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.

Introduction:

In a 1999 Nature Genetics article Bittner, et. al.[1] acknowledged that the volume of data obtained from gene expression analysis using microarrays presented a "mathematical challenge". This followed an earlier argument by, Duggan et. al.[2] that in order for microarray research to achieve true understanding of genome function, it needed to recruit the assistance of statisticians and mathematicians to ponder the problems of data analysis. While the technology is novel, the statistical methods appropriate for analyzing microarray data need not be. Recent literature[3,4,5,6] regarding microarray technology has focused on the need to incorporate classical statistical practices in experimental design in order to utilize more robust, classical statistical methodologies in the analysis of microarray data. The authors of this presentation have demonstrated that classical statistical methods are applicable to analysis of data from Golub, et. al.[7]

Methods and Results:

Prior to analysis, data from the training dataset were filtered and standardized according to Golub's methods as described on his web site (www.genome.wi.mit.edu/MPR). T-tests were computed for the remaining 4892 genes and ranked based on t-value. Table 1 shows the 10 genes with largest t-values favoring larger mean AML gene expression and 10 complementary genes where t-values favor larger mean ALL gene expression. From the t-value ranks, 250 genes with the most positive t-values and 250 genes with the most negative t-values were chosen as starter genes for a stepwise discriminant analysis[8].

The stepwise analysis picked 25 genes, without a stopping criteria (see Table 2). Eighteen genes had larger AML means and 7 genes had larger ALL means (see Figure 1). The 25 genes were then used to derive density estimates from non-parametric discriminant analysis on the training dataset of 38 samples with known classification. The test data of 34 samples were then classified based on result from the training dataset. Classifications outcomes on the test dataset were correct for 32 of the 34 samples (see Table 3).

A stepwise discriminant analysis was then run just on the 25 genes chosen on the original analysis to determine whether a reduced model would classify samples with equal accuracy. In the second stepwise analysis the number of variables in the final model was restricted to 5. Of the genes chosen (highlighted in red in Table 2), one favored larger ALL means and the other 4 favored larger AML means (See Figure 1, red squares). These 5 genes were then used to derive density estimates from the training dataset. Results of the test dataset yielded 33 out of 34 correct classifications (see Table 3).

Table 1: Results from t-test

10 Genes where t-value favors AML			
Gene Accession NUmber	Standardized Mean ALL	Standardized Mean AML	t-value*
M27891_at	-0.50819	1.24738	-9.49089
U50136_rna1_at	-0.50909	1.24958	-8.48094
X95735_at	-0.43795	1.07497	-8.11403
M55150_at	-0.42285	1.03791	-7.74088
Y12670_at	-0.46825	1.14935	-7.69674
M23197_at	-0.50289	1.23435	-7.32637
U46499_at	-0.46931	1.15193	-7.23349
M27783_s_at	-0.48877	1.19972	-7.00355
M81933_at	-0.40481	0.99363	-6.89141
D88422_at	-0.48973	1.20207	-6.69879
10 Genes where t-value favors ALL			
X66533_at	0.34207	-0.83963	6.12532
X82240_rna1_at	0.34809	-0.85439	6.28326
M89957_at	0.35314	-0.86681	6.41754
L13278_at	0.38265	-0.93924	6.45551
Z15115_at	0.41605	-1.02121	6.52462
M77142_at	0.41304	-1.01383	6.56631
M28170_at	0.41224	-1.01187	7.39117
M31523_at	0.4596	-1.12810	7.47356
U22376_cds2_s_at	0.49309	-1.2103	7.78074
X52142_at	0.40777	-1.00089	8.05098

Table 2: 25 Genes obtained from Stepwise Discriminant Analysis

Gene Accession Number	Partial R-Square	F Value	Pr > F
D26579_at	0.3624	20.46	<.0001
D38524_at	0.276	13.72	0.0007
D79997_at	0.1373	5.73	0.022
L08177_at	0.5278	40.24	<.0001
L19872_at	0.3186	16.83	0.0002
M12959_s_at	0.3591	20.17	<.0001
M13792_at	0.4184	25.9	<.0001
M27783_s_at	0.6022	54.51	<.0001
M27891_at	0.651	67.16	<.0001
M62762_at	0.2707	13.36	0.0008
M84526_at	0.6685	72.61	<.0001
M95178_at	0.219	10.1	0.003
M98399_s_at	0.4343	27.64	<.0001
S75256_s_at	0.2378	11.23	0.0019
U25128_at	0.2609	12.71	0.001
X03363_s_at	0.1744	7.61	0.0091
X03934_at	0.1336	5.55	0.024
X16832_at	0.1674	7.24	0.0108
X62654_rna1_at	0.4282	26.96	<.0001
X64072_s_at	0.317	16.71	0.0002
X64364_at	0.2706	13.36	0.0008
X95735_at	0.4835	33.7	<.0001
Y00433_at	0.2362	11.13	0.002
Y00787_s_at	0.5408	42.39	<.0001
Z14982_rna1_at	0.2677	13.16	0.0009

Figure 1: Scatter plot of means of AML and ALL for 25 genes and 5 genes used in discriminant analyses.

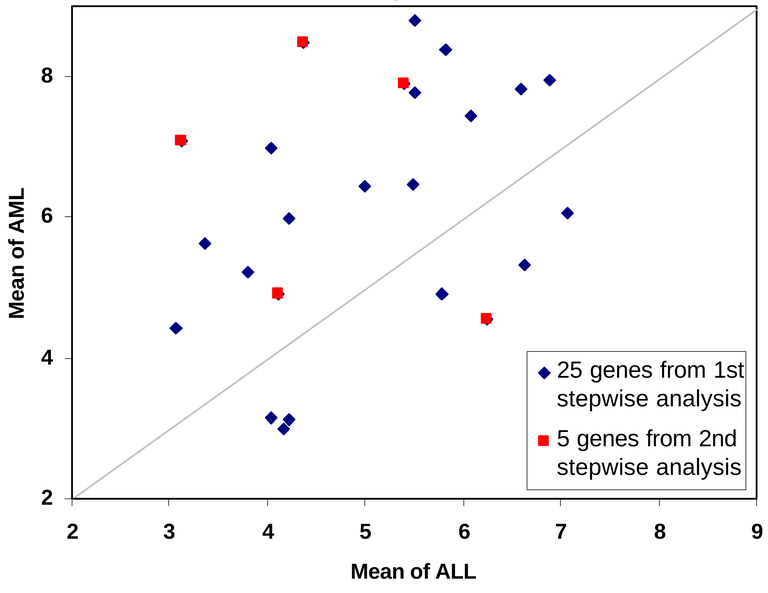

Table 3: Classification Results from Discriminant Analysis

Test Sample	Actual Group Classification	25 Gene Model Classification	5 Gene Model Classification
39	ALL	ALL	ALL
40	ALL	ALL	ALL
41	ALL	ALL	ALL
42	ALL	ALL	ALL
43	ALL	ALL	ALL
44	ALL	ALL	ALL
45	ALL	ALL	ALL
46	ALL	ALL	ALL
47	ALL	ALL	ALL
48	ALL	ALL	ALL
49	ALL	ALL	ALL
50	AML	AML	AML
51	AML	AML	AML
52	AML	AML	AML
53	AML	AML	AML
54	AML	AML	AML
55	ALL	ALL	ALL
56	ALL	ALL	ALL
57	AML	AML	AML
58	AML	AML	AML
59	ALL	ALL	ALL
60	AML	AML	AML
61	AML	*ALL*	AML
62	AML	AML	AML
63	AML	AML	AML
64	AML	AML	AML
65	AML	AML	AML
66	AML	*ALL*	*ALL*
67	ALL	ALL	ALL
68	ALL	ALL	ALL
69	ALL	ALL	ALL
70	ALL	ALL	ALL
71	ALL	ALL	ALL
72	ALL	ALL	ALL

Conclusions:

Simple textbook statistical methods were applied to classify samples from two leukemia types with greater prediction outcomes than the methods used by Golub, et., al. It is notable that only 2 of the 5 genes used in the reduced model are the same as those identified by Golub, suggesting a high level of redundancy among level of gene expression within the leukemia types.

References:

[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., et. al., 1999. Science, Vol 286, pp531-537.
[8]SAS/STAT User's Guide (V6.03), 1988. SAS Institute, Inc., Cary, NC, USA