Analysis of Gene Expression Data **Spring Semester, 2002**

Lecture 7: May 05, 2004

Lecturer: Zohar Yakhini *Scribe: Dror Lupu and Eyal Balla*

7.1 Classification

This scribe is based on lecture presentation by Z. Yakhini.[1]

7.1.1 Overview

Genome projects have produced large amounts of data on the sequences of new genes whose functions are as yet unknown. The functions of new genes are usually inferred by comparing their sequences with those of known genes, but evaluation of the sequence homology of individual genes does not make the most of the available sequence information. Therefore, new methods and tools for extracting more biological information from homology searches would be advantageous.
Classifying genes into groups according to their functionality would greatly support such efforts. Classification of genes extracts information about genes from expression level chips (such as in cDNA chips, Affimetrix chips and Agilent chips - see Figure 7.1). The methods described hereafter use RNA expression data for classifying their carriers to hopefully meaningful classes. We will mention several methods of classification and discuss the algorithms used in some of them.

What can be achieved with Classification?

- Differentiation between normal and tumor tissues.
- Distinguishing between various types of pathology and stages in tumors.
- Help decide of the susceptibility of a treatment for various patients.
- Prognosis of risk levels of patients.

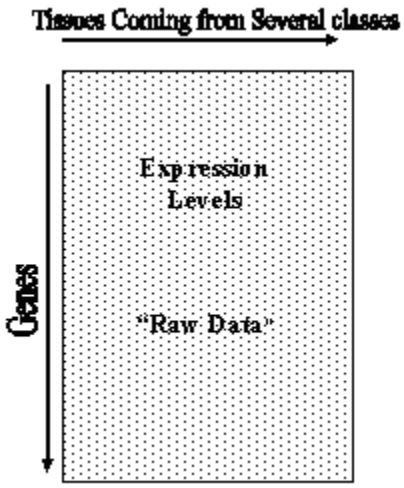

Figure 7.1: The expression levels in a gene array

7.1.2 Informative Genes

Informative Genes are those genes that are differentially expressed in the classes to which the data has been classified into. It is these genes that will be the ones that will give us the insight that will help achieve the goals mentioned in the previous section. Moreover finding such genes can also help to radically reduce the dimensionality of the data we are dealing with (see Figure 7.2).

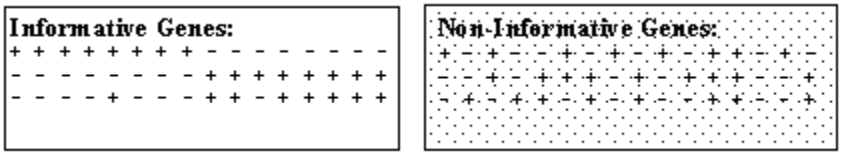

Figure 7.2: The figure shows the expression level of a single gene after sorting the expression levels of the samples and assigning labels according to the known classes of each sample.

7.1.3 Classification Scores

TNOM Score

The TNOM score gives scores to a classifier of two classes. The score is the minimum number of errors that can be achieved using one separator that assigns all samples on its left one class label and the second class label to all samples on its right.
A Perfect Classifier will get the TNOM score of 0, for there will be no errors using the separator that has all of those labelled with the first label on its left and the latter label on its right.

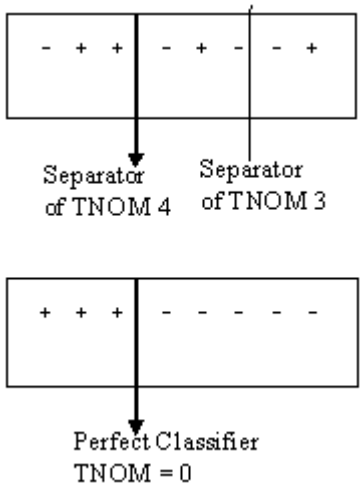

Figure 7.3: The calculation of a TNOM two classifiers.

INFO Score

The Threshold Mutual Information Score known as the *INFO Score* is the minimal conditional entropy of the annotation, given a threshold partition (minimum taken over thresholds).

Separation Score

The Separation score assumes that both classes distribute with gaussian distribution (μ_1, σ_1) and (μ_2, σ_2) accordingly. The separation score is defined to be $\frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$. This score maximizes

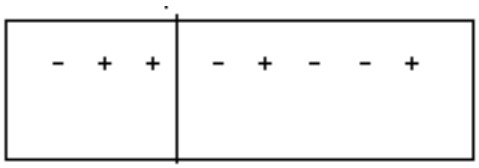

Figure 7.4: The figure shows the threshold with the INFO Score of $\frac{3}{8}H(\frac{1}{3}) + \frac{5}{8}H(\frac{2}{5})$

the likelihood when the score is at its maximum assuming the assumptions about the distribution of the samples are correct.

7.1.4 P-values

Relevance scores are more useful when we can compute their significance. *P-Value* is the probability of finding a gene with a given score if the labelling is random. P-values allow a higher level statistical assessment of the quality of the data and provide a uniform platform for comparing relevance, across data sets.

TNOM P-Values Calculation

In order to find the probability of a gene classifier getting a specific TNOM score the following question arises. What is the probability that a vector with p "+1" and q "-1" which is drawn uniformly is given the TNOM score of t.
Let us look at the corresponding paths in R^2 that begin in (0,0) and for each sample i in vector v assigns position $(i, y(i-1)+v(i))$ where $y(t)$ is the value the path is given at position i and $v(i)$ is the vector value in position i (see figure 7.5).

All paths for given (p,q) are bounded by the paths of the two perfect classifiers (see figures 7.6 and 7.7).

Theorem 7.1 *A vector has a TNOM score $\leq s$ iff the corresponding path crosses either of the lines Y=p-s and Y=s-q.*

Proof: Let $\pi(i)$ be the position in sample i. Then $\pi(i) = p(i) - q(i)$ then lets assume that on left there should be "+". This leads to $TNOM(i) = q(i) + p - p(i) = p - \pi(i) \leq s$.

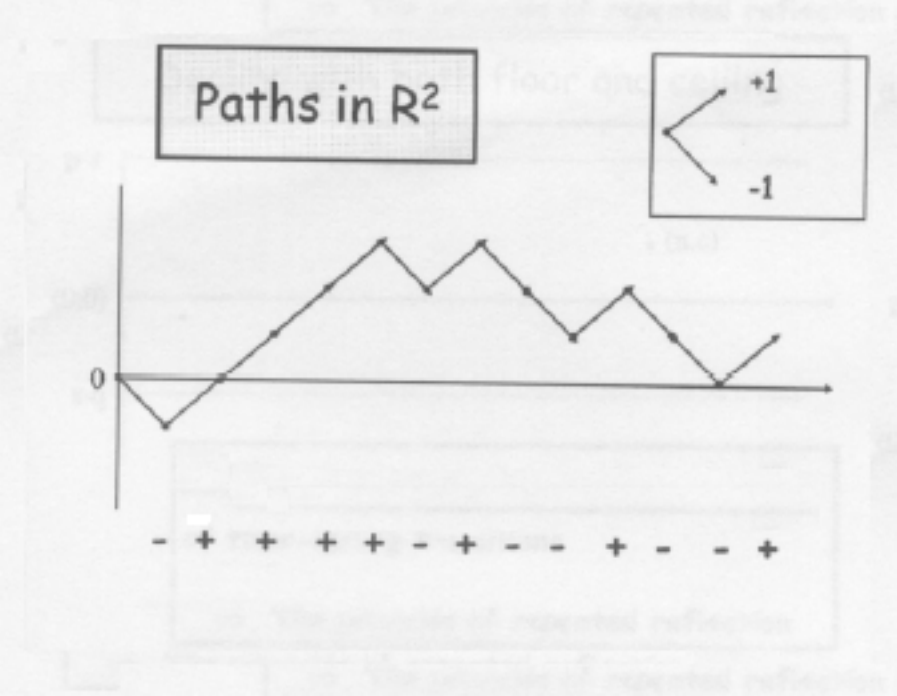

Figure 7.5: The figure shows the path created in R^2 by the vector (-1 +1 +1 +1 +1 -1 +1 -1 -1 +1 -1 -1 +1)

Therefore $\pi(i) \geq p - s$. The same could be done assuming there should be "-" on the left getting $\pi(i) \leq s - q$. ■

In order to evaluate the P-Value of a TNOM score we need to know how many of the paths with the same (p,q) have at most the same TNOM score.

Problem 7.1 Find the number of paths with (p,q) that have a TNOM score of at most s.

The Reflection Principal

Lets assume that we want to find the number of paths with TNOM score $\leq$ s of path with p "+" and q "-". Then using the previous theorem we need to consider the path that cross p-s and s-q. We will show the analysis for p-s.

Theorem 7.2 *There is a one-to-one correspondence and onto mapping between the paths that start with* $(0, 0)$ *and end at* $(p + q, p - q)$ *and cross p-s and the path that start with* $(0, 2 * (p - s))$ *and end in* $(p + q, p - q)$ *(see figure 7.8).*

Proof: Let i be the first position in which the path l crosses $p - s$. Then it will be mapped to the path with the part until i reflected along $p - s$ with the remaining of l untouched.

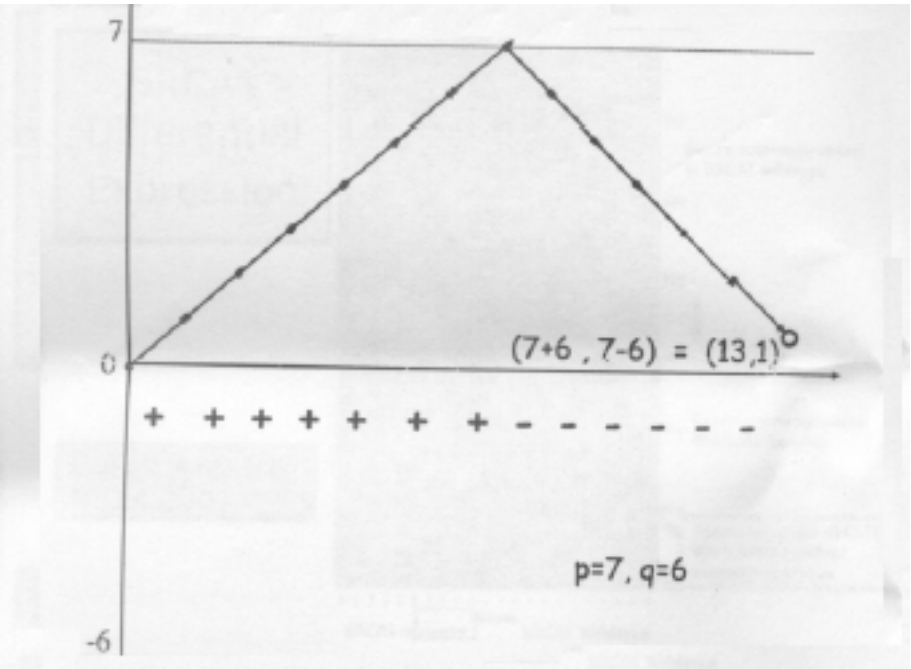

Figure 7.6: The figure shows the path created in R^2 by the Perfect classifier with 7 plus and 6 minus with the plus first

This mapping is onto because each path from $(0, 2 * (p - s))$ crosses $p - s$ to get to $p - q$ and therefore a reflected path from $(0, 0)$ can be found. A one to one correspondence can be shown because either the first part of the path is different (which will lead to a different reflection and therefore another path) or the last part is different (which also leads to a different path) and therefore is a one to one correspondence and onto mapping. ■

Calculating the number of paths starting with $(0, 0)$ and ending with $(p + q, p - q)$ is $\binom{p+q}{p}$. The number of paths starting at $(0, 2 * (p - s))$ and ending at $(p + q, p - q)$ is the same the number of paths from $(0, 0)$ to $(p + q, p - q - 2 * (p - s))$, which is equal to $\binom{p+q}{s}$. The same analysis could be done for the lower bound $s - q$. However doing this some paths are counted twice. The paths that are counted twice are those that cross both bounds. Counting these path is not possible. However we can count the path that first cross the higher bound and then cross the lower bound or vice versa. But again doing so will not suffice because the paths can recross the bounds again and be recounted. This leads us to the full probability theorem. However one does not need all the arguments for the number of recrosses. since the path is $p + q$ long and in order to recross after a first cross the path will grow in at least $p - s - s + q = p + q - 2s$ then the maximum number is bounded.

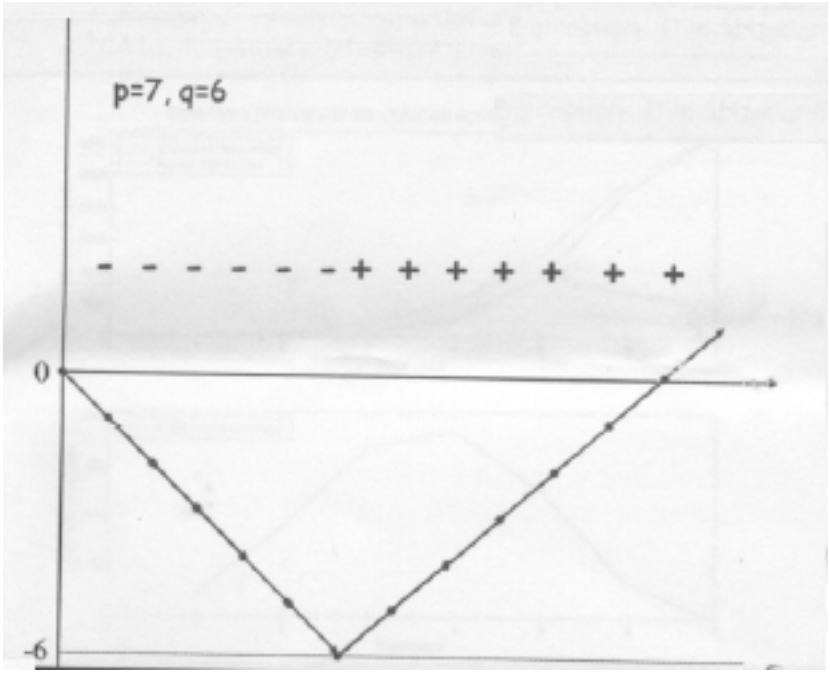

Figure 7.7: The figure shows the path created in R^2 by the Perfect classifier with 7 plus and 6 minus with the minus first

7.1.5 Using P-Values

After calculating the P-values for each score these values could be compared to the ones found in the experiment. If the values found in the experiment are significantly higher than those calculated by the P-Values then it is statistically valid.

7.1.6 INFO Score P-Values

The analysis for INFO Score P-Values calculation is similar to the TNOM analysis in the fact that they too are bounded between the two perfect classifiers. However there is no criteria that holds for all i such that if $\pi(i) \geq k$ then the INFO score $\leq$ then s. Such criteria is i-dependent. Such a criteria is computable using dynamic programming.

7.1.7 Gaussian Error Score

Until now we dealt with two class partitions only, the Gaussian Error Score could be generalized to deal with an arbitrary number of classifiers. First the distribution functions are

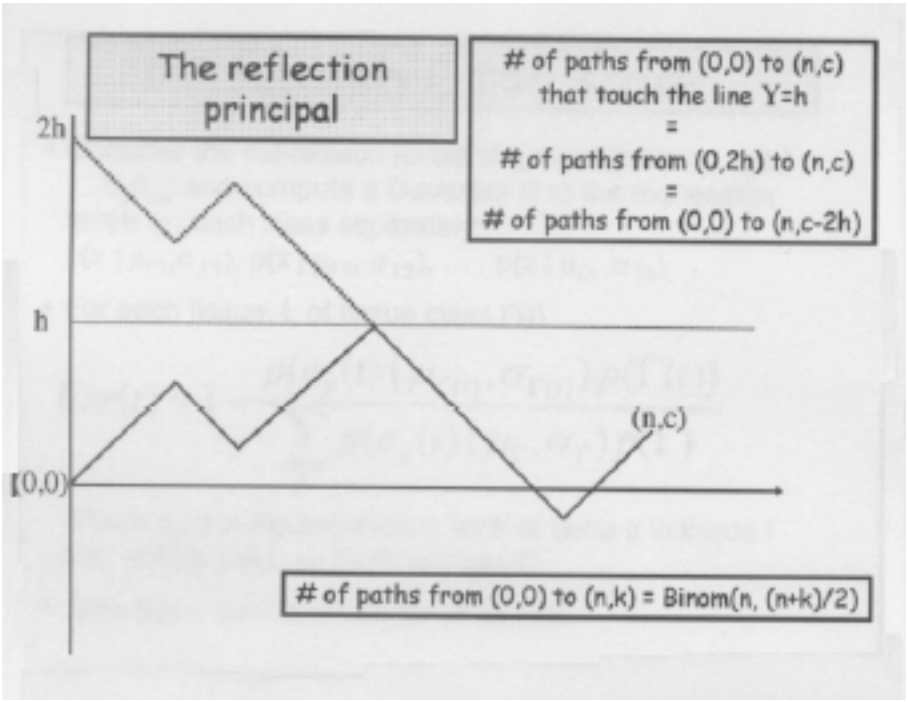

Figure 7.8: The Reflection Principal activated.

calculated according to the classification given. After which the error

$$Err(t) = 1 - \frac{p(e_g(t)|\mu_{\Gamma(t)}, \sigma_{\Gamma(t)})p(\Gamma(t))}{\sum_\Gamma p(e_g(t)|\mu_\Gamma, \sigma_\Gamma)p(\Gamma)}$$

is calculated for each sample and the sum of errors is the score of the classifier. When again a lower score is a better score.

7.1.8 Single Gene Voters

The single voter procedure assigns a score to each gene g for each class T, that represents the magnitude of the sample g coming from class T. The votes are between 0 and 1 and can be calculated using the following calculation:

$$V_g(x) = \frac{p(x|\mu_{\Gamma_1}, \sigma_{\Gamma_1})p(\Gamma_1), p(x|\mu_{\Gamma_2}, \sigma_{\Gamma_2})p(\Gamma_2), ..., p(x|\mu_{\Gamma_s}, \sigma_{\Gamma_s})p(\Gamma_s)}{\sum_\Gamma p(x|\mu_\Gamma, \sigma_\Gamma)p(\Gamma)}$$

7.1.9 Linear Programming Classifier

The Linear Programming Classifier computes a set of weights w_g where g ranges over all genes. These weights are computed by solving the linear program constraints that finds the

max r for which

$$\sum_g w_g(V_g(e_g(t), \Gamma(t)) - V_g(e_g(t), \Gamma)) \geq r$$

for all samples t in the training data and for all classes $\Gamma \neq \Gamma(t)$.
A new sample t will be assigned to the class Γ that maximizes $\sum_g w_g V_g(e_g(t), \Gamma)$.

7.1.10 Other classifying methods

Other classifying methods are available. Among them are:

- Naive Bayesian
- SVM - Support Vector Machines
- K-NN - K- nearest neighbors

7.1.11 Validation of classifiers using LOOCV

Leave One Out Cross Validation (LOOCV) is a method for validating the classifying of using the training data. The method runs the classifying algorithm on the training data without using one (or one part) of the training data. This left-out data set is used to verify that the classification model created is not too specific but can also generalize. This is done leaving out each one of the items (or parts) of the training data when the classification chosen is the one that performed best with the left-out data.

Bibliography

[1] Zohar Yakhini. Gene scoring, classification, class discovery. 2004.