The *GE*-biplot for Microarray Data

[1]Pittelkow, Y.E. [1,2*]Wilson, S.R.

[1]*Centre for Bioinformation Science*

[2]*Centre for Mathematics and its Applications*

[1,2] Mathematical Sciences Institute

[1]John Curtin School of Medical Research

Australian National University

Canberra ACT 0200

Australia

*Correspondence should be addressed to:

Sue.Wilson@anu.edu.au

☎ (61-2) 6125-4460

ABSTRACT

The *GE*-biplot (gene expression biplot) is a useful method for exploratory analysis of microarray data that allows a visual appraisal of both genes and samples. Following a brief outline of the underlying method an application is given showing how to find genes that are differentially up and down regulated in subsets of the chips.

CATEGORY

- Microarray Data Analysis
- Software Application

Keywords: *Microarray, gene expression, biplot, ordination, visualization*

1. INTRODUCTION

Currently most gene expression studies have an exploratory (hypothesis generating) component. Hence we have developed the *GE*-biplot to aid visual exploration of microarray data. The particular emphasis of this tool is to find genes that have a similar pattern of up/down regulation for the samples (regardless of expression level). By simultaneously displaying both the samples and the genes on the same plot it is possible to both visually detect genes which have similar 'profiles' and to interpret this pattern in reference to the samples.

Proc Virt Conf Genom and Bioinf (2):8-11
ISSN 1547-383X

www.virtualgenomics.org

2. METHOD AND IMPLEMENTATION

The results of a gene microarray experiment, after pre-processing including "normalization", can be organized in an expression level matrix, **M**, with N_c rows and N_g columns, where N_c is the number of chips/slides, or samples, and N_g is the number of genes. Simulations (described in [10]) showed that to find genes that are differentially up/down regulated, assuming a multiplicative model, the matrix **Z** is computed from **M** by (i) log transforming (ii) chip (row) standardising these values, and then (iii) column (gene) centring at zero. Next let $\mathbf{U}_{(2)}$ and $\mathbf{V}_{(2)}$ represent the first two columns of the left and right singular vectors (respectively) of **Z**, and let $\mathbf{\Lambda}_{(2)}=diag(\lambda_1, \lambda_2)$, where $\lambda_1 \geq \lambda_2$ are the two largest singular values of **Z**. Then a rank-2 approximation to **Z** can be written as

$$\mathbf{Z}_{(2)} = (\mathbf{U}_{(2)}\mathbf{W}\mathbf{\Lambda}^{\alpha}_{(2)})(\mathbf{\Lambda}^{\mathbf{1-\alpha}}_{(2)}\mathbf{W}^{-1}\mathbf{V}_{(2)}^{\mathbf{T}}) = \mathbf{C}\mathbf{G}^{\mathbf{T}},$$

where $\mathbf{W} = diag((N^*)^{\delta})$, $\mathbf{C}=\mathbf{U}_{(2)}\mathbf{W}\mathbf{\Lambda}^{\alpha}_{(2)}$ represents the chips (rows), and $\mathbf{G}=\mathbf{V}_{(2)}\mathbf{W}^{-1}\mathbf{\Lambda}^{\mathbf{1-\alpha}}_{(2)}$ represents the genes (columns). With δ=0.5, α=0 and $N^*=N_c$ (or N_c-1), the *GE*-biplot is the plot of the two columns of **C** and of **G.** Further details are in [10]. Each row in **C** gives the coordinates for the corresponding chip in two dimensions. The chip is shown in the following plots as a filled circle and a line joining the circle to the origin. Each row of **G** gives the two coordinates for the corresponding gene. The position of the genes is shown in the following plots by the symbol '∼'.

With this factorization of **Z** into **C** and **G** and the choice of **Z** given above, the distance between each pair of chips approximates the standardized Euclidean distance between the chips, and the distance between each pair of genes approximates the variance of their differences. Further, the distance from the origin to a gene approximates the variance of that gene, and the cosine of the angle between two genes approximates the correlation between these genes. So positively correlated genes lie close together and negatively correlated genes lie opposite one another. More variable genes lie further from the origin. Further interpretations are described in [10]. A fit index for the approximation of variances amongst the genes is $I = (\lambda_1^4 + \lambda_2^4) / (\lambda_1^4 + \ldots + \lambda_r^4)$, where r

is the rank of **Z**. Further, any gene which lies close to a chip or group of chips will be relatively highly expressed (up regulated) on that chip or chips, and if it lies opposite, will be relatively lowly expressed (down regulated). This is a consequence of the factorization. The closer the gene is to the chip vector and the greater the distances from the origin, the greater the up-regulation.

The axes (the first two columns of **G** and **C**) are not generally interpreted; rather it is the relationships between genes and chips which are of interest. Rotation or reflection will not alter the distances, but it will change the axes. Clarification of structure, say groups of chips or gradients, is determined by the pattern of high-low expression values for the genes as described in the preceding paragraph. For this reason, and because of the inclusion of both genes and chips on the same plot, we do not label the axes in the *GE*-biplots. The horizontal axis corresponds to the first column of **C** or **G**, while the vertical axis corresponds to the second column in all the plots. Because distances are being interpreted it is important however to show the scales and to have equal scales on both axes.

The algorithm to implement the *GE*-biplot has been programmed in the open source software package, R, and is being made available on the Bioconductor web site (http://www.bioconductor.org). The preprocessed data described in the next section is also available at this site. Further details on the biplot also can be found in Corsten and Gabriel [3], Gabriel [5], [6] and Gower and Hand [9]. A note on an application of a biplot version of PCA to microarray data is given by Chapman et al., [1].

3. APPLICATION

To demonstrate the *GE*-biplot we use the training data set of Golub et al [8]. This data set consists of absolute gene expressions from 38 human Affymetrix high-density oligonucleotide chips. Bone marrow samples from two types of leukaemia patients, Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML), were hybridized to the chips. Of the 27 ALL samples, 8 were T-cells and 19 B-cells. There were 11 AML samples. After preprocessing which involves threshholding and filtering (see Dudoit et al. [4]) the data consist of 3030 genes on 38 chips.

The *GE*-biplot for these data is shown in Figure 1. From the position of the genes relative to the chips it can be deduced that genes which lie approximately close to the AML samples would tend to have higher values on the AML samples compared with ALL samples. Those genes lying in the opposite direction will have higher values for the ALL samples compared with AML samples. Genes lying approximately towards the north or south will tend to differentiate the T-cell from the B-cell ALLs.

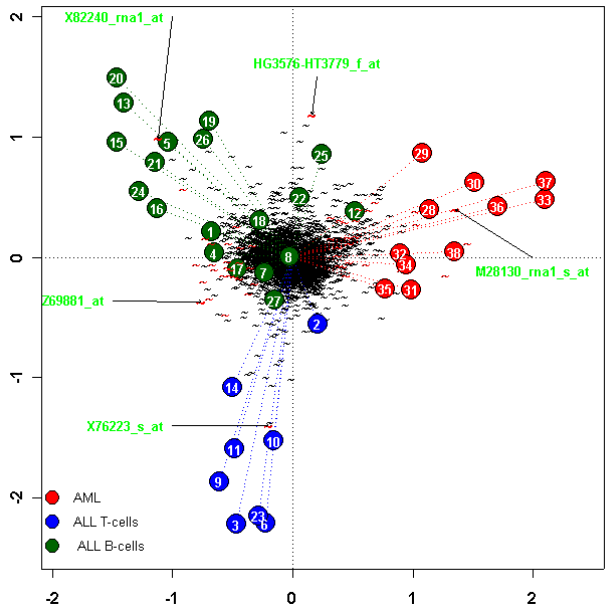

Figure 1. The *GE*-biplot for these data with AML data shown as red, ALL T-cells as blue and ALL B-cells as dark green, and the '~'s represent the genes. One can see that the AML samples (red) fall in a fairly distinct cluster compared with ALL (green B-lineage and blue T-lineage).

If one were interested in finding genes which were differentially expressed between the AML and ALL samples, those genes lying close to the AML samples, or opposite, would be selected for further investigation. Since the genes that are more distant from the origin have greater variance they would tend to have greater differential expression (or it could indicate the presence of abnormal chip/s or genes). For illustrative purposes five genes, as indicated in Figure 1, were examined in more detail. Figure 2 shows the transformed relative intensity values of the genes in the three subgroups, with lines joining the mean values in the B-cell ALL, T-cell ALL and AML samples respectively. The top two graphs show data for two genes which were also chosen by Golub et al., [8] to differentiate the two types of leukaemia. The graphs demonstrate that these genes are differentially expressed between the two types of leukaemia, but in opposite directions as would be deduced from their position on the *GE*-biplot. Statistical analyses support this conclusion, although we do not detail the methods used here, as the intent is to describe and demonstrate the *GE*-biplot. Further, another

advantage of visualization is that appropriate hypotheses and statistical tests are more likely to be used. For example a test assuming equal variance would not always be appropriate. The presence of outliers, especially in the lower graphs, would suggest use of a robust or non-parametric test statistic. Alternatively, a decision might be made to discard the outliers, based perhaps on some auxillary information. Statistical evaluation would normally be followed by further validation in the laboratory.

The bottom three graphs show the data for the three other genes. From their position in Figure 1, it would be expected that gene X76223_s_at would be more highly expressed on ALL T-cells (close to the ALL T samples), X82240_rna1_at, being close to the ALL B-cells, would be expected to more highly expressed in these samples, and HG3576-HT3779_f_at would be expected to be down regulated in the ALL T-cells, but equally expressed on the ALL B-cell and AML samples. All these patterns can be seen in the lower graphs of Figure 2.

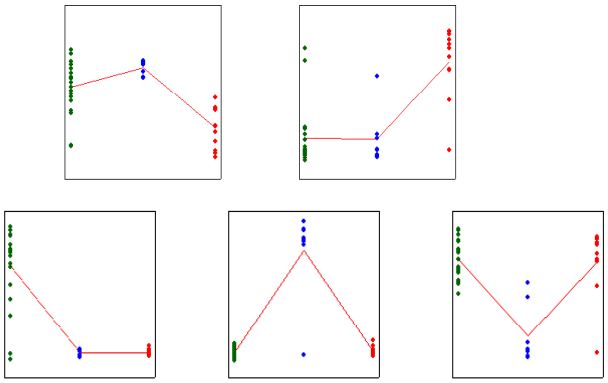

Figure 2 Upper panel: Two of the selected 50 genes in [8] to differentiate AML from ALL, Z69881_at, M28130_rna_s_at; Lower panel from left to right: X82240_rna1_at, X76223_s_at and HG3576-HT3779_f_at. ALL B-cells are shown as dark green, ALL T-cells as blue and AML data as red.

The fit index is 59% for the *GE*-biplot in Figure 1, suggesting only a moderate fit. To improve fit, one might examine the data in higher dimensional space, or examine a subset of the data (gene selection). Since many genes in a microarray experiment are not differentially expressed, it is often preferable to use gene selection methods as an initial strategy to improve fit. Many gene selection methods have been proposed and the ordination does change dependent on the genes selected. For these data Golub et al determined a subset of 50 genes that were most closely correlated with the AML-ALL distinction. These genes are highlighted in red in Figure 1. Choosing these genes, the resultant *GE*-biplot is shown in Figure 3.

Additional information on the samples, have been added to the *GE*-biplot. The gender of the patient, where known is shown as 'M' or 'F'. The success (+) or otherwise (-) of chemotherapy applied to AML adult patients is indicated. The area of the circles for the samples is proportional to the corresponding prediction strength score derived in [8]. The fit is excellent (98%) and as expected, the genes lie towards, or away from, the AML cluster. Further ordinations of these data are given in [10].

The *GE*-biplot is proving to be very useful visualization tool for understanding microarray data. Genes can be selected for further detailed analyses following a visual inspection of the *GE*-biplot. The *GE*-biplot can also be used to detect outliers, to visually present the results from other analyses (as in figure 3), and as an exploratory visualization tool. An additional advantage compared with 'standard' color maps is that there is no need to find an explicit ordering of the genes and the samples. Further, by incorporating a third axis, the above can be generalized to a bimodal (which is a three-dimensional analogue of a biplot, see [9]). Then application of visual techniques, such as rotation, can be used to look for 'clusters'.

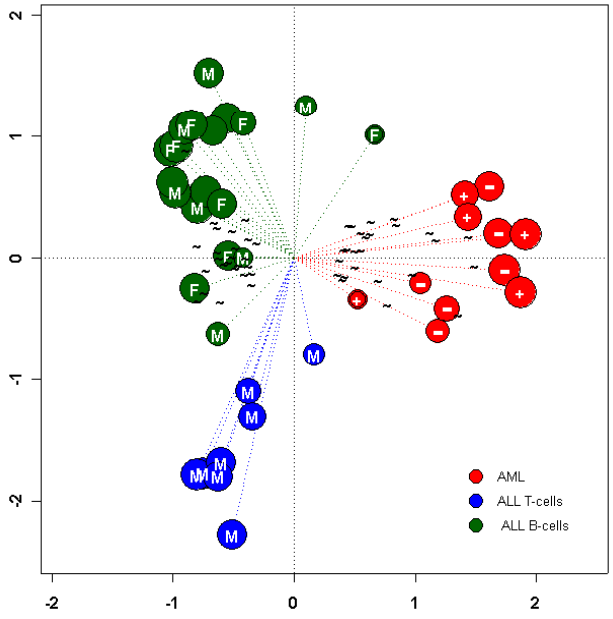

Figure 3. *GE*-biplot of the 38 chips using the 50 genes selected in [8].

Visualisations, such as the *GE*-biplot, also can be used to help refine statistical analyses. In the Golub example, the *GE*-biplot clearly demonstrates the separation of the T-cell and B-cell samples, and if this had been observed, it is unlikely that as much effort would have given to

analyses of these data that concentrated on only the AML-ALL structure (for example Ge et al [7]). Further, if the confounding of gender with T-cell and B-cell samples had been noted by Chow et al [2] in their extended data set, they might not have looked (unnecessarily) for both T-cell/B-cell markers and Male/Female markers.

Other ordination methods based on the appropriate singular value decomposition are also useful additions to the microarray analyst's toolkit; see for example [10]. The advantage of biplots over some alternative ordination techniques is that patterns discerned on biplots can be translated algebraically into a model for the data. Note that simulations (not shown here) showed that quality of the ordinations proposed in the literature vary markedly. The ultimate guide to the use of any data analysis method is its impact on our biological understanding. Currently our knowledge of the underlying biology, particularly of gene regulation, is limited. Hence, exploratory techniques such as the *GE*-biplot are invaluable.

6. REFERENCES

[1] Chapman, S., Schenk, P, Kazan, K, Manners, J., 2001. Using Biplots to interpret gene expression patterns in plants. Bioinformatics, 8, 1, 202-204.

[2] Chow, M.L., Moler, E.J., Mian, I.S., 2001. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. Physiol. Genomics, 5, 99-111.

[3] Corsten, L.C.A., Gabriel, K.R., 1976. Graphical exploration in comparing variance matrices. Biometrics, 32, 4, 851-863.

[4] Dudoit, S., Fridlyand, J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumours using gene expression data. J. Am. Statist. Soc., 97, 77–87.

[5] Gabriel, K.R, 1971. The biplot graphical display of matrices with applications to principal component analysis. Biometrika, 58, 453-467.

[6] Gabriel, K.R., 1981. Biplot display of multivariate matrices for inspection of data and diagnosis. In: V. Barnett (Ed), Interpreting Multivariate Data Wiley, New York, 147-173.

[7] Ge, Y., Dudoit, S., Speed, T.P., 2003. Resampling-based multiple testing for microarray data analysis (with discussion). TEST, 12, 1-78.

[8] 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., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286, 531-537.

[9] Gower, J.C., Hand, D.J., 1996. Biplots. Chapman & Hall, London.

[10] Pittelkow, Y.E., Wilson, S.R., 2003. Visualisation of gene expression data – the GE-biplot, the Chip-plot and the Gene-plot. Statistical Applications in Genetics and Molecular Biology, 2, 1, 6.