Classification and Outlier Detection of Microarray Data

Raymond Wan
rwan@kuicr.kyoto-u.ac.jp

Bioinformatics Centre, Kyoto University, Japan

August 21, 2007

DNA Microarrays

Decision Trees

DT - Results

Outlier Detection

References

DNA Microarrays

Microarrays are high-throughput technologies that permit expression levels of thousands of genes to be analyzed simultaneously under a given set of conditions.

Through microarrays, one can identify the subset of genes that are expressed or repressed under conditions such as cancer or environmental changes like heat.

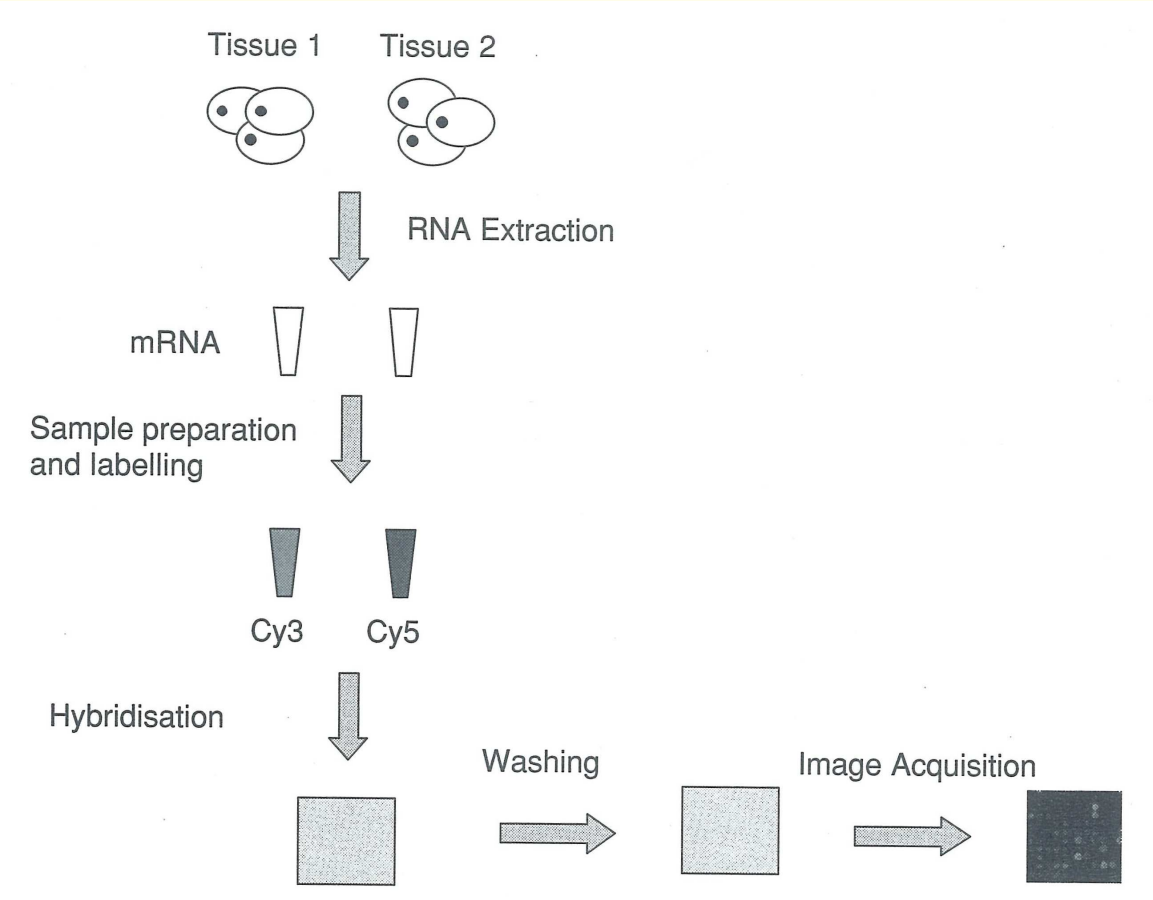


Figure 1: Two-channel microarray

Source: Stekel [2003, pg. 10].

...To Dry Lab

After the last step (image acquisition), the data set is represented as a two-dimensional table of expression levels whose dimensions are typically thousands of genes (up 20,000 or more) by a handful of conditions.

After possible further normalisation by the experimenter, analysis of the data begins.

Single vs Two-Channel

Two main categories of microarrays are:

Two-channel

- the level of hybridisation of DNA samples from two conditions (i.e., cancerous vs healthy tissue) with the probes on the slide are compared
- an expression level is the relative difference between them

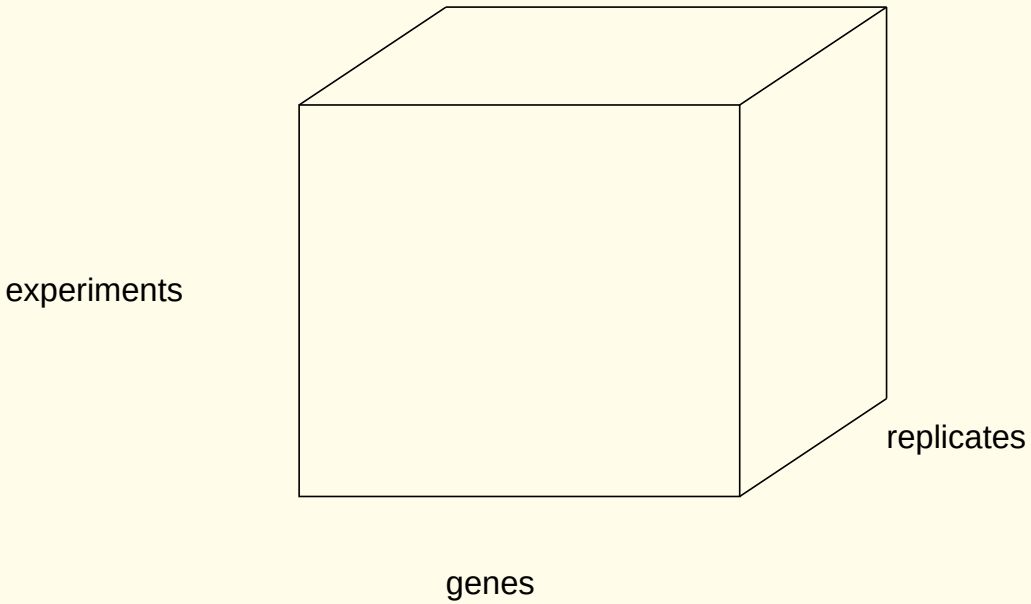
Single-channel

- a single condition's level of gene expression is reported on the slide
- two slides required to compare two conditions

We view a microarray data set $\mathcal{D}$ as:

	Gene (Attribute, Feature)							Disease (Class)
Sample (Example)	A_1	A_2	A_3	A_4	A_5	$\ldots$	A_n	
S_1								tumor
S_2								normal
S_3								tumor
S_4								tumor
S_5								normal
...								tumor
S_m								normal

Later, when we refer to replicates, we mean:

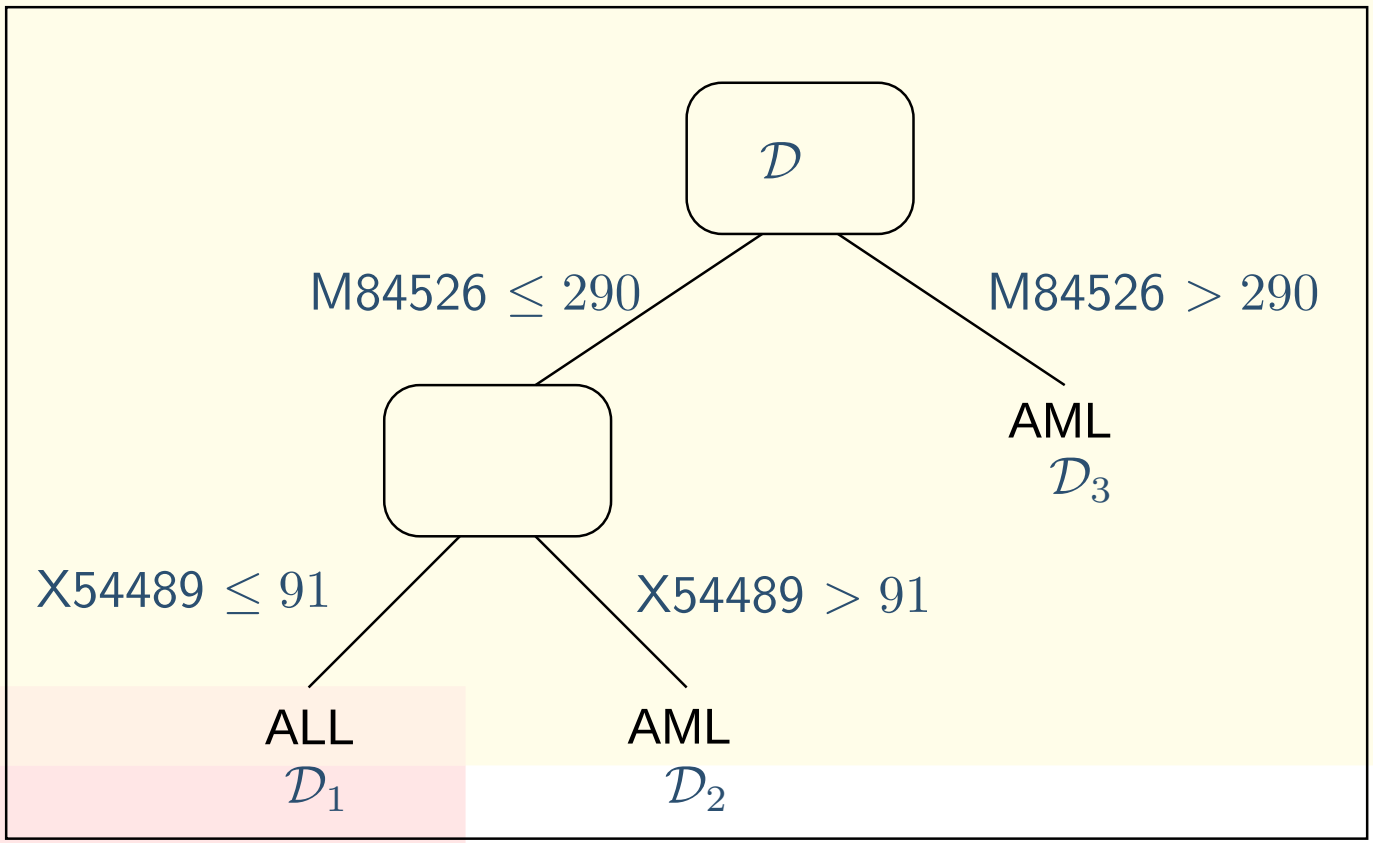


Two separate problems are considered in this talk:

1. The application of decision trees to microarray classification.
   - We show that its accuracy can be improved by augmenting two Gaussian-distribution dependant splitting criteria.
2. Current work on the outlier detection of entire microarrays.
   - Initially motivated by the existance of large repositories of microarray data.

Decision Trees

Decision trees is a supervised machine learning technique used for classification (i.e., determining whether a sample is "tumor" or "normal"). Like other supervised machine learning techniques, a portion of the data is used for training, and the remainder is kept for testing.



It is known that decision trees perform worse than other algorithms such as support vector machines (SVM) for numerical data such as microarray data.

However, they possess other traits that make them useful...

Comparison of Machine Learning Algorithms

Characteristic	DT	SVM	Neural Nets	k-NN
Data of mixed type	●	○	○	○
Missing values	●	○	○	●
Robustness to outliers	●	○	○	●
Insensitive to monotone transformations	●	○	○	○
Scalability (large N)	●	○	○	○
Dealing with irrelevant inputs	●	○	○	○
Linear combinations of features	○	●	●	◐
Interpretability	◐	○	○	○
Predictive power	○	●	●	●

Table 1: Good: ●; Fair: ◐; Poor: ○

(Source: "Decision Tree Methods in Pharmaceutical Research" [Blower and Cross, 2006], adapted from Hastie et al. [2001].)

DT with Small Samples

We realized that for data sets (like microarrays) which have:

- small number of examples (patient samples)
- large number of features (genes)

prediction performance is noticably poor.

Upon closer inspection of implementations such as C4.5, we realized (and confirmed) that its intended use was for large data sets.

C4.5 and Gain Ratio

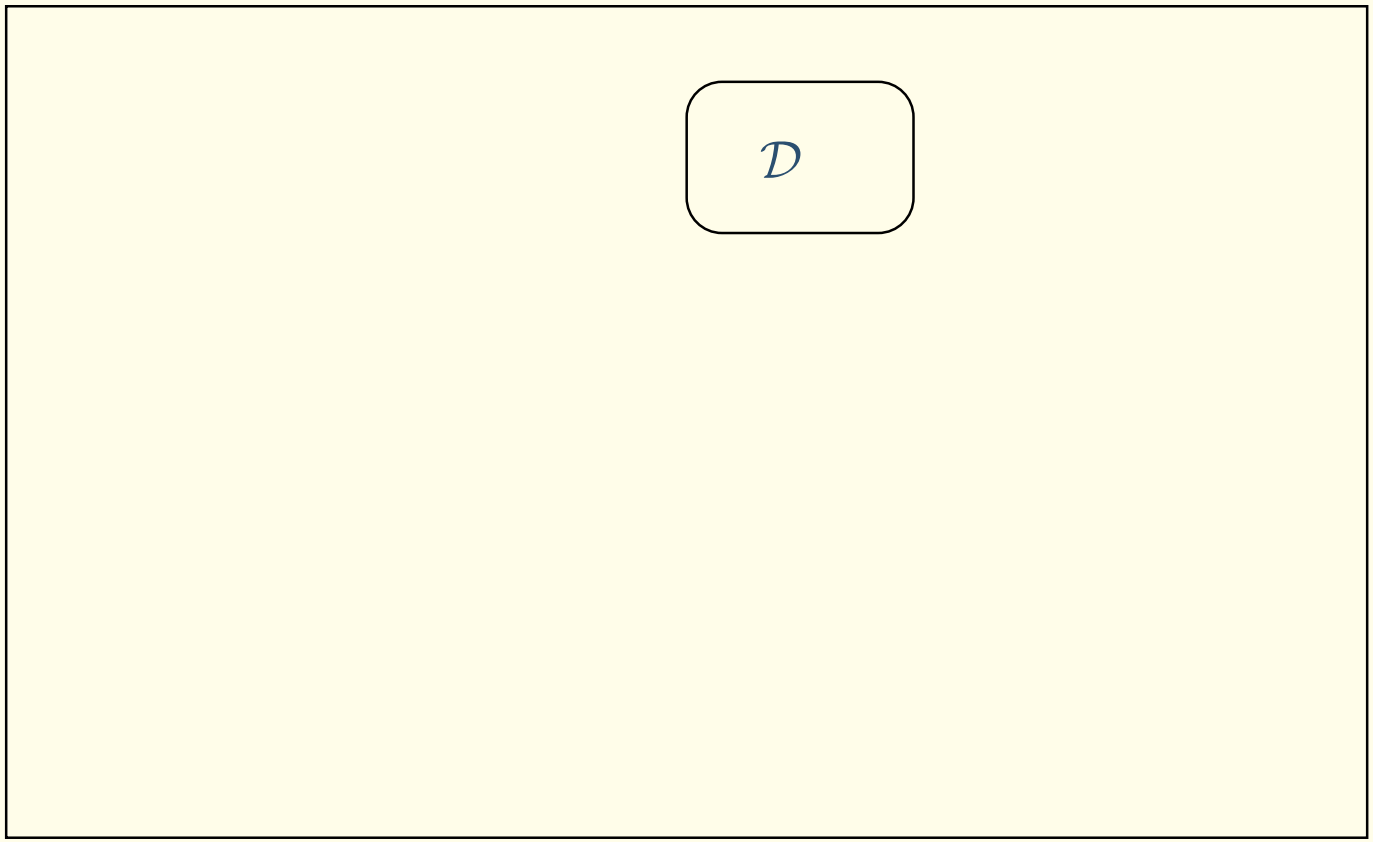
Our focus is on C4.5 Release 8 [Quinlan, 1996]. Other implementations exist (such as J4.8 of WEKA [Witten and Frank, 2005]) and we confirmed their implementations were the same.

C4.5 uses the gain ratio to select the splitting attribute for each node.

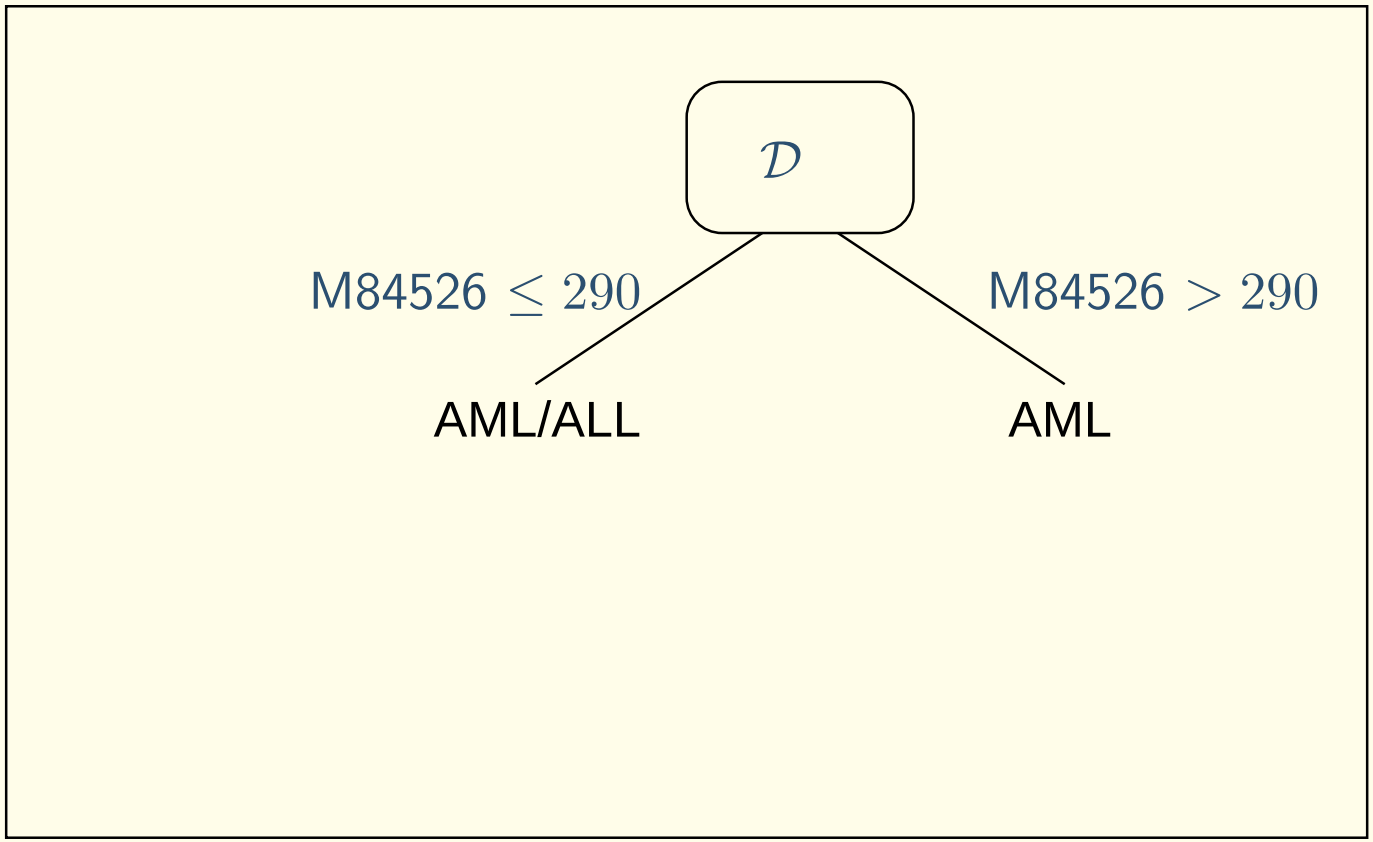
- Each attribute is examined one-by-one.
- For continuous values, each attribute's values are sorted and every possible cut point is tested.

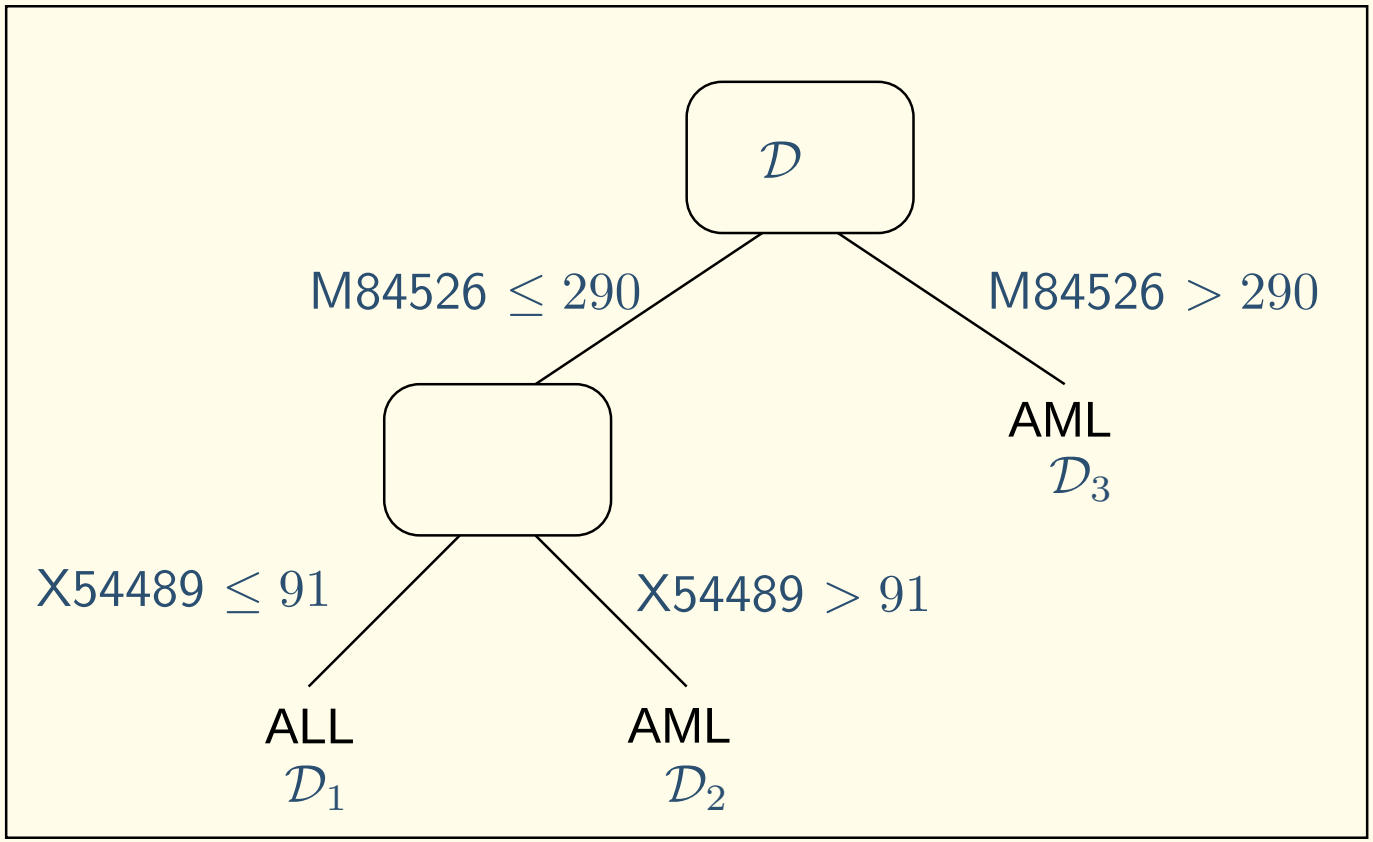
The attribute associated with the cut point that gives the best gain ratio is chosen. This process is repeated until $\mathcal{D}$ is divided into numerous subsets which are uniform in class.

Example



Example





Each node represents an attribute and (for continuous values) a cut point.

Patient ID	Gene A	Gene B	Gene C	Class
1	10	10	10	tumor
2	20	20	15	tumor
3	40	30	20	tumor
4	30	40	400	normal
5	50	50	500	normal
6	60	60	600	normal

In this example, either Gene B or C would give the best split in class. That is, either:

- $B \leq 30 \rightarrow$ tumor
- $C \leq 20 \rightarrow$ tumor

Problem: If the number of genes (1,000s) is far larger than the number of samples (10 or 20), then it is very likely that multiple genes will yield the same gain ratio. Decision tree implementations such as C4.5 break such ties by arbitrarily selecting the "first" attribute.

Idea: Augment the gain ratio with Gaussian distribution-dependent criteria that focus on the attributes instead of the classes. Our aim is to improve prediction accuracy over the original method (though, improving over SVM would be nice...).

Assume: Each gene is independent of each other when splitting and each gene A_k is made up of two normal distributions, one for each class.

There is extensive work in computer science on the addition of various statistical methods to decision trees. Our motivation, here, is on the application to high-dimensional biological data sets, like microarrays.

More specific to microarrays:

- Zhang et al. [2001] used decision trees to classify microarray data. The splitting criterion was similar to the gain ratio but was further refined using cross-validation.
- They later built a forest of decision trees and took the top two levels of each to form a fingerprint for the tumor type [Zhang et al., 2003].
- The RankGene software [Su et al., 2003] includes several algorithms, including the gain ratio (information gain) and t-test to rank genes.

- Combine the gain ratio with the Student's t-test and the Kullback-Leibler divergence.
- While the gain ratio focuses on the class distribution, the other two look at the expression levels' distribution.

	A_1	A_2	...	A_k	...	A_n	Class
$\mathcal{N}_1(\mu_1, \sigma_1)$	v_{11}	v_{12}	...	v_{1k}	...	v_{1n}	tumor
	v_{21}	v_{22}	...	v_{2k}	...	v_{2n}	tumor
	v_{31}	v_{32}	...	v_{3k}	...	v_{3n}	tumor
$\mathcal{N}_2(\mu_2, \sigma_2)$	v_{41}	v_{42}	...	v_{4k}	...	v_{4n}	normal
	v_{51}	v_{52}	...	v_{5k}	...	v_{5n}	normal
	v_{61}	v_{62}	...	v_{6k}	...	v_{6n}	normal

Each gene A_k is assigned three scores: $gain_s$, $ttest_s$, and KL_s.

No changes to the gain ratio; each attribute's gain ratio is also its score.

Several forms of Student's t-test exist. We employed a two-tailed, two sample t-test for unequal variance. The t-test score is simply its t-statistic.

The Kullback-Leibler (KL) divergence calculates the difference between two distributions.

$$D(p||q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx \,. \tag{1}$$

If both are normally distributed, their KL divergence can be simplified to:

$$D(p||q) = \frac{1}{2\sigma_2^2} (\mu_1 - \mu_2)^2 + \frac{(\sigma_1 - \sigma_2)(\sigma_1 + \sigma_2)}{2\sigma_2^2} + \log \left(\frac{\sigma_2}{\sigma_1} \right) . \tag{2}$$

Since the KL divergence is non-symmetric, we calculate KL_s as half of $D(p||q) + D(q||p)$ [Jeffreys, 1946].

Since each score has different ranges, we convert each to ranks: $gain_r$, $ttest_r$, and KL_r. The ranks are assigned from n down to 1. Tied scores are assigned the same rank.

After ranking, we introduce three parameters to weight the ranks ($\{\alpha_1, \alpha_2, \alpha_3\}$) to derive a final score. The highest score becomes the splitting attribute.

$$\text{score}(A_k) = \alpha_1 \times \text{gain}_r + \alpha_2 \times \text{ttest}_r + \alpha_3 \times \text{KL}_r$$
$$\text{such that } \sum_{i=1}^{3} \alpha_i = 1. \tag{3}$$

We evaluate several combinations of these parameters.

DT - Results

Data Sets

Name	Classes		Number of samples				Number of genes
	Class 1	Class 2	Proportion			Total	
Colon	Tumor	Normal	40	:	22	(62)	2,000
Leukemia	ALL	AML	47	:	25	(72)	7,129
Lung	ADCA	MPM	150	:	31	(181)	12,533
CNS	Success	Fail	39	:	21	(60)	7,129
Multiple	Tumor	Normal	190	:	90	(280)	16,063
Lymph	DLBCL	FL	58	:	19	(77)	7,129
Prostate	Tumor	Normal	52	:	50	(102)	12,600

Inverse Cross-Validation

	$\alpha_1, \alpha_2, \alpha_3$	Colon	Leukemia	Lung	CNS	Multiple	Lymph	Prostate	Avg
10	$\{1,0,0\}$	59.5	58.8	75.8	52.6	59.3	65.5	51.6	60.5
	$\{0,1,0\}$	+1.4	+10.2	+4.6	+2.2	+2.0	+3.7	+12.1	+5.2
		(1.81)	(13.20)	(9.27)	(3.94)	(8.79)	(5.27)	(23.26)	
	$\{0,0,1\}$	+0.8	+6.1	+6.9	+4.9	-0.7	+3.1	+3.4	+3.5
		(0.89)	(10.16)	(17.50)	(6.09)	(-2.94)	(4.04)	(6.92)	
	$\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$	+2.3	+11.4	+5.3	+3.4	+1.5	+6.6	+7.6	+5.4
		(3.23)	(16.26)	(11.94)	(5.43)	(6.00)	(9.15)	(16.87)	
	$\{0, \frac{1}{2}, \frac{1}{2}\}$	+2.0	+11.7	+5.6	+3.6	+1.7	+7.0	+9.4	+5.8
		(2.74)	(15.09)	(13.70)	(5.60)	(6.76)	(9.59)	(17.84)	
	Train:Test	10:52	10:62	10:171	10:50	10:270	10:67	10:92	
20	$\{1,0,0\}$	68.2	82.7	81.1	55.2	63.3	74.4	68.0	70.4
	$\{0,1,0\}$	-2.0	-3.0	+5.4	+0.5	+0.5	+3.8	+5.8	+1.6
		(-2.33)	(-3.11)	(11.92)	(0.65)	(1.55)	(5.28)	(7.55)	
	$\{0,0,1\}$	-2.0	-6.0	+3.9	+1.9	-1.8	-2.2	-7.4	-1.9
		(-1.87)	(-5.70)	(9.13)	(2.29)	(-6.24)	(-2.58)	(-8.59)	
	$\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$	+0.9	+2.5	+6.8	+0.6	+1.0	+4.3	+5.4	+3.1
		(0.93)	(2.51)	(15.72)	(0.68)	(3.79)	(5.96)	(7.97)	
	$\{0, \frac{1}{2}, \frac{1}{2}\}$	-0.1	+1.2	+7.1	+0.8	+1.1	+4.3	+5.3	+2.8
		(-0.13)	(1.32)	(16.88)	(0.83)	(3.85)	(5.38)	(8.04)	
	Train:Test	20:42	20:52	20:161	20:40	20:260	20:57	20:82	

- 50 repetitions
- Red indicates significance at the 99% confidence interval (the two-tailed t-statistic is 2.6800).

Cross-Validation

	$\alpha_1, \alpha_2, \alpha_3$	Colon	Leukemia	Lung	CNS	Multiple	Lymph	Prostate	Avg
5	$\{1,0,0\}$	76.5	84.8	93.4	56.2	78.5	79.2	82.4	78.7
	$\{0,1,0\}$	-2.5	+2.5	+2.1	+0.8	-1.8	+11.4	+1.9	+2.1
		(-2.42)	(3.96)	(7.78)	(0.66)	(-3.96)	(15.43)	(2.90)	
	$\{0,0,1\}$	-6.6	+2.9	-1.0	+2.5	-3.3	-5.1	-10.3	-3.0
		(-7.12)	(4.64)	(-2.70)	(2.07)	(-7.26)	(-5.91)	(-13.72)	
	$\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\}$	-3.6	+3.6	+1.9	+4.0	-0.1	+2.5	-0.4	+1.1
		(-4.58)	(5.95)	(8.21)	(3.63)	(-0.13)	(3.62)	(-0.73)	
	$\{0,\frac{1}{2},\frac{1}{2}\}$	-3.6	+6.0	+3.6	+2.0	-0.2	+3.4	+1.2	+1.8
		(-4.73)	(10.41)	(14.41)	(1.99)	(-0.51)	(5.25)	(1.95)	
	Train:Test	50:12	58:14	145:36	48:12	224:56	62:15	82:20	
10	$\{1,0,0\}$	77.5	80.4	93.7	58.3	78.7	81.1	83.3	79.0
	$\{0,1,0\}$	+0.6	+6.4	+2.2	-2.8	-2.0	+12.1	+0.9	+2.5
		(0.68)	(10.20)	(9.44)	(-2.69)	(-5.14)	(21.72)	(1.52)	
	$\{0,0,1\}$	-5.2	+10.1	-1.2	+3.5	-3.2	-6.0	-11.9	-2.0
		(-6.72)	(16.97)	(-4.03)	(3.73)	(-6.94)	(-9.59)	(-18.23)	
	$\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\}$	-4.9	+7.8	+1.7	+1.5	-0.3	+0.3	-1.7	+0.6
		(-6.38)	(12.69)	(6.81)	(1.39)	(-0.88)	(0.40)	(-3.52)	
	$\{0,\frac{1}{2},\frac{1}{2}\}$	-2.7	+11.3	+3.7	-0.1	+0.3	+1.5	+0.6	+2.1
		(-3.89)	(23.32)	(16.14)	(-0.07)	(0.76)	(2.55)	(1.14)	
	Train:Test	56:6	65:7	163:18	54:6	252:28	69:8	92:10	

	Training size / Folds	$\{1,0,0\}$	$\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\}$	$\{0,\frac{1}{2},\frac{1}{2}\}$
Inverse CV	10	60.5	+5.4	+5.8
	20	70.4	+3.1	+2.8
CV	5	78.7	+1.1	+1.8
	10	79.0	+0.6	+2.1

	Training size / Folds	$\{1,0,0\}$	J4.8	Naive Bayes	SVM
Inverse CV	10	60.5	+0.1	+7.0	+12.2
	20	70.4	+0.2	+2.5	+10.5
CV	5	78.7	+0.0	-1.7	+10.5
	10	79.0	+0.3	-2.1	+10.6

- Our modification to C4.5 has improved prediction accuracy over the original C4.5 .
- As expected, less improvement when the training set is larger (has more examples).
- Prediction accuracy still poorer than SVM.

Our method assumes that genes are independent, which is not true. Ideally, we should split on groups of attributes (genes) instead of single ones. Thus, some preliminary gene clustering is needed.

Some strengths in decision trees have not been exploited:

- Presenting a list of important genes.
- Incorporating additional information which could be continuous (patient age) or categorical (gender, race, family history, tumor type, etc.).

Combining our method with forests of trees or consensus trees for microarray data.

$\Longrightarrow$ Ideally, we hope to reduce the size of the forests and still arrive at the same conclusion (prediction accuracy).

Outlier Detection

Microarray Repositories

Microarray data have been around for about 15-20 years. Since then, many repositories have been built to store them NCBI's Gene Expression Omnibus (GEO) has 3,635 platforms and 163,356 experiments.

The meta-analysis of microarray data aims to combine results from multiple laboratories that are using the same microarray platform.

For example, Stevens and Doerge [2007] showed how Affymetrix (single channel) microarray results from multiple laboratories can be combined and that others have combined p-values from multiple microarrays to improve the significance levels of the list of expressed genes.

Outlier Detection

In microarrays, outliers can be:

- expression levels which differ from other values on the same microarray slide, or
- microarray slides which differ from other slides (i.e., replicates that did not hybridised correctly).

After identifying outliers, further investigation is needed to determine if they are:

- erroneous or
- interesting.

In knowledge discovery and data mining (KDD), distance-based outliers examine the distance between a record and its nearest neighbors. Bay and Schwabacher [2003] lists several "popular definitions" in the literature:

1. Outliers are the examples for which there are fewer than p other examples within distance d.
2. Outliers are the top n examples whose distance to the kth nearest neighbor is greatest.
3. Outliers are the top n examples whose average distance to the k nearest neighbors is greatest.

So, examples in a data set are compared against all examples in the same data set (with pruning to improve efficiency).

Assist (wet lab) scientists by identifying microarray slides (that is, not individual expression levels) which might be problematic, by using data already found in microarray repositories as a guide. A microarray slide is scored based on the number of expression levels that appear to be outliers.

Expression levels of a gene are not compared to every other gene, but only those indicated by other experiments in the repository using distance-based outlier methods as the basis.

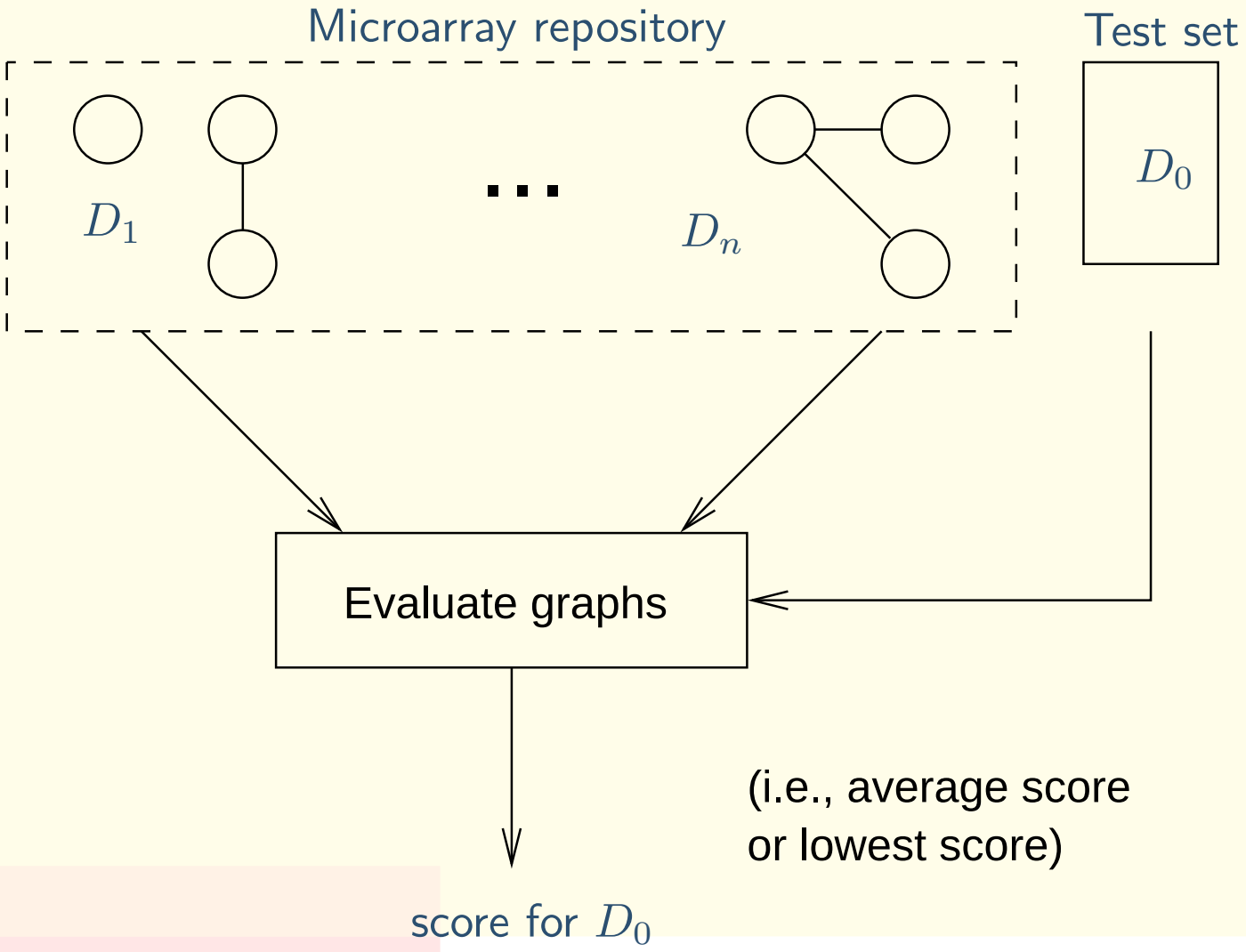
From: "Outliers are the examples for which there are fewer than p other examples within distance d."

For each data set in the repository, we assign a gene neighbourhood to each gene, based on the Euclidean distance between genes and some threshold $d_{\text{repository}}$.

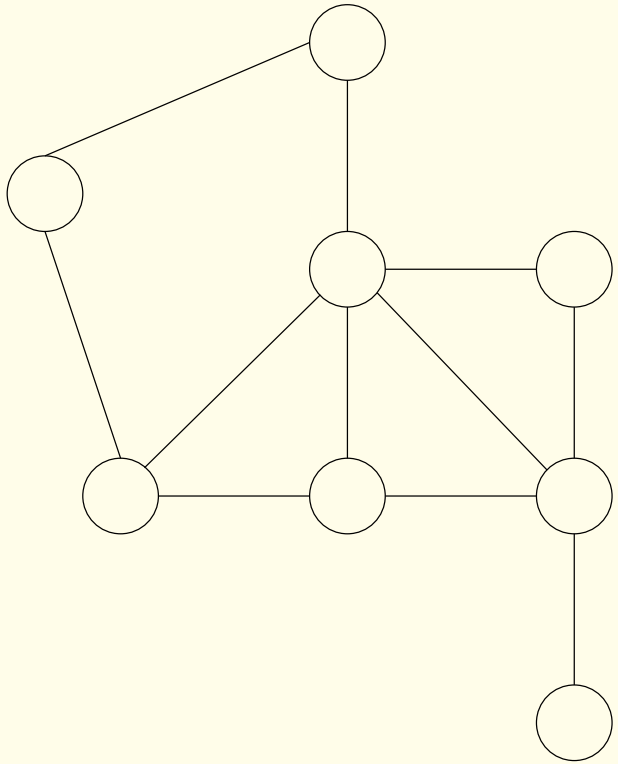
To: "Outliers are genes for which there are fewer than $p\%$ of its neighbours within a distance d_{test}." (where $d_{\text{repository}} = d_{\text{test}}$)

Informally, "if genes A and B had similar expression levels in an experiment from the repository, they should have similar values in the current experiment".

If we represent neighbourhoods as an undirected graph:

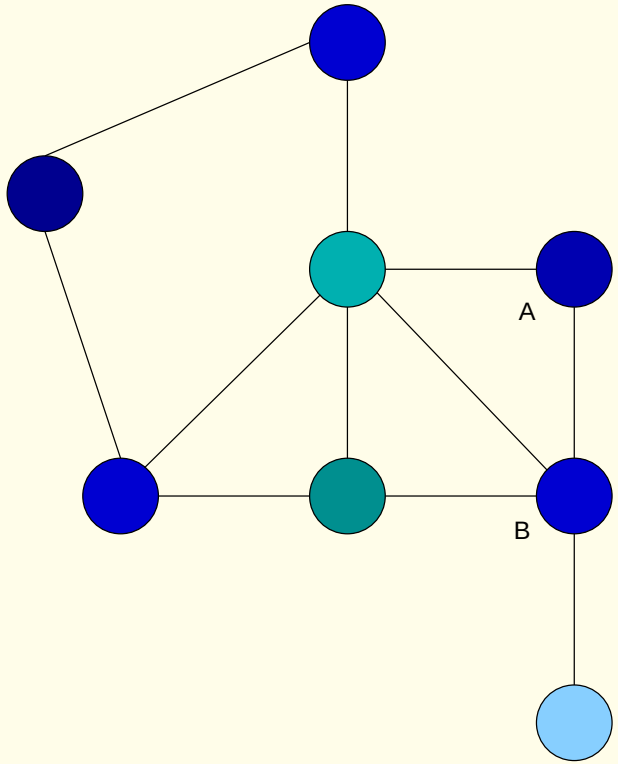


For each data set D_i, where i is size of the repository:



This graph has nodes as genes and edges indicate similarity in the repository data set using Euclidean distance.

Now, if the test set D_0 is inserted into this graph:



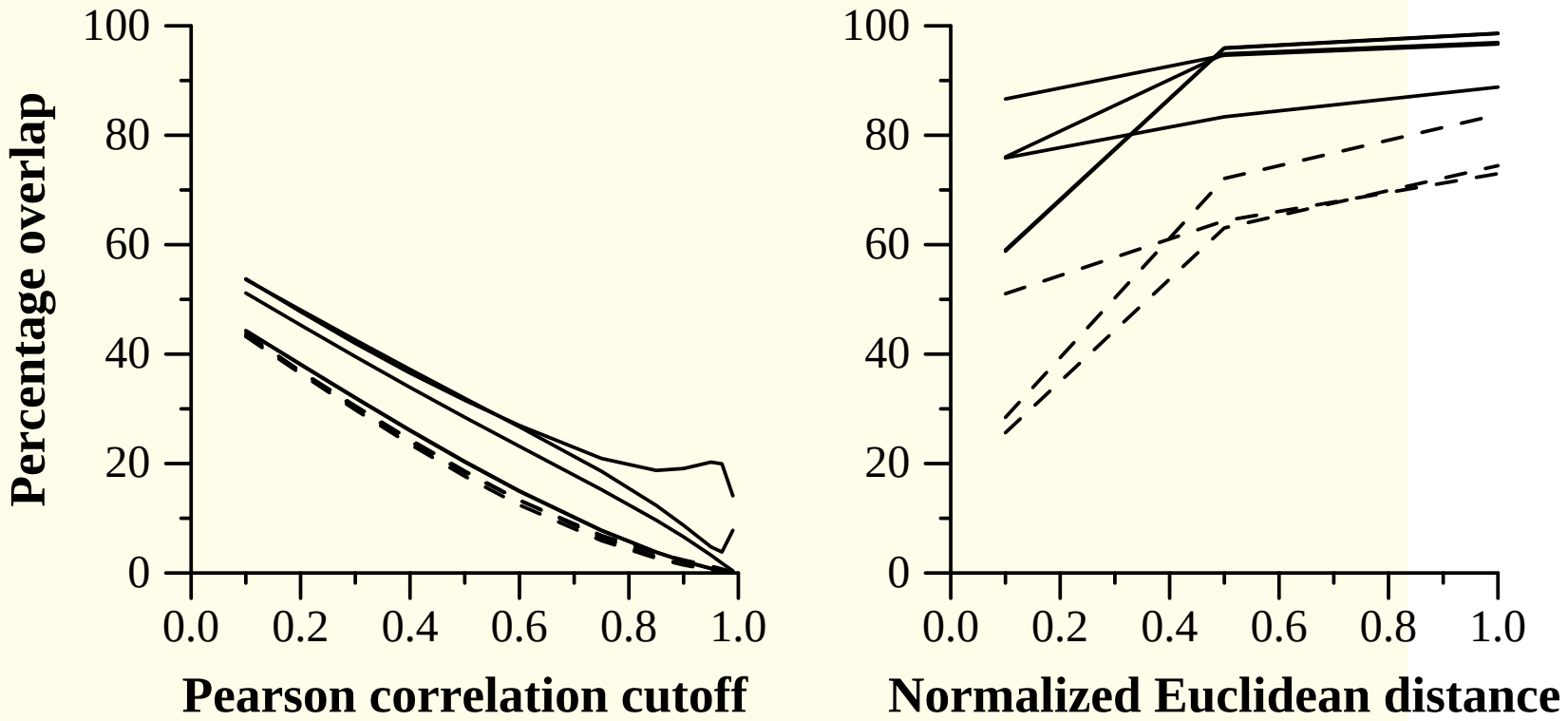
Percentage of outliers for Gene A: $\frac{1}{2}$; Gene B: $\frac{3}{4}$.

Obtained 3 data sets from NCBI's Gene Expression Omnibus (GEO).

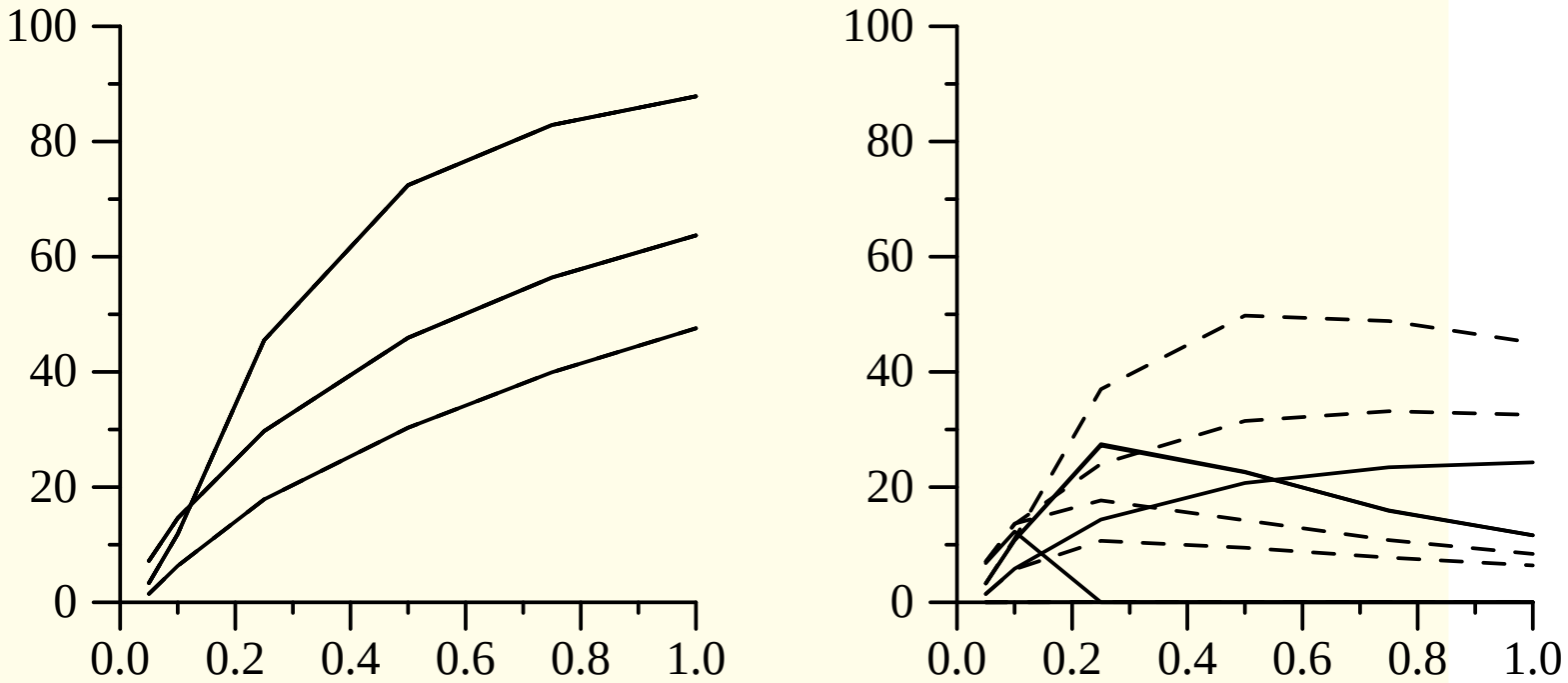
- GDS1515, GDS1689, and GDS1757
- Organism: Arabidopsis
- Genes: 22,814
- All based on the GPL198 platform (Affymetrix GeneChip Arabidopsis ATH1 Genome Array)
- Replicates: GDS1515 (2), GDS1689 (3), and GDS1757 (3)

Selected one replicate from each data set for the "repository" and another for the "test set".

Overlap of gene pairs (edges):

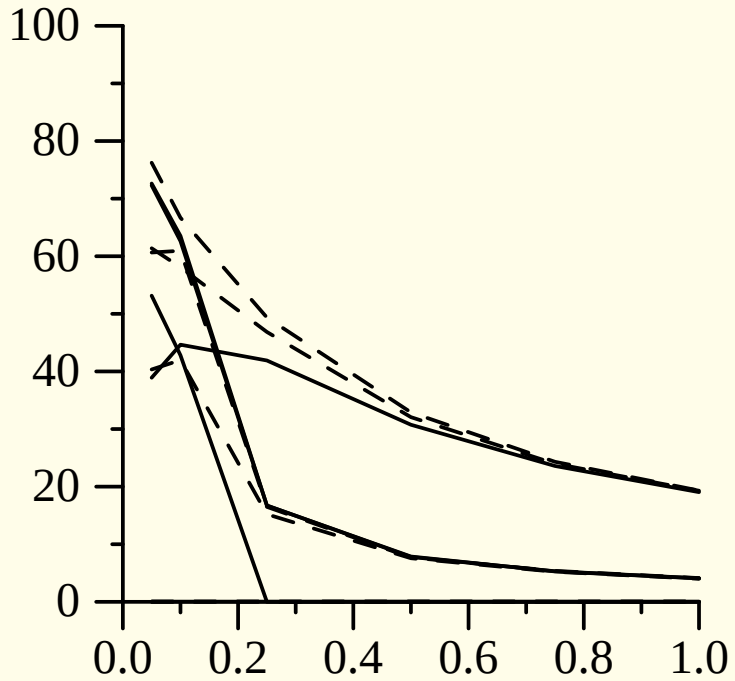


Percentage of edges vs normalized Euclidean distance:



Left: Percentage of edges in repository data.
Right: Percentage of edges which are $< d_{\text{repository}}$, but $> d_{\text{test}}$ (outlying genes).

Results so far not encouraging, though... Percentage of outliers vs normalized Euclidean distance, with a percentage of 10%[1].:



[1] Outliers are genes for which there are fewer than 10% of its neighbours within a distance $d_{\text{repository}} = d_{\text{test}}$.

Still a lot to do...

- Assuming the same distance threshold ($d_{\text{repository}} = d_{\text{test}}$) might be too strict; it might have to be adjusted for each data set.
- Consider using other data to build the graphs instead of the repositories (i.e., Gene Ontology (GO) terms, sequence alignment) to act as a "gold standard" to compare to.

$\Longrightarrow$ Microarrays are known to be noisy and variability in results can occur for many reasons... Stevens and Doerge [2007] noted that different laboratories can report on a different set of statistically significant genes.

Acknowledgements

Collaborators:

- Prof. Hiroshi Mamitsuka
- Dr. Ichigaku Takigawa
- Dr. Åsa M. Wheelock (Karolinska Institute, Sweden)
- Dr. Matthew J. Bartosiewicz (was University of California, Davis)

Data and Software:

- Various microarray sources, including NCBI's GEO
- C4.5 Release 8
- WEKA machine learning program

References

P. E. Blower and K. P. Cross. Decision tree methods in pharmaceutical research. *Current Topics in Medicinal Chemistry*, 6(1):31–39, 2006

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001

J. R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996. Source available from: `http://www.rulequest.com/Personal/`

J. R. Stevens and R. W. Doerge. Combining Affymetrix microarray results. *BMC Bioinformatics*, 6(57), 2007

Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif. RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, 19(12):1578–1579, 2003. Software available from `http://genomics10.bu.edu/yangsu/rankgene/`

I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers, second edition, 2005

R. Wan, I. Takigawa, and H. Mamitsuka. Applying Gaussian distribution-dependent criteria to decision trees for high-dimensional microarray data. In M. M. Dalkilic, S. Kim, and J. Yang, editors, *Proc. Data Mining in Bioinformatics (VDMB) Workshop at the 32nd International Conference on Very Large Data Bases*, volume 4316 of *LNBI*, pages 40–49. Springer-Verlag, September 2006

H. Zhang, C.-Y. Yu, B. Singer, and M. Xiong. Recursive partitioning for tumor classification with gene expression microarray data. *Proc. National Academy of Sciences USA*, 98(12): 6730–6735, June 2001

H. Zhang, C.-Y. Yu, and B. Singer. Cell and tumor classification using gene expression data: Construction of forests. *Proc. National Academy of Sciences USA*, 100(7):4168–4172, April 2003

Data Set Sources

U. Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. National Academy of Sciences USA*, 96(12):6745–6750, June 1999. Data:
`http://microarray.princeton.edu/oncology/affydata/index.html`

T. R. Golub et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999. Data:
`http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

G. J. Gordon et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62(17): 4963–4967, September 2002. Data:
`http://www.chestsurg.org/publications/2002-microarray.aspx`

S. L. Pomeroy et al. Prediction of central nervous system embryonal tumour outcome based on gene expresion. *Nature*, 415(6870):436–442, January 2002. Data:
`http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

S. Ramaswamy et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. National Academy of Sciences USA*, 98(26):15149–15154, December 2001. Data:
`http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

M. A. Shipp et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, January 2002. Data:
`http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

D. Singh et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1 (2):203–209, March 2002. Data:
`http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`