

Gene Clustering Algorithm GS4M5D

Wim de Boer
VU Medical Center and
Dept. of Molecular Cell Physiology
BioCenter Amsterdam
Vrije Universiteit Amsterdam
De Boelelaan 1087
1081 HV Amsterdam, The Netherlands
wdeboer@wanadoo.nl

Jan Lankelma
VU Medical Center and
Tumor Cell Biology group
Dept. of Molecular Cell Physiology
Vrije Universiteit Amsterdam
De Boelelaan 1087
1081 HV Amsterdam, the Netherlands
lankelma@bio.vu.nl

ABSTRACT

DNA micro-arrays allow quantitative measurement of the expression of thousands of genes in a biological sample. DNA micro-arrays can be used to measure changes in gene expression

1. Of related samples, such as clinical samples taken from different cancer patients, or
2. During a biological process, such as the reaction of a patient to chemotherapy.

Gene shaving

The gene shaving method recently proposed by R. Tibshirani and others from Stanford University^[1] is designed to identify clusters of genes with *coherent* expression patterns and *large variation* across the samples. Genes with large variation are found by invoking the singular value decomposition^[2] (SVD) of the entire expression matrix, each row of the matrix containing the expression of a gene over the samples. It can readily be shown that the so-called “eigen gene” - a linear combination of all rows with weight factors equal to the components of the first left singular vector of the expression matrix - has maximum variance over all genes. Many of the weight factors are vanishing small, so setting the smallest say 10% of them equal to zero will have a minor effect on the variance of the eigen gene and reduces the number of genes involved. Now recalculating the eigen gene for this reduced set of genes, again setting the smallest weight factors equal to zero, and so on, one finally has one gene left, which by construction will have a large variance. For more detail, see^[3].

Algorithm GS4M5D

The present algorithm is based on the “Gene Shaving” method but differs in that the Spearman rank-correlation test^[4] is used to determine cluster size.

Thus the present algorithm runs as follows:

1. Start with the entire expression matrix, each row centered to have a zero mean.
2. Compute the leading principle component of the rows of the expression matrix.
3. Shave off a proportion (typically 10%) of the rows having smallest inner product with the leading principle component.
4. Repeat steps 2 and 3 until only one row remains.
5. Compute for each row the Spearman rank-order correlation coefficient with the one row and select those genes to belong

to the cluster that meet a statistical significance level, chosen a priori.

6. Orthogonalize each row of the expression matrix with respect to the average gene in the cluster.
7. Repeat steps 1 to 4 with the orthogonalized data and repeat step 5, to find a second cluster. This process is continued until the vector space spanned by the genes is fully covered.

Use of the Spearman rank-order correlation method ensures that only those genes are selected that are really correlated, that is, to the significance level chosen. Accordingly the clusters will comprise all genes, *including those of small variance*, that are positively or negatively correlated with the one gene found by shaving.

Noise on the gene data could give rise to spurious clustering, so it is a good idea to rerun the algorithm on data with noise added.

Application

The algorithm has been applied to the diffuse large B-cell lymphoma data set^[5] and some results will be shown.

REFERENCES

- [1] Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D, and Brown P: Clustering methods for the analysis of DNA microarray data. Technical report. Stanford University: Department of Statistics, 1999.
- [2] Golub GH, Van Loan CF: Matrix Computations, The John Hopkins University Press, 1983, chapter 2.3.
- [3] Hastie T, Tibshirani R, Eisen M, Brown P, Scherf U, Weinstein J, Alizadeh A, Staudt L, Botstein D: Gene shaving: a new class of clustering methods for expression arrays. Technical report. Stanford University, 2000.
- [4] Press WH, Flannery BP, Teukolsky SA, Vetterling WT: Numerical Recipes, Cambridge University Press, 1987, chapter 13.8.
- [5] Alizadeh AA, and others: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403 (2000) 503-511