Slide 1

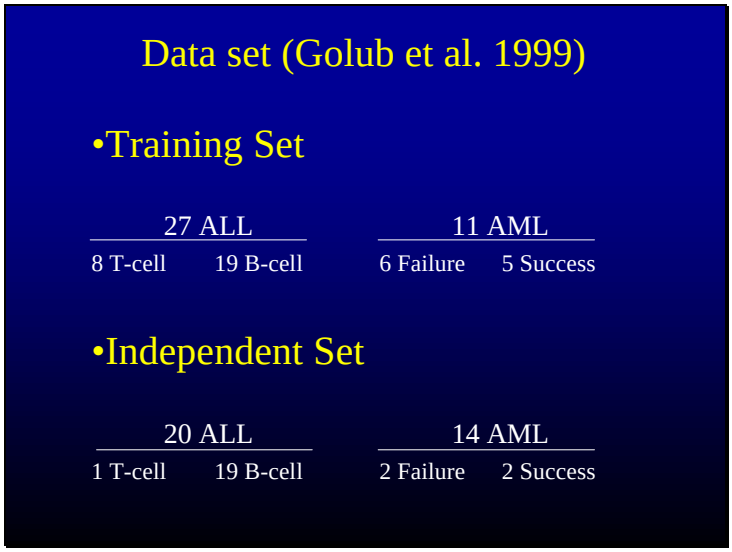
A Ranking Method to Improve Detection of Disease Using Selectively Expressed Genes in Microarray Data

Virginie Aris[1], and Michael Recce[2]

Center for Applied Genomics[1]

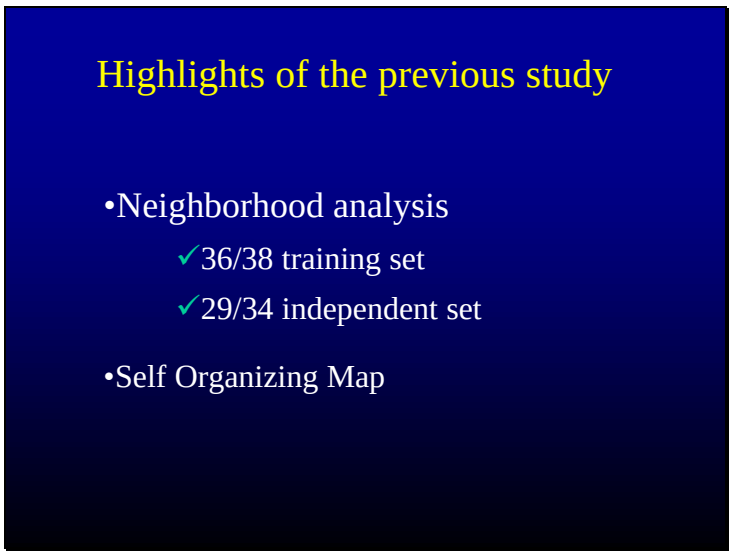
Center for Computational Biology and Bioengineering[2]

Slide 2



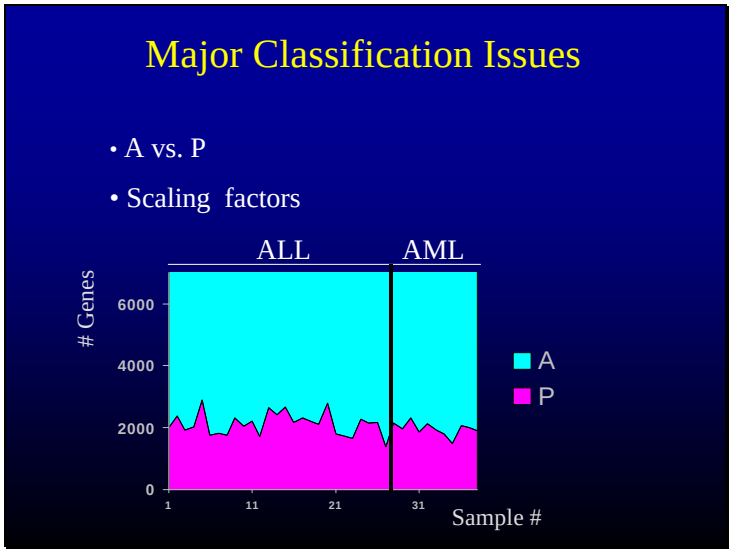
We chose to use the Golub and al. data set.
As a brief summary the training set used to develop a method and a set of classifying parameters, was composed of Bone marrow samples from patients suffering from acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The ALL comported to subtypes: T-cell and B-cell, and on the AML information about the treatment failure or success was recorded. The and the Independent set was use to test that method, and some of its samples were derived from peripheral blood.

Slide 3



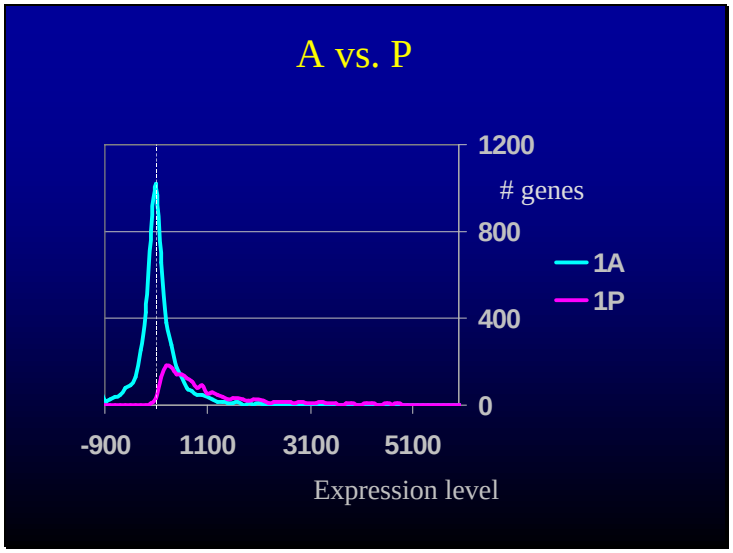
Using the neighborhood analysis they were able to classify 36 of the 38 samples in the training set and 29 of the 34 independent samples.
They also use self organizing map for automatic discovery of the classes.

Slide 4



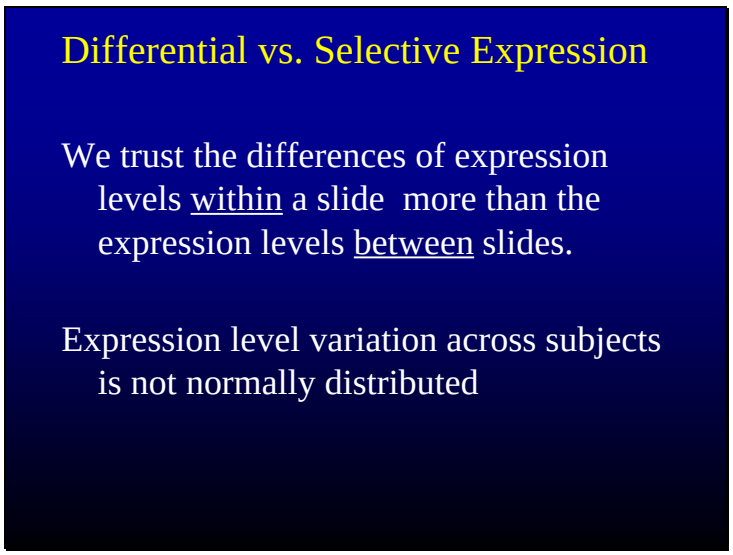
Affymetrix outputs contains a Present (P) or Absent (A) call for each gene.
Can those calls give out interesting information? How shall they be used in an analysis?
Another concern was the normalization factor from slide to slide. Does it really work? How reliable is it?
As we can see on this graph (the samples are on the X-axis and separated into ALL and AML patients, and then the number of genes is on the left) Absent calls are predominant. We can also notice that there is a large variation of the number of genes expressed from sample to sample: 1352 genes for the lowest and 2877 for the highest with an average of approx. 2000.

Slide 5



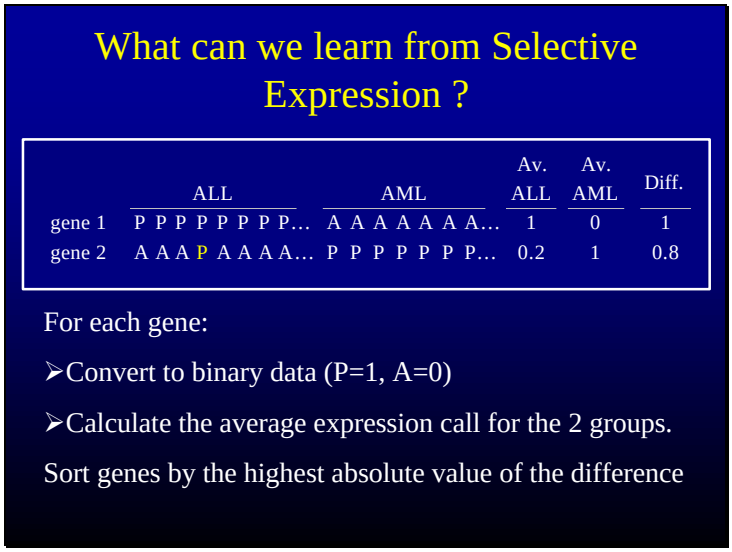
This graph represent the distribution frequency of the Expression levels of the A and P calls for the sample #1.
The expression levels are on the X-axis and the frequency distribution is on the Y axis.
We can see that A has a cusp shape around 0.
P is asymmetric and has a long tail. The two distributions are very different and they overlap.
Any threshold based solely on the expression level will contain a mix of this population which would make them difficult to model.

Slide 6



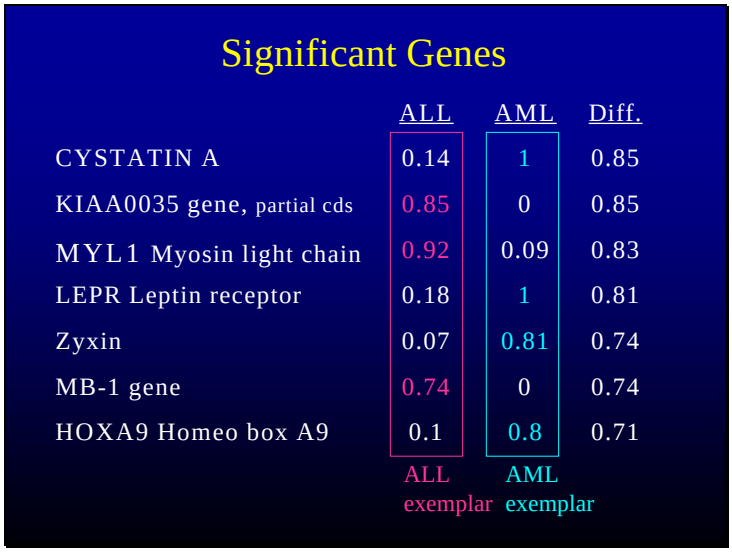
We trust the differences of expression levels within a slide more than the expression levels between the slides.
The second point is that the expression level variation across subjects is not normally distributed

Slide 7



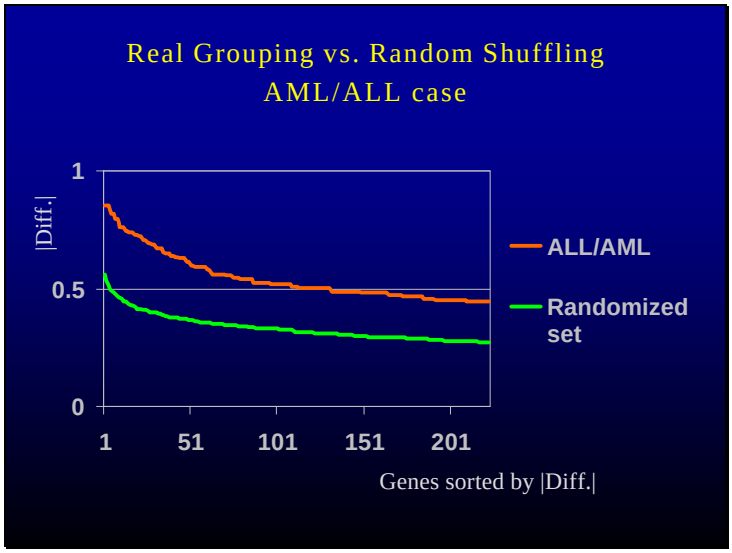
Can we learn something with the presence and absence calls (selective expression)?
So for each gene in each sample we considered only the Present or Absent call.
We looked for genes that were consistently present for a group and absent in the other one.
Converted the calls into binary data, and took the average difference for each group then we took the absolute difference value of those 2 average difference.
We performed this for all 7129 genes and we sorted all the genes according to the highest difference.

Slide 8



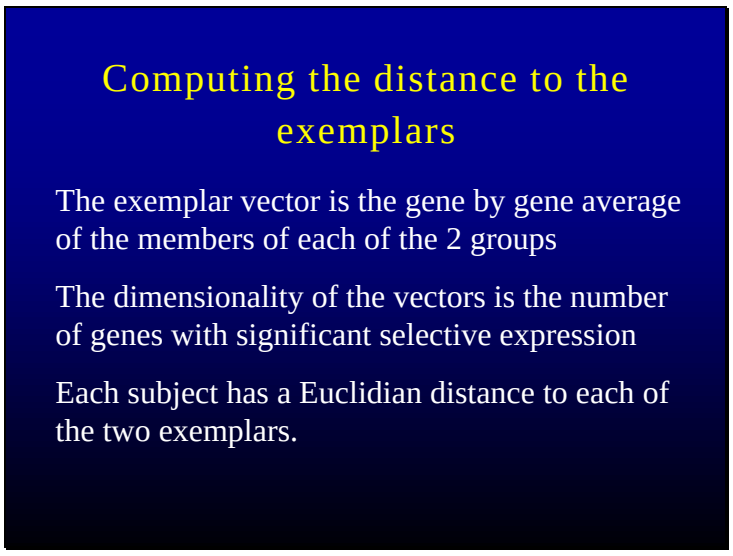
This slide represents part of the genes selected in our method and we can see that some of them were also selected in other studies (Golub et al.).
So the average selective expression value for one group represent sort of the “ideal behavior” of a sample in a group. We call this also an exemplar.
Later on wel compute the distance of the training and independent samples to those 2 exemplars.
The fact that we were selecting some of the same genes was good news but wasn’t enough to validate the method on its own.

Slide 9



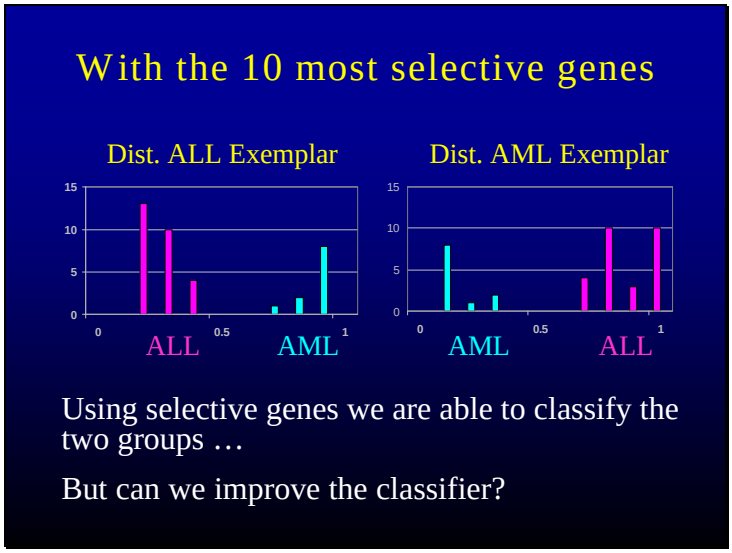
We performed a random shuffling of the samples within the categories. On the Y-axis we have the absolute difference of the average of 1 and 0 for each group and on the X-axis we have the number of genes (sorted by their higher absolute value difference).
The AML/ALL difference curve is 6 standard deviation above the random difference curve.

Slide 10



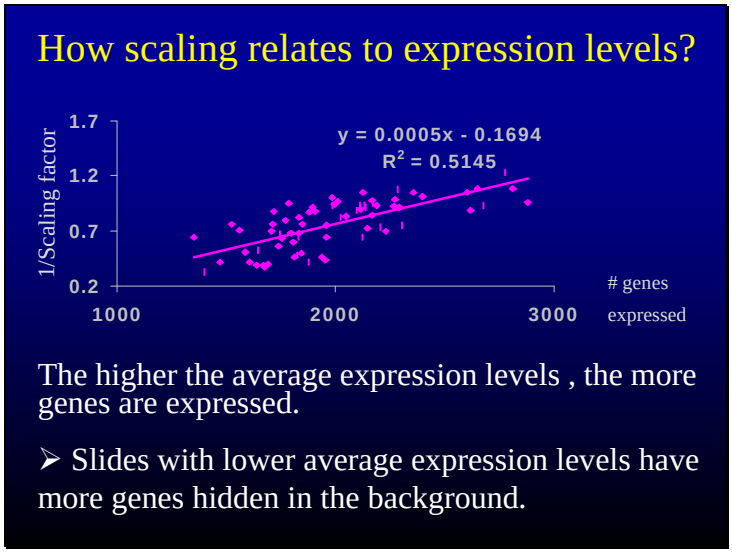
We then went on computing the distance to the AML and ALL exemplars.
The dimensionality of the exemplar vector is the number of genes we want to include to discriminate between the two groups (10, 20, 30, 50, 100).
We take the distance for each subject to the exemplar or “Ideal Case”.

Slide 11



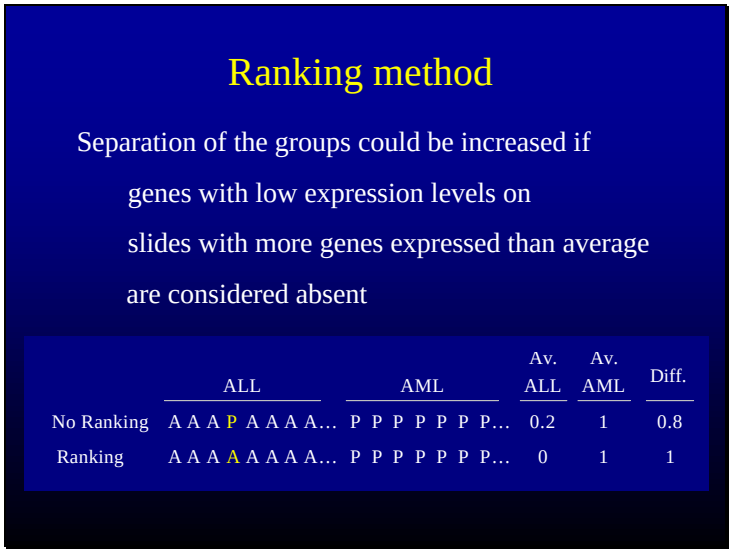
With the ten most selective genes we obtain those 2 graphs. On the X-axis is the distance to the ALL exemplar on the left and AML exemplar on the right. On the Y-axis we have the frequency distribution of the samples. In pink we have the ALL training samples and in blue the AML ones.
We can see that the ALL samples are closer to the ALL exemplar that the AML samples and Vice-Versa.
So using selectively expressed genes we are able to classify the training data.
But can we do better?

Slide 12



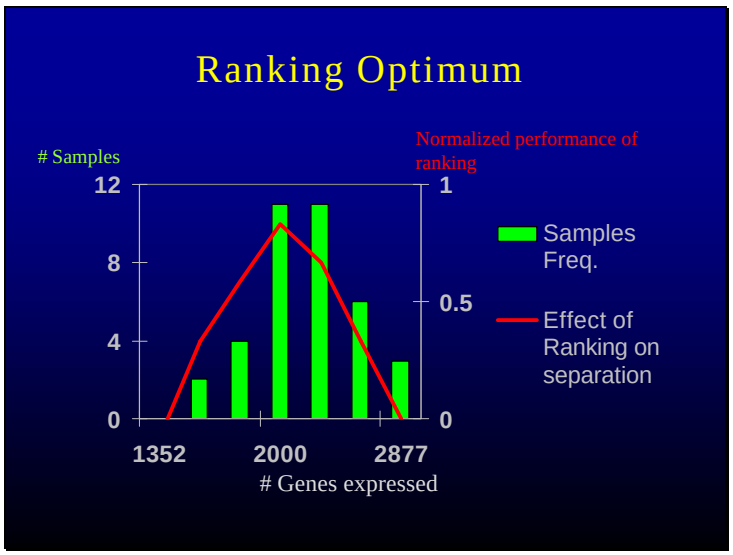
A few slides ago I mentioned the difference between the number of genes expressed between samples. The scaling factor is based upon the average expression level. There seems to be a quite straight forward correlation between the average expression level and the number of genes expressed.
This implies that slides with lower average expression levels have more genes hidden in the background.

Slide 13



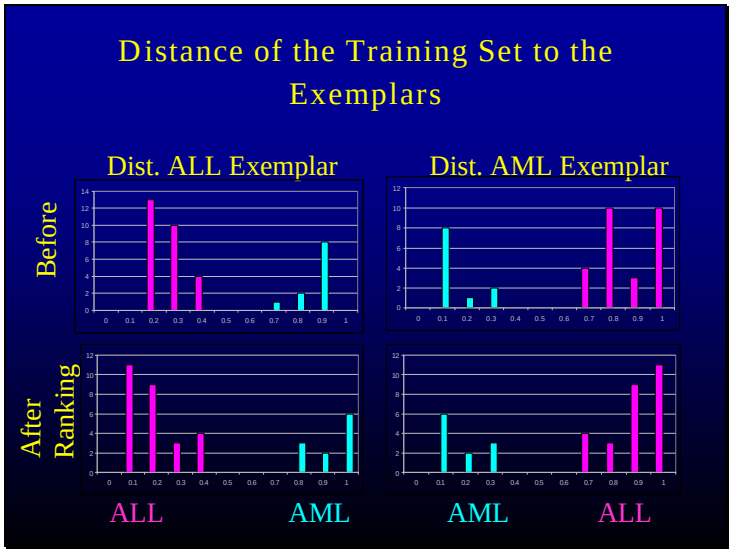
So instead of scaling up, we scaled the distribution down by turning off the genes that are low expressed.
In other words, we're going to take the expression value within a slide (that we trust) and rank them from highest expression level to lowest, and we set to 0 the later genes on slides that have more than average number of genes expressed. The net effect of this for a sample that has more P values, might set the low expressors to 0 (A) and make the gene more selective.

Slide 14



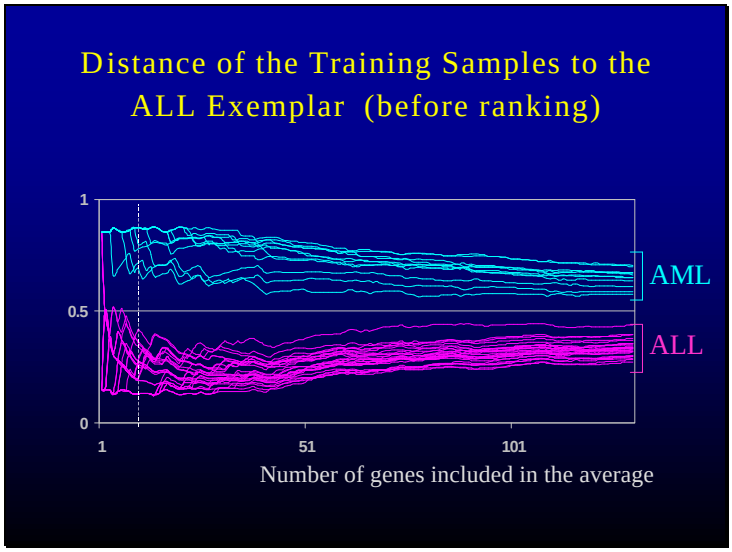
In green we have the distribution frequency of the samples by their number of genes expressed. We designed a metric to find the optimum number of genes to keep in order to improve the separation, in this case we found the optimum to be 2000.

Slide 15



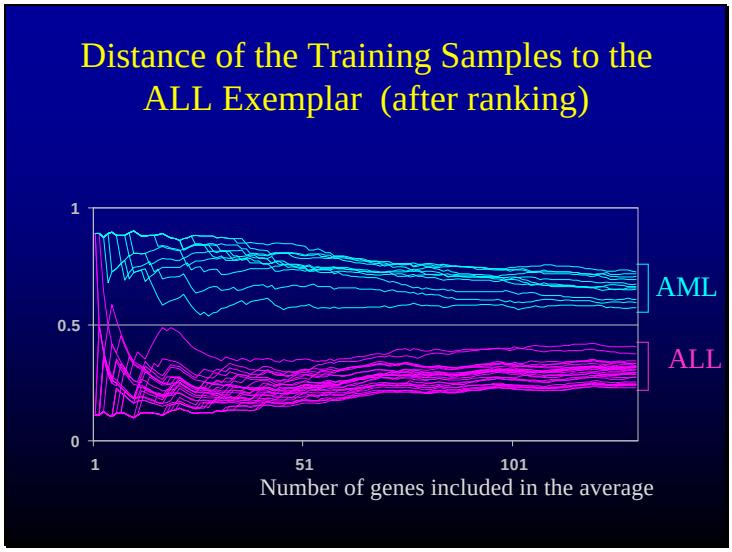
This graph is similar to the one I've shown you before and we can see that with ranking, we move the ALL and AML clusters apart.
We have a good separation with 10 genes on the training set. Our next question was: How does this hold if we change the number of genes selected for the separation?
So we took the first graph and expended it by increasing its dimensionality by changing the number of genes in the exemplars.

Slide 16



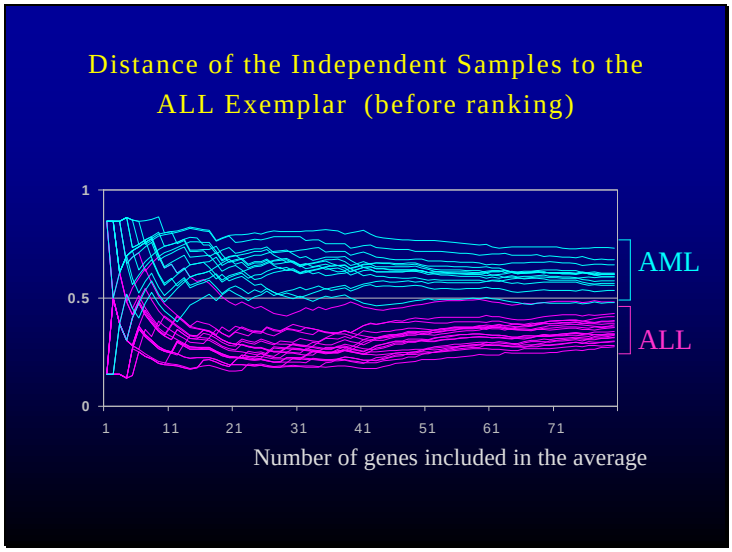
On the Y-axis we have the distance to the ALL exemplar, and on the X-axis we have the number of genes taken into account.
We have a very good separation of the two groups with ALL being closer to the ALL exemplar.
It's easily separated with the threshold 0.5 .
Because we equally weighted the genes used in the average the 2 distribution eventually converge as we add less significant genes.

Slide 17



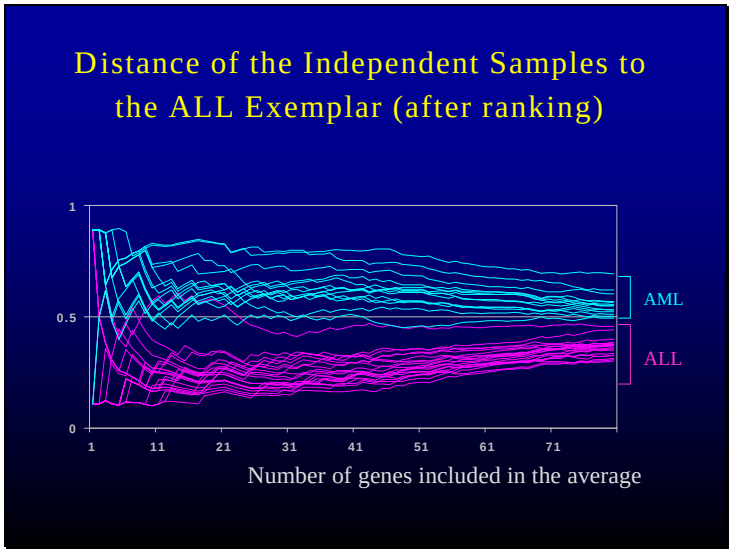
With ranking on the training set we obtain a comparably good separation.

Slide 18



This is the distance of the independent samples to the ALL exemplar.
We see that both groups are well separated except for one sample (66).

Slide 19



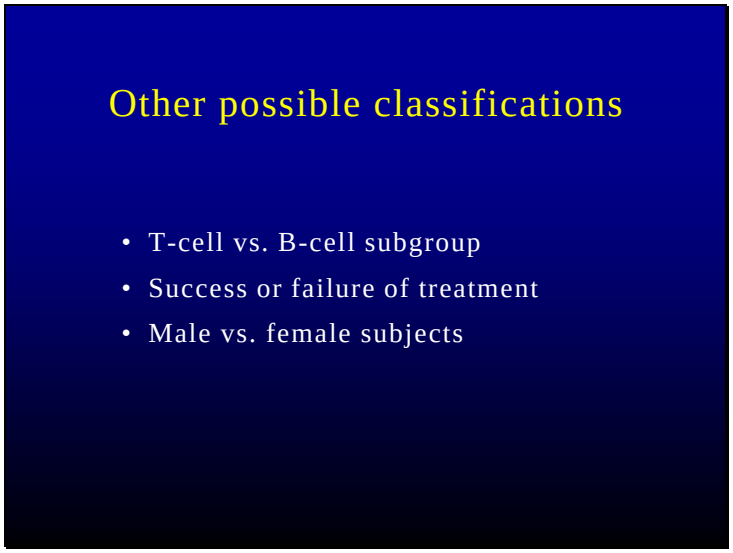
After ranking we have a tighter clustering of the ALL samples.

Slide 20

Results from ALL/AML classification

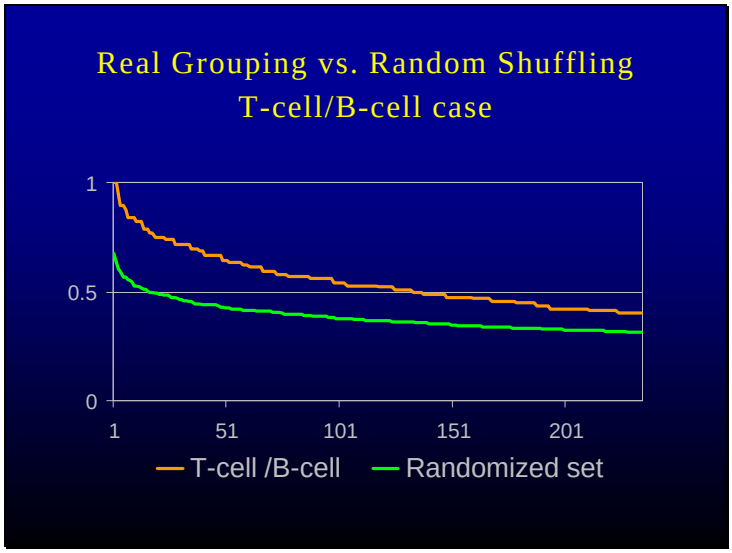
- We obtained a perfect separation of the training set with and without ranking.
- We were able to classify 33 out of 34 independent samples.

Slide 21



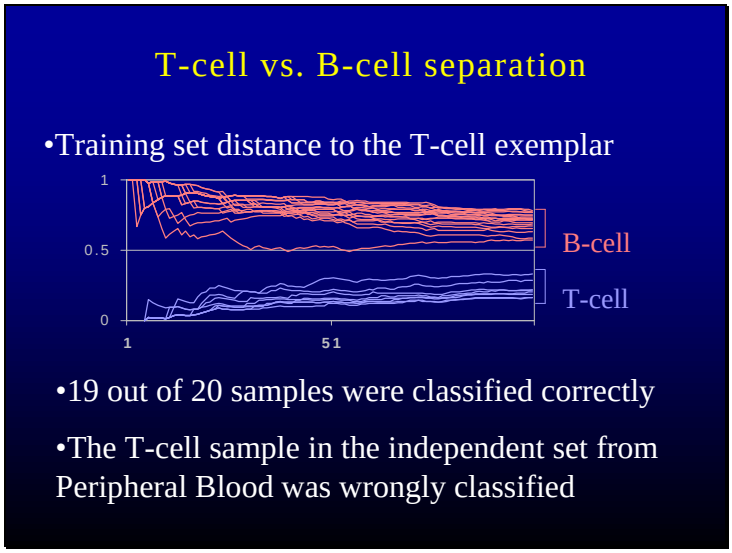
We then wanted to look at other possible classifications: T-cell vs. B-cell, success vs. failure and Male vs. female.

Slide 22



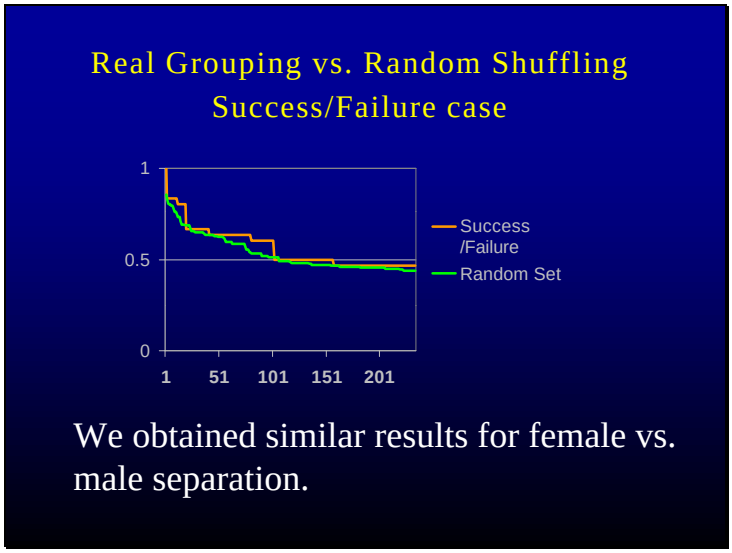
In the T-cell vs. B-cell case we also have a very good separation of the absolute value of the difference of sorted genes of the real grouping compared to a random shuffled grouping.

Slide 23



We obtained a good separation of the training samples and were able to classify 19 out of the 20 independent samples.

Slide 24



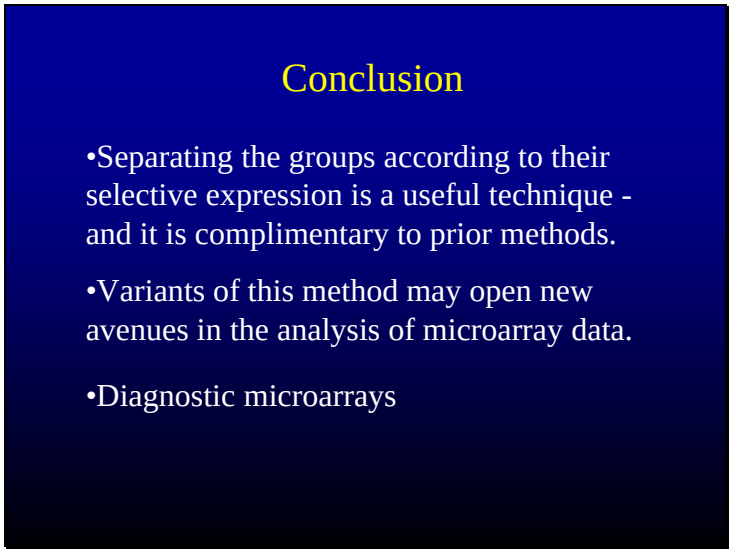
In the success vs. failure and in the female vs. male classification we did not obtained significant differences between the real grouping and randomized groups.

Slide 25

Results

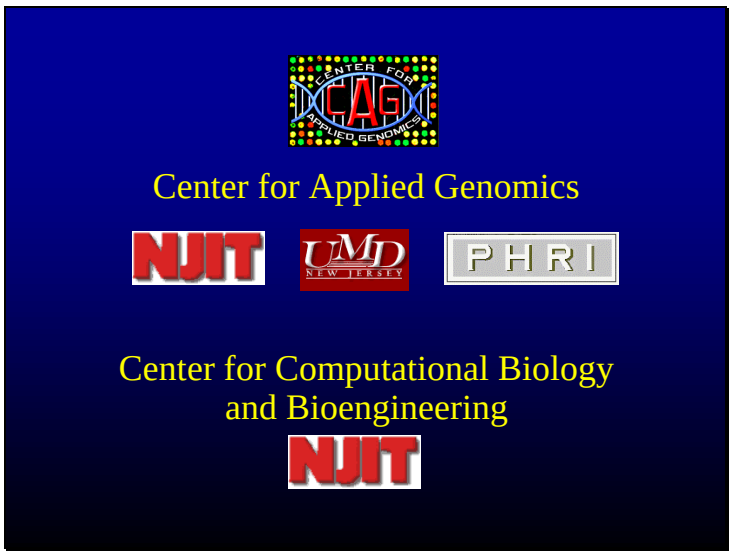
- T-cell and B-cell subgroups were well classified by this method.
- Absence of distinction in the male vs. female and success vs. failure of treatment was identified
- Ranking makes suggestive improvements, that warrant further investigation.

Slide 26



Separating the groups according to their selective expression is useful.
This approach is orthogonal to other approaches and complementary to them.
The combination of those approaches could open new avenues in analyzing microarray data.
This method is simple, robust and easy to use to develop a diagnostic microarray since it would contain redundant strongly expressed robust genes.

Slide 27



I am a graduate student at the Center for Applied Genomics which is part of the following organizations: the New jersey Institute of Technology, the University of Medecine and Dentistry of New Jersey and the Public Health Research Institute. And my advisor Dr. Recce is at the Center for Computational Biology and Bioengineering from NJIT.