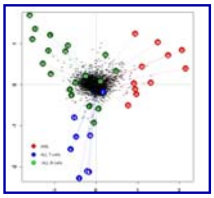
GE-biplot Microarray data

"An approach to the ordination of Gene Expression Data - the GE-biplot"

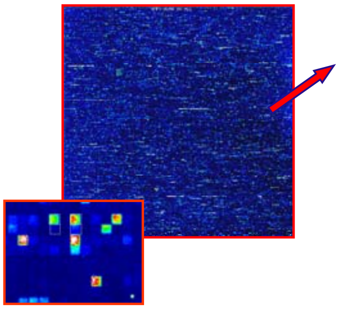
Y.Pittelkow
S.R.Wilson
CBiS
MSI ANU

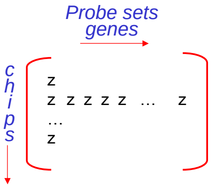
Outline

1. The Bi-plot
   - *GE*-Biplot
2. Applications
   - Simulated Data
   - Colon Data
   - Leukemia Data



GeneChip Data



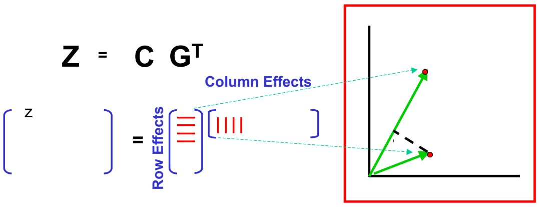


Expression values
Summary of the PM and/or MM Intensity values

Biplot

Graphical Display of a Rank 2 matrix

$$\mathbf{Z} = \mathbf{C}\ \mathbf{G}^T$$



Rao, 1965
Gabriel, 1971 Biometrika

Inner Product
- Proportional
- Zero entries

Biplot - Microarray Data

$$\mathbf{Z} = \mathbf{C}\ \mathbf{G}^T$$
$$= \{CR^T\}\{GR^{-1}\}^T$$

- **Factorization**
  - metric in which to represent the data.
- **Approximation of Z** by a matrix of rank 2
  - SVD
- **Choice of Z**
  - Expression level transformations ? (logs)
  - Row (chip) standardizing? (Scaling, Normalization)
  - Column (gene) mean correcting or standardization?
  - Gene selection (filtering) ?

Approximate Biplot

$$Z_{(2)} = U_{(2)}\Lambda_{(2)}V_{(2)}^T$$

Eckhart and Young 1939
Good 1969

- cols of V = right singular vectors
  = eigenvectors of Z^tZ
- cols of U = left singular vectors
  = eigenvectors of ZZ^t

Λ = diag$(\lambda_1,\lambda_2, \ldots)$ $\lambda_1>\lambda_2 \ldots>0$
= diag ($\surd$ eigenvalues)

- Minimizes $\Sigma_i\Sigma_j(z_{ij}-z_{(k)ij})^2$
- Goodness of Fit is
$\Sigma^k_{i=1}\lambda_i^2/\Sigma^r_{i=1}\lambda_i^2$

$$Z = \underbrace{\begin{bmatrix} u_1 & u_2 \\ | & | \end{bmatrix}\begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix}^{\alpha}}_{C}\ \underbrace{\begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix}^{1-\alpha}\begin{bmatrix} v_1 & — \\ v_2 & — \end{bmatrix}}_{G^T}$$

GE-biplot

Factorization

$Z=CG^T= (U\Lambda^{\alpha})(\Lambda^{1-\alpha}V^T)$

- **If** $\alpha=0$, $C= U$ and $G^T=\Lambda V^T$

Then if Z is mean corrected

and $C= \sqrt{N_c}U$ and $G=(1/\sqrt{N_c})\Lambda V^T$

$GG^T= S_g$

Goodness of Fit variances

$\Sigma^k_{i=1}\lambda_i^4/\Sigma^r_{i=1}\lambda_i^4$

Matrix Z

•Log 2 expression level

•Chips - standardized

•Genes - mean corrected

Display

•Chips - vectors with annotation.

•Genes - points or symbols

Variant of the h-plot

Corsten and Gabriel (1976)

Focus: Similarity of up down regulation.

Simulation Study

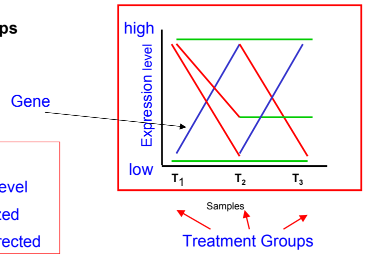
Multiplicative model

'True' expression values

low (1), medium (3), or high (9)

- **27 genes**
- **3 treatment groups**
- **3 replicates**

Every combination



Matrix Z

1. Log 2 expression level
2. Chips - standardized
3. Genes - mean corrected

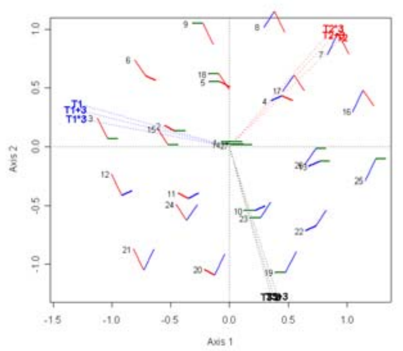


Fig. 1 GE-biplot Simulated data

Goodness of fit .98

Note

Chips

Replicates

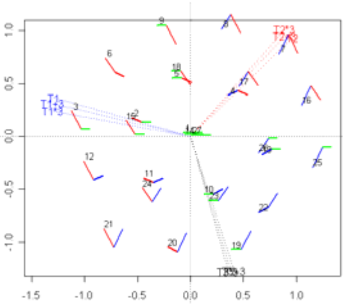
Different treatments

Genes

•Similar up down regulation

•Non responding genes

Some Interpretations



Genes

- σ Euclidean distance
- ρ Cosine of the angle

Chips

- Δ Mahalanobis Distance

Inner products

- $z_{ij} = c_i^T g_j$
- Difference vectors

$z_{ij}-z_{ij'} = c_i^T(g_j-g_{j'})$

$z_{ij}-z_{i'j} = (c_i^T- c_{i'}^T)g_j$

Colon Data

Data **62 x 1988**

- 22 matched normal and tumour colon tissue samples
- 18 unmatched tumour colon tissues

Pre-processing steps

- Filtering ; selection of the 2000 'highest minimal intensity' genes.
- considerable skewness

Problems

•mixture of matched and unmatched samples

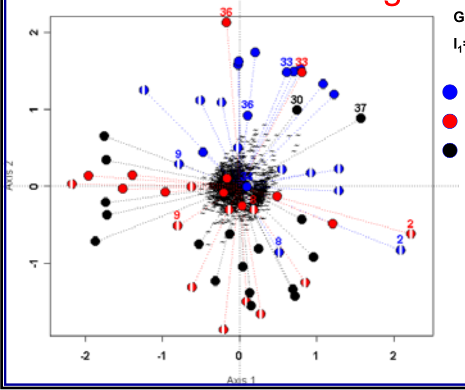
•change of protocol after the first 11 samples of matched pair data

•contamination with muscle tissue (Alon et al., 1999).

http://www-genome.wi.mit.edu

Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., Levine A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumour and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A. 96(12) 6745-50.

Colon data -1988 genes



Goodness of Fit

l_1**= .26 (.51)**

Early

● **Normal** ◐

● **Matched tumors** ◐

● **Un-matched tumors**

Alon et al.

Getz et al.

Chow et al.

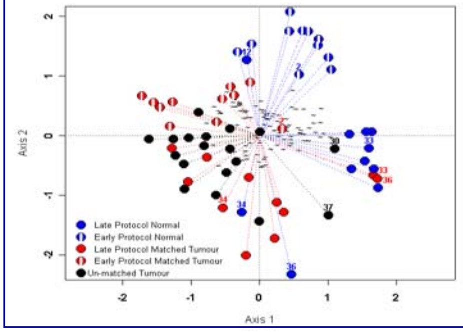
McLachlan et al.

Ben-Dor et al.

Gene Selection
MAD 100 genes

Goodness of Fit

l_1= .41 (.79)



Leukaemia Training Data Set

Data **38 x 7129**

- **•11 Acute Lymphoblastic Leukaemia (AML)**
- **• 27 Acute Myeloid Leukaemia (ALL).**
  - **• 8 T-cells**
  - **• 19 B-cells**
- **•Pre-processing steps as described in Dudoit et al (2002) + removal of controls**
  - •Thresholding : floor 100 ceiling 16000
  - •Filtering : max/min $\leq$ 5 and max-min $\leq$ 100
  - •3030 genes Log$_2$
- **•Censored data**

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., Lander, E.S., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression profiling. Science, 286, 531-537.

Leukaemia Data (3030 genes)

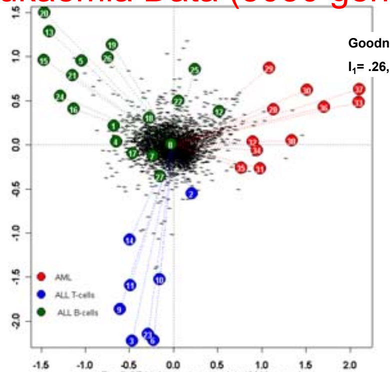


Fig. 5 GE-biplot Leukaemia data (3030 genes)

Goodness of Fit

l_1= .26, l_2=.6

Golub et al.

Dudoit et al.

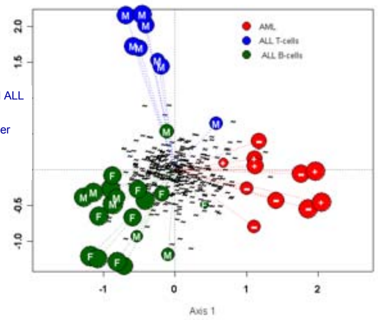
Ge et al.

Korenberg

10% Most varying genes

Childhood ALL patients

Dana-Faber Cancer Institute protocol



Goodness of Fit

l_2= .74

Adult samples Cancer and Leukemia Group B Leukemia bank.

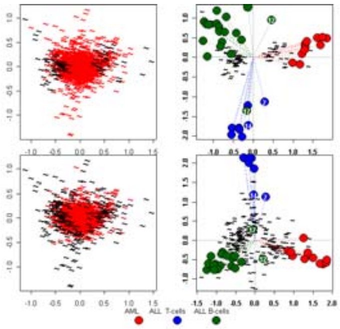


Fig. 7 . Top: Ge et al selection (92 genes) . Bottom: Anova Gene selection (245 genes)
Left: Gene-plot Right: GE-biplot

Gene Selection

Ge et al.

l_1=.65 ,l_2=.97

Adjusted p <.05

Independent F-tests

3 groups

l_1=.63 ,l_2=.97

p< .00001

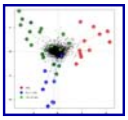
Summary

1. The biplot is a useful visualization tool for microarray data

   Simultaneous plotting of the genes and chips on the same plot

2. Many types of biplots

   - *Factorization,Rank 2 approximation,Matrix*
   - *GE*-Biplot



CSDA -Computational Statistics & Data Analysis Journal

- Special issue on bioinformatics & biostatistics'.
- Possibly an issue on the analysis of microarray data