead to the results that: $\mu_1 = \mu_2, \mu_2 = \mu_3$ but $\mu_1 \neq \mu_3$.

An alternative approach, which we use, is based on the result that a formal comparison of two nested models can be expressed in terms of a likelihood ratio test. This approach is usually more accurate even for small sample sizes than the Normal approximation. We compare the likelihoods between a (sub) model representing a particular pattern and a (full) model which assumes all means are different. Such an approach considers the modeling task as one of determining which, of all the models possible, is the most parsimonious model that best describes the data. For both the log-Normal and the gamma, a dispersion parameter needs to be estimated so an *F*-test can be applied. If the relevant *p*-value is not judged to be significant at some pre-defined level, then the sub-model is assumed adequate and the gene is assigned to the corresponding pattern, with the following provisos. If there is more than one adequate sub-model, then the one with the fewest degrees of freedom is used to determine the final pattern assignment. Occasionally for gene expression data more than one sub-model with the same degrees of freedom may adequately describe the pattern of response. Since sample sizes are (usually) small for microarray experiments we have chosen to use Akaike's Information Criterion (AIC)

to select the most parsimonious model, and the gene is assigned to the corresponding pattern class.

A general algorithm can be developed using either contrasts or a log likelihood approach (or any other relevant approach) by noting that the pattern between the means of the gene's expression values in the different groups (treatments) can be described by their ranks. For example, a pattern 1234 refers to a pattern with monotonically increasing means in the four groups. If, say, the first and second means only are determined to be equal within sampling variability, then their ranks would be set equal. For this example the new ranks would be 1, 1, 2 and 3, respectively, and 1123 would describe the pattern. The pattern can be used to create a graphical icon, which describes the response pattern in terms of up/down and no difference. The pattern can also be used for generating gene lists.

2 Application to an *Arabidopsis thaliana* light adaptation experiment.

In this study a number of leaves at the same developmental stage were harvested from different *Arabidopsis thaliana* plants. Plants were grown for 7 weeks at low light (50 mmol quanta $m^{-2}s^{-1}$) with a ten hour photoperiod and day/night temperature of 21° in 70% humidity. Plants were transferred into high light (550 mmol quanta $m^{-2}s^{-1}$ with the same photoperiod). RNA was extracted after 0 (shade), half, two and six hours of high light and hybridized to Affymetrix GeneChips using standard Affymetrix protocols. Some of the responses to high light are expected to be transient and choosing the appropriate time points for the experiment is crucial. Initial RT-PCR results on some genes have since supported the design. The gene expression measure used in this study was the Affymetrix MAS 5.0 signal and the chips were normalized to the same mean expression on each chip (150). To allow the log-Normal model to be fitted, zero estimates of gene expression were set to 0.1 in all analyses.

The aim of this study (Rosche [14]) was to identify early changes in gene expression after the transfer of shade-adapted plants into different high light conditions (treatment). In this paper, however we are interested in identifying and describing *all* the different responses to treatment. Not all changes will be associated with light adaptation. For example, some of the changes in gene expression might be due to various stresses incurred during the transfer, such as mechanical stimulation, a change in temperature or the generation of reactive oxygen species due to an excess of irradiation on shade-adapted leaves.

Neither the same leaves, nor the same plants can be used at the different time points because of the known effect of harvesting on gene expression. Further, the number of available leaves is restricted. Hence the experimental design consisted of three biological replicate chips in the shade condition and two in each of the three light treatment groups.

Only a proportion of the 22,800 genes and ESTs (expressed sequence tags) represented on the chip would be expressed. There were14,669 genes in this data set that were detected as being present by the MAS 5.0 algorithm in all replicates in the shade or any one of the light treatments. These genes were used in subsequent analyses.

Patterns of 'no change', 1111, can be very common, so genes having this response pattern are most efficiently treated separately to the remaining genes. As mentioned in section 1.2, the null

model is the natural model to use for assigning genes to the 'no change' pattern class (i.e. no difference between groups).

Table 1 shows a comparison of the number of genes which would be declared 'significant', and thus other than pattern '1111', given the use of either the log-Normal or the gamma models, for different choices of *p*-values.

Log-Normal Model	Gamma Model					
p-value range	(0, 0.001]	(0.001, 0.005]	(0.005, 0.01]	(0.01, 0.05]	(0.05, 1]	Total
(0, 0.001]	258	22	0	0	0	280
(0.001, 0.005]	22	548	56	1	0	627
(0.005, 0.01]	0	46	423	86	0	555
(0.01, 0.05]	0	1	79	2074	116	2270
(0.05, 1]	0	0	0	139	10798	10937
Total	280	617	558	2300	10914	14669

Table 1: Comparison of the different numbers of genes which would be declared 'significant' depending on the *p*-value and the model chosen.

Thus, if a multiplicative ANOVA model, with an overall *p*-value of 0.005 was used, $14669 - 907 = 13762$ genes would be assigned a pattern of 1111, with ten more if an additive model had been used. $14669 - 850 = 13819$ genes would be assigned a pattern of 1111 by either model.

We selected an overall *p*-value of 0.005 for assigning genes to the 1111 pattern and a *p*-value of 0.01 for choosing the most parsimonious model, given detection of differential expression. Applying the algorithm, assuming a multiplicative model, 40 patterns of response (from the potential 75) were distinguished for these data (including 1111).

In Figure 2, a *Gene*-plot [12] is shown where each gene is represented by a colored icon, showing the pattern identified using the multiplicative model. The 39 different colors from the spectrum represent the 39 different pattern classes identified by the multiplicative model. Genes with the 'no change' pattern 1111 are not included. It is not necessarily easy to distinguish the colors although one can see that icons of similar shape and colour are located in pie shaped segments centred at the origin. Icons which are most distant from the origin have greater variance often indicating larger fold changes in expression. The *Gene*-plot has been developed as an exploratory method to display patterns of gene expression from a microarray experiment. This plot is similar in essence to the *h*-plot of Corsten and Gabriel [6]. The actual coordinates are at the left hand point for the icon. On this type of plot, genes which lie close together have similar correlations for their expression values. Genes lying on a straight line passing through the origin will have correlated values, with variance decreasing towards the origin. Genes which lie opposite each other have negatively correlated values, indicating a reverse pattern of up and down regulation across the chips. We note that if the plot is done on a different scale (such as on the untransformed gene expression data) its initial visual impact can appear to be rather different, although the interpretations are as just described. Further details on the *Gene*-plot can be found in Pittelkow and Wilson [12].

In Figure 2, the clustering of genes of similar response patterns is clearly visible. Some of the highly responsive genes at the top of the plot are 'shock' genes *i.e.* genes which respond following handling and harvesting (such as touch and stress genes).

Multiplicative	Additive	Model												
1111	1111	*13,715*												
1111	1112	*27*	1121	*4*	1122	*3*	1211	*7*	2111	*3*	2112	*1*	2211	*2*
1112	1112	***238***	1122	*2*	1123	*1*								
1121	1121	***13***												
1122	1111	*3*	1122	***101***	1222	*1*								
1123	1123	***10***												
1211	1211	***28***	1221	*1*										
1212	1111	*1*	1212	***3***										
1221	1111	*3*	1221	***39***	1222	*1*								
1222	1111	*3*	1112	*1*	1222	***38***								
1223	1112	*1*	1223	***3***										
1231	1231	***1***												
1232	1121	*2*	1232	***3***										
1233	1222	*1*												
1322	1322	***1***												
2111	1111	*2*	2111	***55***	2211	*3*	3211	*1*						
2112	1111	*2*	1112	*1*	2112	***38***	2212	*1*						
2113	2113	***7***												
2121	2121	***4***												
2122	1111	*1*	1112	*1*	2122	***12***								
2123	2123	***3***												
2132	2132	***1***												
2133	2133	***1***												
2134	2134	***1***												
2211	1111	*7*	2211	***87***	2221	*3*								
2212	1111	*3*	2212	***22***										
2213	2213	***3***												
2221	1111	*32*	1121	*1*	1211	*2*	2221	***79***						
2231	2231	***1***												
2311	2311	***1***												
2312	2312	***1***												
2321	1211	*1*	2321	***5***										
2331	2331	***2***												
3121	3121	***1***												
3123	3123	***1***												
3211	3211	***1***												
3212	3212	***3***												
3221	3221	***4***												
3312	3312	***1***												
3321	3321	***12***												
4213	4213	***1***												

Table 2: Pattern classes found using a Multiplicative (log-Normal) Model and an Additive (gamma) Model. A *p*-value cut-off of 0.005 for assigning genes to the 1111 pattern and a *p*-value cut-off of 0.01 for choosing amongst the patterns were used. The left hand column shows patterns distinguished by a Multiplicative Model, followed by the range of different patterns resulting from the use of an Additive Model. The number of genes allocated to the different pattern classes is shown in italics following the pattern label. The bolded values correspond to the identical patterns for both Models (ignoring pattern 1111).

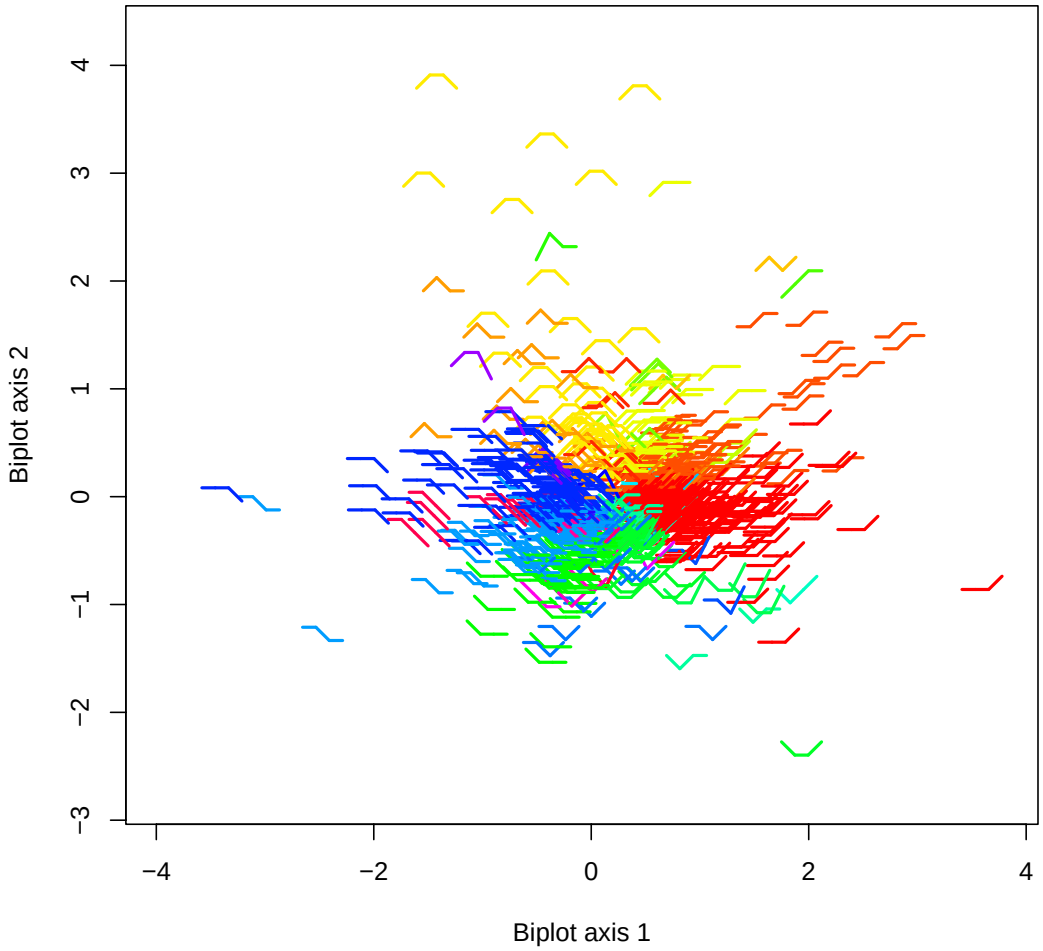

Figure 2: ***Gene*-plot for the *Arabidopsis thaliana* gene expression data. Each gene is represented by a colored icon, showing the patterns identified using a multiplicative model. Each color from the spectrum represents a different pattern class. The actual coordinates are at the left hand point for the icon. Genes with pattern 1111 are not shown.**

When an additive model is used, there are 38 patterns distinguished from amongst the 74 possible (excluding 1111). Figure 3 shows the patterns for both the multiplicative and additive models in a grid layout. The numbers of genes assigned to each pattern by each model are shown in brackets. The icons indicate the pattern (based on ranks) of up, down and no difference. The actual mean gene expression in each of the treatment groups can be quite varied and if all the genes with a given pattern were plotted simultaneously, the interpretation of a common pattern, in terms of up/down and no difference can be missed.

Only 129 genes have a different pattern depending on which model is used. Most of these differences are a consequence of the different assignments to the pattern of no change, 1111, between the two models. The differences are tabulated in Table 2. The left hand column shows the patterns distinguished by a multiplicative model, followed by the range of different patterns assigned when an additive model is assumed. The number of genes allocated to the different patterns is shown (in italics) following the pattern label, with the values for identical patterns for the two models in bold.

The most common response was a pattern which showed no difference until sometime between 2 and 6 hours after exposure to high light when expression was up regulated (1112). There were 241 and 269 genes with this pattern in the multiplicative and additive model respectively. The 'opposite' pattern, 2221, was close to being the next most common response, with frequencies of 114 and 82 respectively. The next most common response was one which indicated a detectable response to light occurring between half an hour and 2 hours (i.e. patterns 1122 and 2211). Altogether the gene expression values for (105, 108) and (97, 92) genes were found to fit these two patterns when assuming multiplicative and additive models respectively.

Generally the differences in pattern allocation between the two models tended to be that the log-Normal model differentiated between the lower ranked means more often than the gamma model.

Initial biological interpretation of some of the patterns is promising. For example, the early light induced protein (ELIP1), one of the genes showing an early stress response pattern (1233), has been described by Adamska [4] as a component of the photo protection mechanism. A mitochondrial glutathione peroxidasea with a 1234 modeled base pattern has been reported by Navabpour et al. [11] to be induced by oxidative stress. There was also a relatively high percentage of up-regulated genes involved in transcription, communication and stress-response which are thought to be components of early signalling pathways leading to morphological and physiological adaptations to the new environment.

3 General Discussion and Conclusion

Even before recent technological advances enabled the simultaneous measurement of an extremely large number of measurements on each individual sample, the literature on multiple comparisons was already voluminous; see Cook and Dunnett [5]. Indeed, Tukey acknowledged the 'usefulness of - and need for - a variety of procedures'. Now, the additional challenge for the problem of multiple comparisons lies in also accommodating the problem of multiple hypothesis testing associated with the simultaneous evaluation of large numbers (literally thousands, or more) of genes. In the context of microarray gene expression data, a relatively recent review of the many approaches proposed and under development for multiple hypothesis testing can

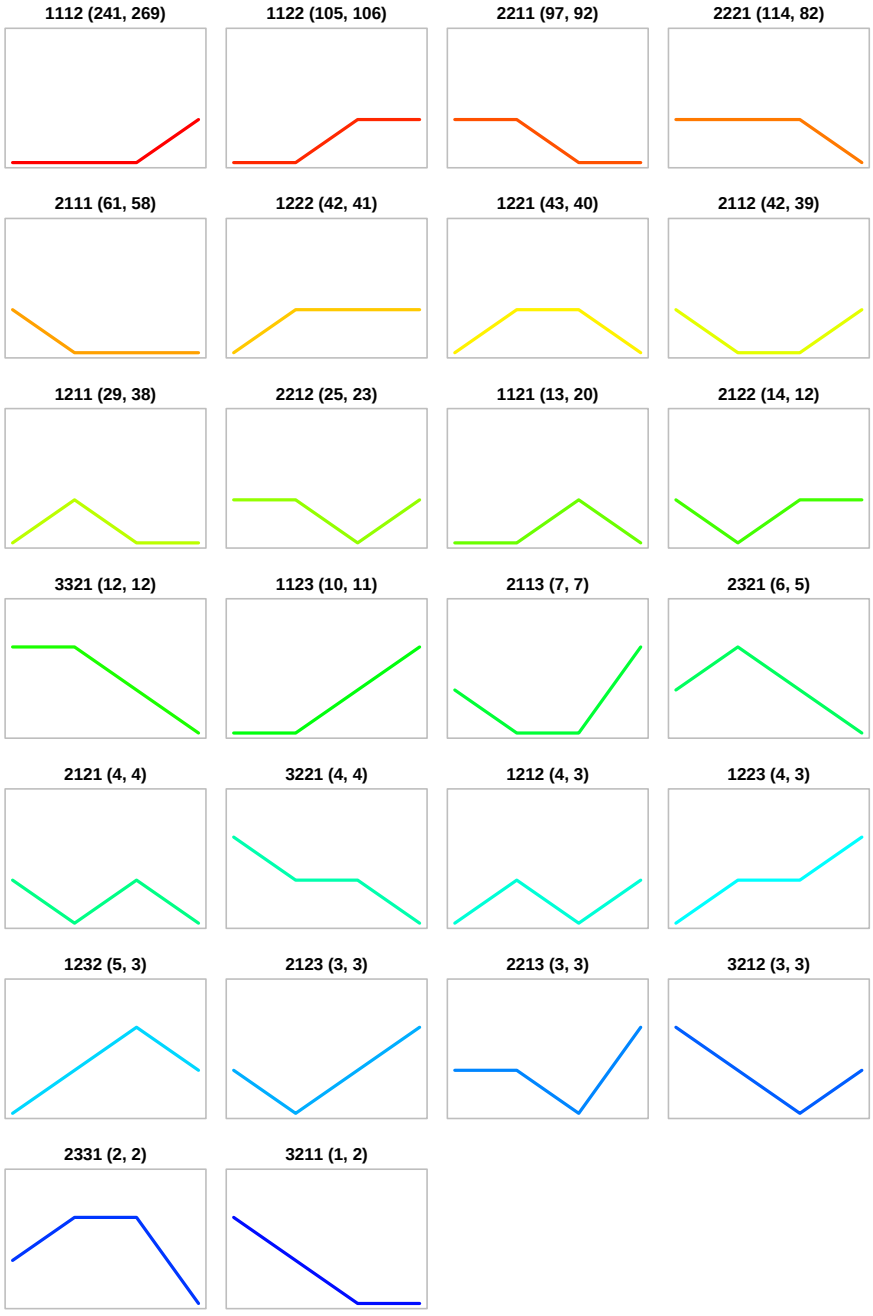

Figure 3: Grid layout of 26 of the gene pattern classes for the *Arabidopsis thaliana* gene expression data; patterns that were only observed once (see Table 2) are not shown here. The numbers of genes assigned to each pattern class for (i) the multiplicative, (ii) the additive model are shown in brackets. The icons indicate the patterns of up, down and no difference.

be found in McLachlan et al [10] and references therein. Note, to avoid confusion, we find it helpful to use the term 'multiple comparisons' for distinguishing between groups of samples (here arrays/ chips) and the term 'multiple hypothesis testing (evaluation)' for the simultaneous consideration of many variables (gene expression and EST values).

In this paper we did not attempt to give the definitive approaches to multiple comparisons and to multiple hypothesis evaluation that naturally arise when one wishes to describe the biologically important problem of determining differential expression patterns for microarray data. For both these problems, considered separately as well as together, we would apply the Tukey quote above. Our preferred approach for the type of data used in the study described in section 2, is basically motivated by the Fisherian evidential *p*-values that in turn were originally motivated by applied researchers (Hubbard and Bayarri [7]). We are well aware that selection of the values of *p* is arbitrary, but we found the values we chose here to provide results of interest to the experimental scientists. Another choice we made, among the many available, was the likelihood modeling approach and the use of AIC. Applying different approaches will certainly alter the precise details in the results, but not the overriding conclusions that we found. Further, a final step that is beyond the scope of this paper is to organize the genes within each pattern in terms of their biological interest, and/or size of the effects.

Motivated by (a) the use of both the log-Normal model (commonly) and the gamma model (less commonly) in the microarray literature, (b) the development of Affymetrix's recent PLIER (Probe Logarithmic Intensity ERror) algorithm that essentially proposes the measurement model changes from 'additive' for low values to 'multiplicative' for values far from zero, and (c) the sometimes spectacular differences in interpretation between the log-Normal (multiplicative) model and the gamma (additive) model that we had previously observed (see Figure 1), we wished to compare these two models for the plant data. To do so we devised the methods outlined in this paper. Interestingly, as can be seen from Table 2 and Figure 3, the results from using both models are reasonably homogeneous, and after examining gene expression values for the genes which had different patterns we found no example of the spectacular differences in response profiles as seen in Figure 1. Here the types of experiments were quite different, possibly reflecting different underlying biological processes. The changes in gene expression in response to exposure to (natural) light in *Arabidopsis* plants were found to be generally slight while changes in gene expression in a stimulation experiment may be quite dramatic for some genes. Further, it is worth noting that often many differences occur due to the multiplicative model tending to differentiate the lower ranked means, but such differences may correspond to the gene expression values where the additive model is actually more appropriate.

In exploring gene expression microarray data, as well as interpreting and communicating the results of microarray modeling to biologists, visualisation of the data is fundamentally important. Here we have presented three useful ideas, namely (i) the *Gene*-plot, (ii) a grid layout and (iii) the use of 'icons', to communicate the different observed patterns of relatively up, down and no difference between the different treatments. The *Gene*-plot is one method of display that can provide a visual summary of patterns along with information on variance and correlation. Other dimension reduction or mapping methods can be used for displaying gene expression patterns by augmenting the usual displays with appropriate icons. For all such methods, care needs to be taken in choosing the appropriate centring and scaling (Pittelkow and Wilson [13]). Canonical variation (bi)plots are another alternative for some designs.

Various methods could be chosen to communicate the different patterns. We have found a

visual presentation using icons which communicate the results of statistical evaluation to be very effective, and note that alternative formats than those given here could be used to display the information. For example, for these *Arabidopsis* data one could alter the relative placement of the treatments to, say, reflect the different times of light exposure for the three 'treatments' rather than use equal spacing as in Figure 3, although we note that for the scale appropriate for these data, the visual patterns were not as clear. In other experimental situations, where the treatments are not ordered, other considerations would indicate which order would be best for communicating results.

In conclusion, we would like to emphasise that it is our experience there is no 'one-size-fits-all' approach to modeling microarray data. The approaches we have presented were selected as giving results that have proven useful for experimental biologists.

References

[1] Affymetrix, Affymetrix Data Mining Tool User's Guide, *Affymetrix Inc.* 2001.

[2] Affymetrix New Statistical Algorithms for Monitoring Gene Expression on GeneChip, *Affymetrix Inc.*, 2001.

[3] M. Aitkin. Statistical modelling: the likelihood approach. *The Statistician* **35**:103-113, 1986.

[4] I. Adamska. ELIPs: light-induced stress proteins. *Physiol Plant*, **100**:794-805, 1997.

[5] R.J. Cook and C.W. Dunnett, Multiple Comparisons., in *Encyclopediaof Biostatistics.*, Eds: Armitage, P. and Colton, T., *Wiley*:3383-3393, 2005.

[6] L.C.A. Corsten and K.R. Gabriel. Graphical exploration in comparing variance matrices. *Biometrics*, **32**(4):851-863, 1976.

[7] R. Hubbard and M.J. Bayarri, Confusion over measures of evidence (*p*'s) versus errors ($\alpha's$) in classical statistical testing., *The American Statistician* **57**(3):171-182, 2003.

[8] I. Lönnstedt and T.P. Speed, Replicated microarray data., *Statistica Sinica* **12**:31-46, 2002.

[9] P. McCullagh, and J.A. Nelder, *Generalized Linear Models*., *Chapman and Hall*, 1989.

[10] G.J. McLachlan, , K-H Do and C. Ambroise, *Analyzing Microarray Gene Expression Data*., *Wiley*, 2004.

[11] S. Navabpour, K. Morris, R. Allen, E. Harrison, S. A-H-Mackerness and V. Buchanan-Wollaston, Expression of senescence-enhanced genes in response to oxidative stress. *J Exp. Bot.* **54**(391):2285-2292, 2003.

[12] Y.E. Pittelkow and S.R. Wilson, Visualization of Gene expression Data - the *GE*-biplot, the *Chip*-plot and the *Gene*-plot., *Statistical Applications in Genetics and Molecular Biology*, **2**(1), 6, 2003.

[13] Y.E. Pittelkow and S.R. Wilson, Use of Principal Component Analysis and of the GE-biplot for the graphical exploration of gene expression data. *Biometrics*, **61**(2):630-632, 2005.

[14] E. Rosche, Screening and analysis of transcription factors in *Arabidopsis thaliana*. *In Preparation.*, 2006.

[15] J.W. Tukey, The Philosophy of Multiple Comparisons., *Statistical Science* **6**(1):100-116, 1991.