Classical Statistical Approaches to Molecular Classification of Cancer using Microarray Technology

Susan Spruill
PPGx
3500 Paramount Parkway Morrisville, NC 27560
USA
919-463-6719
919-379-6102
susan.spruill@ppgx.com
CAMDA00 Dataset 2: Leukemia

Lu J., Hardy S., Tao W. Muse S. Weir B. Spruill S.

Classical Statistical Approaches to Molecular Classification of Cancer From Gene expression Profiling. J. Lu†, S. Hardy†, W. Tao†, S. Muse†, B. Weir† and S. Spruill‡ †N.C. State University Bioinformatics Program and ‡PPGx. Inc.

In a 1999 Nature Genetics article Bittner, et. al. acknowledged that the volume of data obtained from gene expression analysis using microarrays or DNA chip presented a "mathematical challenge". This followed an earlier argument by Duggan et. al that in order to achieve true understanding of genome function, it needed to recruit the assistance of statisticians and mathematicians to ponder the problems of data analysis. Recent literature , , , regarding this new technology has focused on the need to incorporate classical statistical practices in experimental design in order to utilize more robust, classical statistical methologies in data analysis.

The authors of this presentation have demonstrated that classical statistical methods are applicable to analysis of data from Golub, et. al . Our preliminary analysis of all 6817 genes involves simple t-tests for statistically significant separation of means of gene expression level in two cancer types. Our subset of genes which distinguish AML types from ALL types are relatively consistent with those published by Golub. We choose those predictor genes based on the t-values, and evaluate its performance in predicting 34 test samples by linear discriminant analysis. The preliminary result shows that 50 predictor genes can give us 31 correct predictions in 34 samples (which is compared to 29 correct predictions by Golub's method). Three samples that were not correctly predicted are sample 54, 57 and 60, which have not been surely assigned by Golub as well. We chose 20, 50, and 200 genes as predictors and had the same results. We will evaluate the parsimony of our model by evaluating, through a stepwise method, the minimum number of genes required to maintain a high level of accuracy in predicting cancer types.

1 Bittner, M., et. al., 1999, Nature Genetics. Vol 22, pp213-215. 2 Duggan, D.J., et. al., 1999, Nature Genetics, Vol 21, pp10-14. 3 Kaminski, N., et. al. 2000, PNAS, Vol 97.4,pp1778-1783 4 Hilsenbeck, S.G., et. al., 1999, J Nat. Cancer Inst, Vol 91.5, pp453-459 5 Dudoit, S, et. al., 2000, Technical Report #578, www.stat.Berkeley.EDU/users/terry/zarray/Html/matt.html 6 Kerr, M.K., et. al. 2000. Submitted manuscript. www.jax.org/research/churchill/pubs/index.html. 7 Golub, T.R., el. al., 1999. Science, Vol 286, pp531-537.

Keywords

microarray, SAS, discriminant analysis, leukemia

Tools