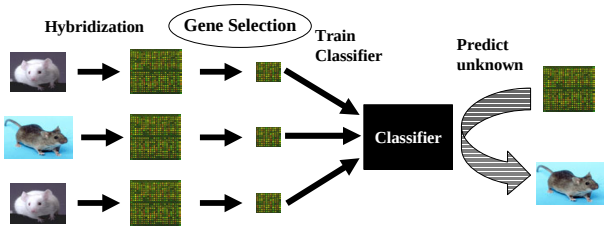
IMPROVED GENE SELECTION FOR CLASSIFICATION OF MICROARRAYS

J. JÄGER[1,2], R. SENGUPTA[1,3,4], W.L. RUZZO[1,4]

[1]Department of Computer Science & Engineering, University of Washington, Seattle, WA
[2]Current address: Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin
[3]Current address: Dept of Computer Science and Engineering, IIT, Kharagpur 721302, INDIA
[4]Department of Genome Sciences, University of Washington, Seattle, WA

Please see PSB03 proceedings for details.

Challenges in Gene Selection

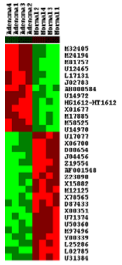


Genes are usually selected ranked by a statistical test score such as a t-test

Description	Adenoma				Normal				t-test P-val		Description	t-test P-val
Human CCG1 mRNA	9.8	13.0	-32.0	17.0	8.7	20.0	17.4	10.7	0.366	X	Macmarcks	0.000
Serine Kinase Psk-H1	-10.3	-35.3	-24.7	-41.2	-2.4	-7.4	-21.4	2.9	0.054	X	Calmodulin Type I	0.005
Mucin (Gb:M22406)	198.1	199.6	160.8	141.1	225.3	184.4	121.6	40.2	0.499	X	Ras-Like Protein Tc10	0.032
Cystatin D	2.3	0.0	1.1	-6.8	-3.4	10.0	4.0	0.5	0.344	X		
Ras-Like Protein Tc21	5.1	12.5	-43.6	7.2	9.2	3.2	-3.5	8.2	0.546	X		
Ras-Like Protein Tc4	59.2	60.2	51.0	30.2	35.0	38.5	42.9	44.1	0.248	X		
Utrophin	87.2	16.3	114.0	48.4	32.5	24.2	23.9	14.5	0.140	X		
Macmarcks	139.9	158.4	144.5	135.1	16.0	59.5	39.9	27.6	0.000			
Elastase 1	74.6	65.6	139.8	81.2	95.7	86.9	95.7	77.0	0.937	X		
Desmoplakin I	28.9	39.6	83.5	19.5	6.3	13.2	17.9	16.0	0.127	X		
Calmodulin Type I	85.8	65.1	54.1	58.6	129.2	118.5	148.5	181.2	0.005			
Ras-Like Protein Tc10	3.7	9.2	6.8	11.9	20.4	35.3	15.9	19.9	0.032			

Problem: This may lead to the selection of many highly correlated genes
Naïve solution of just selecting more genes in order to capture all relevant genes has several problems:

- higher computational cost of classification
- higher cost for biological verification
- possible skew of the classification result



Expression profile for the top genes in the Notterman Adenoma data set. Not surprisingly many genes show a very similar expression profile pattern and have a very high correlation. The additional value of having two "similar" genes is small. It would be better to include different genes that can form a stronger predictor together.

	Exp1	Exp2	Exp3	Exp4	Ctrl1	Ctrl2	Ctrl3	Ctrl4	t-test P-val
Gene A	0.7	-0.2	0.1	0.6	0.1	-0.4	0.5	-0.1	0.3706823
Gene B	-0.3	0.6	0.3	-0.2	-0.4	0.1	-0.8	-0.2	0.1857501
Gene A+B	0.4	0.4	0.4	0.4	-0.3	-0.3	-0.3	-0.3	2.646E-50

Artificial data example showing that two weak genes together can form a very strong predictor

Proposed solution

Overall classification accuracy can be improved by deliberate, careful, selection of gene sets that are not highly correlated. In particular, selecting representatives from different clusters will give improved classifiers.

Approach:

- group similar genes using **clustering or correlation**
- select only representative and informative genes from these groups to **avoid redundancy**
- find parameters for optimal classification

Comparison of classifiers using five test statistics:

- Fisher
- Golub
- Wilcoxon
- TNoM
- t-test

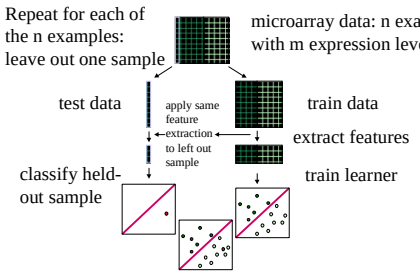
Used data sets:

- Golub et al. (47 ALL and 25 AML leukemia samples)
- Alon et al. (40 Adenocarcinoma and 22 normal samples)
- Notterman et al. (4 Adenoma and 4 Normal tissues)

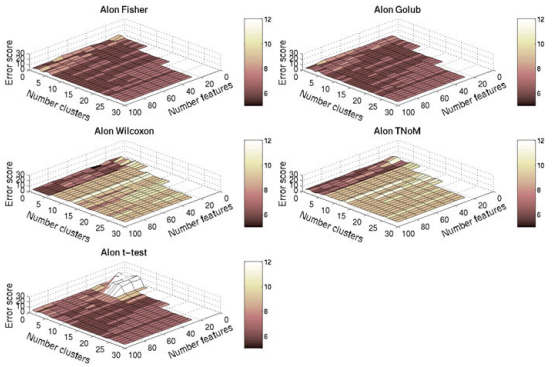
We propose three algorithms:

- clustering of the genes and selecting from each cluster
- clustering and selecting only from "non noise" clusters
- using correlation analysis to select dissimilar genes

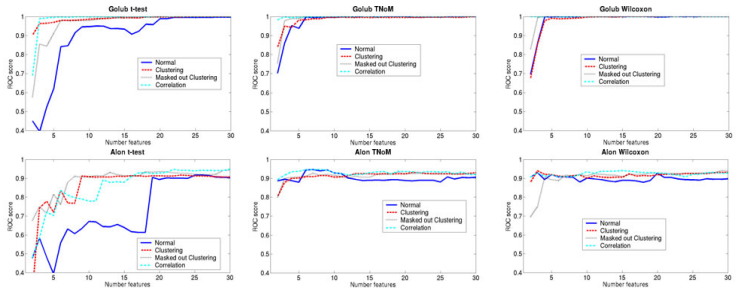
Performance measure:
support vector machines and leave one out cross validation



Results



Comparing different test statistics for Alons data set. The classification error is plotted on the z-axis and also color coded (white means more errors). Fisher, Golub and t-test achieve the best results with 6-25 clusters whereas TNoM and Wilcoxon get by with fewer clusters.



Comparison of the leave one out classification performance of our proposed methods versus the conventional methods (in blue). Plotted is a receiver operator curves (ROC) score (the area under the ROC graph).

Conclusion

In almost all cases, our methods identify sets of genes that are stronger predictors than similarly sized sets found by standard methods. This should be of significant value for diagnostic purposes as well as for guiding further exploration of the underlying biology.

Acknowledgment: Thanks to Bill Noble and Rainer Spang for their helpful suggestions. JJ and RS were supported in part by NIH grant NHGRI/K01-02350. WLR and JJ were supported in part by NSF DBI-9974498.