Examining Microarray Experiments

Within the past decade, advances in biological technology have allowed researchers to penetrate the foundations of cellular biology, culminating with the publication of the draft sequence of the human genome. Researchers from both private and public institutions have succeeded in identifying thousands of coding regions of genetic sequences commonly referred to as genes. The genes of yeast and other simple organisms have been completely identified and an estimated two-thirds of the human genome has been identified. Despite this widely heralded effort, sequencing the genome and identifying its coding regions are only the first steps in understanding the wildly turbid and complicated functions of genes at a cellular level.

Furthermore, gene array technologies, developed over the past six years, have begun to produce valuable information for understanding the functions of genes at a cellular level. The arrays simultaneously measure cellular concentrations of thousands of messenger RNAs or mRNAs since transcription from DNA to mRNA is the initial step in the creation of a protein from a gene. Measurement of mRNA provides a proxy for the activity level of a particular gene and is referred to as an *expression level* or gene expression. Prior to the advent of array technologies, researchers could only measure expression levels for individual genes through techniques such as Northern blots. Array technologies allow investigation of expression levels for all known genes in the cell and permit a more thorough characterization of the transcriptional state of the cell. So far array technologies have been used to characterize functions of newly found genes, to understand the machinery of cell maintenance and regulation and to classify certain types of cancer. Insofar as this new technology has become useful, it has required a good deal of statistical and computational methods aimed at understanding how the technology both succeeds and fails at representing the state of the cell accurately.

There are two differnet array technologies currently used. Both measure the propensity of mRNA or cDNA to hybridize with single strands of complementary DNA sequestered and immobilized on a solid substrate. The technologies differ in the length of the sequesterd DNA strands and the number of spots used to detect the expression level of each gene. In this paper the focus is on the GeneChip, developed by Affymetrix Inc. It would not be an efficient use of space explaining how the GeneChip actually measures the expression levels of the genes. However, it is

necessary to explain what the output of the GeneChip looks like and how faithfully it represents the actual level of mRNA for each gene.

The GeneChip array uses 40 different oligonucleotide sequences at 40 adjacent spots on the array to detect each type of mRNA molecule. 20 of these spots contain *perfect match* sequences that are exactly complementary to the subsequences of the target gene's mRNA. The other 20 contain *mismatch* sequences which differ by a single base pair from its corresponding perfect match sequence. Affymetrix measures cross-hybridization, the non-specific hybridization that can occur with genes of similar sequence, by subtracting the perfect match and mismatch sequences as the mismatch sequence is intended to represent this non-specific hybridization. The differences are then averaged to produce an expression level specific to each gene.

Gene array experiments make use of a large number of steps that are not insulated by any means from error which, in turn, lead to noise in the resulting expression levels. Current practice addresses problems of noisy data heuristically but there has been a growing literature that deals with noise estimation specifically. In this paper I follow the approaches of Wong and Li (2001) who model the distribution of individual perfect match – mismatch probe differences. Once expression levels have been estimated from the perfect match – mistmatch values, a matrix consisting of gene expressions for each gene in each sample is filtered and clustered. The filtering is necessary to exclude large amounts of data that remain unchanged from experimental condition to experimental condition. Usually, change in expression is measured by that of a statistical magnitude test like that of a *t*-test which relies on estimating the distribution of expression levels across experiments for each gene. A simple test like the ratio of sample standard deviation over mean expression value across experiments can also be used to exclude genes. Clustering algorithms, for example hierarchical clustering, have been used ubiquitously to group those genes that are manifesting change between experiments. Two frequently cited examples of clustering methods are due to Eisen (1998) who uses agglomerative hierarchical clustering to group genes and Tavazoie (1999) who uses k-means clustering. Agglomerative clustering starts at the lowest level of dissimilarity of expression values (namely zero for each singleton cluster consisting of only one gene) and at each level recursively merges a selected pair of clusters into a single cluster which produces a grouping at the next higher level with one less cluster. K-means works by

choosing *K* initial means for the data and moving the means around to minimize the variance of the mean and the points within the cluster corresponding to the particular mean. Clustering is in itself a large field and there remains no consensus which clustering method is most suitable for gene expression data. Suffice it to say, many researchers use heirarchical clustering so that the number of naturally ocurring clusters within the data may be inferred rather than imputed. The technique by Eisen is probably the most widely used clustering method of expression data. Thus, since the scope of this paper goes beyond clustering, the results presented in this paper are based on this widely used method.

As stated microarray experiments require researchers to analyze large amounts of gene expression data. The number of gene expressions recorded may be in the tens of thousands and the number of experiments may well be in the hundreds. This approach leads to a disturbingly high dimension if one is interested in modeling gene to gene interactions and even intimidating for within sample analysis. Even after filtering and clustering are used to identify tenuous groups of genes in experiments that are changing, it is instructive for the biologist to know not only which genes are clustering but also the effect different experiments and their corresponding treatment conditions have on such clusters. This thesis addresses the following questions: what are most important experiments for the inducement of a specific cluster?, how dissimilar are the gene expressions in this cluster compared to those in other clusters? These are just some of the prevailing questions that are of significant importance to the researchers pursuing functional relationships of genes via clusters.

In this thesis we approach these questions by initially filtering and clustering the data with the agglomerative hierarchical clustering algorithm proposed by Eisen. For each cluster of data a ranking of the most informative experiments is computed so that we may query a given set of conditions and find the clusters whos information ranking corresponds to those experiments. Principal component projections are then made to reduce the experimental space to an orthogonal subspace. This has the advantage that highly correlated experiments will not bias the formation of possible groups of data in the subspace as it does in the experimental space. Methods of unsupervised learning are then used to discern possible modes in the projected data so that we can compare or justify the clusters in the experimental space with the groups of data found in the orthogonal subspace.

Entropy and Mutual Information

Given an *n* x *m* matrix of genome expression data consisting of *k* clusters with $g_1,...,g_n$ genes in $s_1,...,s_m$ experiments , biologists are interested in knowing which experiments could be perceived as being the most important in the inducement of the cluster. That is, if one were to choose one of the $k = 1,...,K$ clusters of genes can one estimate how dissimilar the gene expressions or distribution of genes expressions in cluster *k* are from those in clusters $1,...,k-1,k+1,...,K$. Once this is done for each cluster over each experiment we have an ordinal *ranking* of the importance of each experiment.

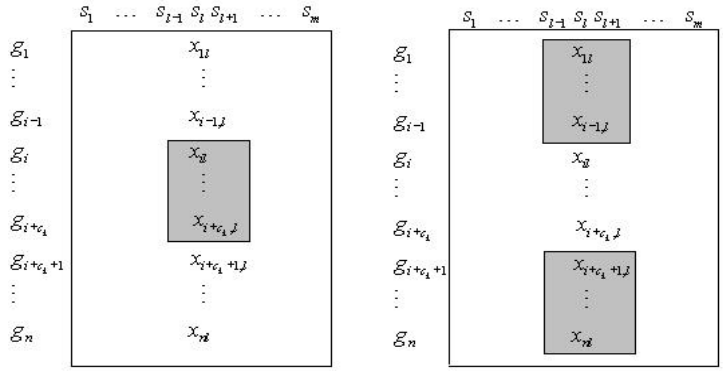

figure 1: genes in cluster and genes outside of cluster, respecitvely, for cluster *k* containing genes with indices $i,...,i+c_k$

The question then becomes, what is an appropriate estimation of dissimilarity between gene expression distributions for gene expressions within a given cluster to those in the other clusters. If one can estimate the distribution of the set of gene expressions exclusive of the genes of the cluster in question, how dissimilar will the estimated distribution be? This motivates the use of i*nformation*, more precisely *mutual information* or *relative entropy* (equivalent to the well known Kullback-Leibler distance) thoroughly presented in Cover and Thomas (1991) and Papoulis (1991) and derived from the work of Shannon and Weaver (1949). Historically the term *entropy* refers to the *entropy of partition U* such that the entire experiment space will be contained in the union of the mutually exclusive *U*s. Referring to figure 1, the union of the grey-shaded areas constitute the entirety of the experiment space s_l .

Entropy was derived to satisfy a number of postulates based on a heuristic understanding of uncertainty. If entropy of partition *U* is denoted $h(U)$ then h must satisfy the following postulates:

1. $h(U)$ is a continuous function of the probability $p_i = P(X_i)$ of an event X_i over partition *U*.
2. If $p_1 = \cdots = p_N = 1/N$ then $h(U)$ is an increasing function of *N*
3. If a new partition *B* is formed by subdividing one of the sets of *U*, then $h(B) \geq h(U)$.

It can be shown in Papoulis (1991) that $h(X) = h(U_X) = -\sum_{i=1}^{N} p_i \log p_i$ satisfies the above postulates for discrete random variable *X* over partition *U* and

$h(X) = -\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx$ where $P(-\infty \leq X < \infty) = \int_{-\infty}^{\infty} f(x)\, dx$ for continous random variable *X*. Similarily, if *X* and *Y* are two discrete type random variables such that $P(X = x_i, Y = y_j) = p_{ij}$ then the *joint entropy* is defined by the entropy of the product of their respective partitions, that is $U_X \cdot U_y = \{X = x_i, Y = y_j\}$ so that

$h(X,Y) = h(U_X, U_y) = -\sum_i \sum_j p_{ij} \ln p_{ij} = E\left[-\ln p(X,Y)\right]$ in the discrete case and

$h(X,Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \ln f(x,y) = E\left[-\ln f(X,Y)\right]$ for the continuous case, where

$E[\,\cdot\,]$ represents the expected value of the term in brackets. *Conditional entropy*, as expected, works analogously. However the probability of one random variable will be conditioned on the assumption that either $Y = y_j$ or U_Y is known. The former condition yields

$h(X \mid Y = y_j) = -\sum_i (p_{ji} / p_j) \ln (p_{ji} / p_j)$ and the latter condition

$h(X \mid Y) = -\sum_i p_j h(X \mid Y = y_j) = -\sum_i p_j \ln (p_{ji} / p_j)$ for the discrete case. The continuous case yields

$h(X \mid Y = y_j) = -\int_{-\infty}^{\infty} f(x \mid y) \ln f(x \mid y)\, dx = E\left[-\ln f(X \mid Y) \mid Y = y\right]$ and

$$h(X|Y) = -\int_{-\infty}^{\infty} f(y)\ln h(X|y)\,dy = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ln f(x|y)\,dxdy$$
$$= E\left[-\ln f(X|Y)\right] = E\left[E\left[-\ln f(X|Y)|Y=y\right]\right]$$

Mutual Information is now almost trivially defined as the entropy over the partition $U_X \cup U_Y \setminus U_X \cap U_Y$,

$$I(U_X,U_Y) = h(U_X \cup U_Y \setminus U_X \cap U_Y) = h(U_X) + h(U_Y) - h(U_x \cdot U_y)$$
$$= h(X) + h(Y) - h(X,Y)$$
$$= h(X) - h(X|Y)$$

since $f(X,Y) = f(X|Y)f(Y)$. In terms of expectations, the mutual information can be written $I(X,Y) = E\left[\ln \frac{f(X,Y)}{f(X)f(Y)}\right]$, the discrete case yields a similar expression.

Thus, if continuous random variables *X* and *Y* are independent then $f(X,Y) = f(X)f(Y)$ so that their shared or *mutual information* is zero. In the parlance of our experiments, the more the independent the random variable whos distribution is estimated from genes $g_i,\ldots,g_{i+c_k}$ is from the random variable whos distribution is estimated from genes $g_1,\ldots,g_{i-1},g_{i+c_k+1},\ldots,g_n$ the less the information shared between the two random variables. Below is a depiction of the data transformation from an *K*x*m* mutual information matrix to an *K*x1 mutual information ranking.

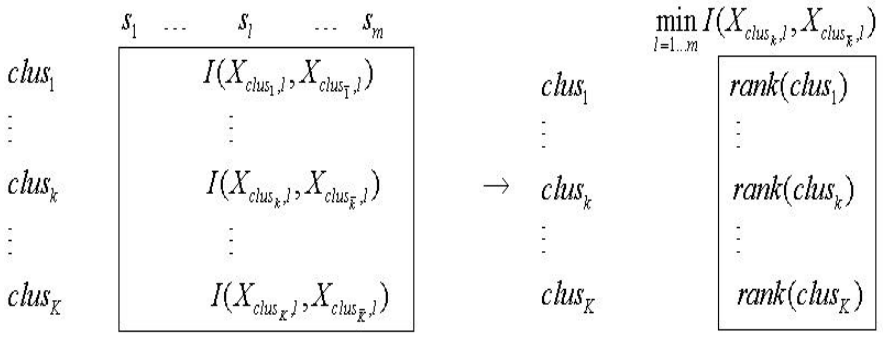

figure 2 *K* x *m* matrix of information on the left leads two a ranking of the experiments by taking the minimum information over all experiments for each cluster

In figure 2, $X_{clus_k,l}$ represents the random variable of gene expressions in cluster *k* of experiment *l* and $X_{clus_{\bar{k}},l}$ the random variable of gene expressions not in cluster *k* of

experiment *l*. $rank(clus_k)$ is the minimum of the mutual information for cluster *k* across all experiments *l* = 1,…,*m*. *rank*( ) can either represent the one experiment of lowest information content or can just sort the experiments *l* = 1,…,*m* in ascending order of mutual information.

Many researchers take natural or base-2 log transfroms of the raw expression data to impose normality on expression levels since all expression levels are greater than zero and assumed to have a lognormal distribution. If expression levels can indeed be imputed to be lognormal (normal after log-transforms) than the mutual information content $I(X_{clus_k,l}, X_{clus_{\bar{k}},l})$ becomes extremely simple to compute. For continuous random variables *X* and *Y* we can remove the mean and compute

$I(X,Y) = h(X) + h(Y) - h(X,Y)$

by $E[-\ln f(X)] + E[-\ln f(X)] - E[-\ln f(X,Y)]$ where

$$E[-\ln f(X,Y)] = E\left[\ln 2p s_X s_Y (1-r^2)^{1/2}\right] + \frac{1}{2(1-r^2)} E\left[\frac{X^2}{s_X^2} - \frac{2r s_{XY}}{s_X s_Y} + \frac{X^2}{s_Y^2}\right]$$

$$= \ln\left(2p s_X s_Y (1-r^2)^{1/2}\right) + \frac{1}{2(1-r^2)}\left(\frac{s_X^2}{s_X^2} - \frac{2r^2 s_X s_Y}{s_X s_Y} + \frac{s_Y^2}{s_Y^2}\right)$$

$$= \ln\left(2p s_X s_Y (1-r^2)^{1/2}\right) + 1$$

$$= \frac{1}{2}\ln\left((2p e)^2 s_X^2 s_Y^2 (1-r^2)\right)$$

where $s_{XY} = r s_X s_Y$ for correlation r. It is worthwhile to notice that $s_X^2 s_Y^2 (1-r^2)$ is the determinant of the 2x2 covariance matrix for X and *Y* if we substitute the off-diagonal s_{XY} terms with the equivalent $r s_X s_Y$ terms. In fact, the general form for entropy $h(X_1,..,X_n)$ of stationary random variables $\{X_1,...,X_n\}$ can be written $\frac{1}{2}\ln\left((2p e)^n \Delta\right)$ where Δ is the determinant of the *n*x*n* covariance matrix with off-diagonal substitutions $r_{ij} s_{X_i} s_{X_j}$ $i, j = 1,..,n$ Papoulis (1991). Now the mutual information is $I(X,Y) = \frac{1}{2}\ln(2p e) s_X + \frac{1}{2}\ln(2p e) s_Y - \frac{1}{2}\ln((2p e)^2 |\Delta_{X,Y}|)$. Since the number expression values from $clus_k$ will not be equal in number to those of $clus_{\bar{k}}$, to compute the

sample covariance matrix one must resample the cluster with the fewer number of expression values via the nonparametric bootstrap (Efron 1993).

There is no dearth of literature disputing the most appropriate distribution of gene expressions. Mutual information is most easily computed for normal data but we can compute the mutual information between two random variables with any parametric or non-parametric distribution by using numerical quadrature. Silverman (1982) presents a very efficient method for non-parametric density estimation which reduces to binning the raw data and convolving the binned data with a gaussian kernel by using a fast fourier transform. In practice, some of the clusters that need to be evaluated have not more than a few data points so parametric assumptions are necessary. Figure 3 represents the diagram of a non-parametric estimate using Silverman's algorithm together with a maximum-likelihood fitted gaussian distribution. There seems to be not much difference in appearance thus justifying a gaussian distribution to represent the log-transformed data.

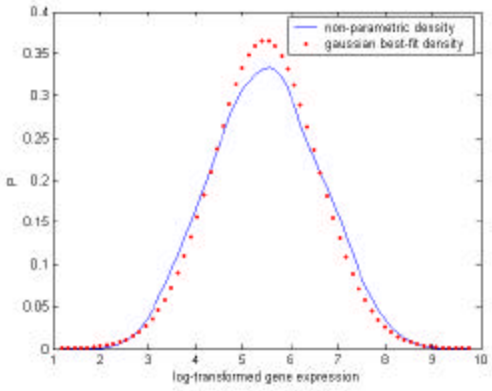

figure 3 solid curve represents non-parametric fit to data
dotted curve represents best gaussian fit

Using the gaussian estimation for the distribution, the mutual information matrix is

	s_1	...	s_j	...	s_m
$chus_1$	...	$\frac{1}{2}\ln(2\pi e)\sigma_{X_{chus_1 1}} + \frac{1}{2}\ln(2\pi e)\sigma_{X_{chus_1 1}} - \frac{1}{2}\ln((2\pi e)^2\lvert\Delta_{X_{chus_1 1}X_{chus_1 1}}\rvert)$			
$chus_k$	...	$\frac{1}{2}\ln(2\pi e)\sigma_{X_{chus_k j}} + \frac{1}{2}\ln(2\pi e)\sigma_{X_{chus_k k}} - \frac{1}{2}\ln((2\pi e)^2\lvert\Delta_{X_{chus_k k}X_{chus_k k}}\rvert)$			
$chus_K$	...	$\frac{1}{2}\ln(2\pi e)\sigma_{X_{chus_K 1}} + \frac{1}{2}\ln(2\pi e)\sigma_{X_{chus_K 1}} - \frac{1}{2}\ln((2\pi e)^2\lvert\Delta_{X_{chus_K 1}X_{chus_K 1}}\rvert)$			

.

Now we have a way of determining which experiments most likely induce any particular cluster of the expression data. If one experiment seems to induce most or all of the clusters then it can likely be inferred that this specific experiment and corresponding treatment conditions tend to overshadow the effects of the other experiments and their corresponding treatment conditions. Removal of this experiment may result in finding different rankings of the experiments with more subtle effects for different clusters and thus a possible causal relationship between the set of genes in the cluster and the experiment with its coinciding treatment conditions. Rankings of experiments in this way may also suggest which treatments researchers should include or exclude in an experiment to obtain clusters of genes that reflect more subtle regulations of genes by treatment conditions. It is sometimes the case that two or more experiments are highly correlated because of one or the confluence of two treatments shared by all. Thus the added cost and laboratory time is lost because the data from this experiment sheds little new light. From this perspective, it may be instructive to take the *n*x*m* matrix of gene expressions and map it to an *n*x*p* matrix in a *p*-dimensional orthogonal subspace, $p \le m$. This motivates the use of principal component projection. Other projection methods such as exploratory projection pursuit and independent component analysis are also used to find "interesting" projections of the data (projections that yield highly irregular patterns that deviate severly from normality) but are much more sophisticated and may not be necessary if the relatively simple principal component projection produces desirable results. Finally methods in pattern classification can be used to find distinctions in the projected data. All this is to say that it may be more useful to examine projections of the data instead of the experimental data itself – especially when the data form different experiments is highly correlated.

Principal Component Analysis

Principal components analysis involves the spectral decomposition of the symmetric covariance matrix $\Sigma = ADA^t$ where *D* is the diagonal matrix of eigenvalues and *A* the corresponding orthogonal matrix of eigenvectors. If the eigenvalues are distinct and ordered $0 \le l_1 \le l_2 \le ... \le l_m$, then the columns $A_1,..,A_m$ of *A* are unique up to sign. The random variables $V_j = A_j^t X$, $j = 1,..,m$ have covariance matrix $A^t \Sigma A = D$. These random variables are called principal components. They are uncorrelated and increase in variance from V_1 to

V_m. It may be of interest to compute the correlation of the *j*th principal component with that of the *i*th experiment in order to ascribe experimental meaning to the *j*th principal component. Consider the *j*th component $V_j = A_j^t X$ and $\mathrm{cov}(I_{\mathrm{nxn}}\mathrm{X}, V_j) = I_{\mathrm{nxn}}\Sigma A_j = \Sigma A_j = l_j A_j$ since A_j is the eigenvector of sample covariance matrix Σ and I_{nxn} the *n*x*n* identity matrix. The covariance of X_i and V_j is the *i*th element of $l_j A_j$, namely $l_j A_{ij}$. The standard deviation of X_i is $\sqrt{s_{ii}^2}$ and that of A_j is $\sqrt{l_j}$ so that the correlation between the *i*th experiment and *j*th principal component is $r_{X_i V_j} = \dfrac{l_j A_{ij}}{\sqrt{s_{ii}^2 l_j}} = A_{ij}\dfrac{\sqrt{l_j}}{\sqrt{s_{ii}^2}}$. The univariate measure $R_{X_i}^2$ depicting how each experiment X_i alone relates to each $V_1,..,V_p$ can be computed $R_{X_i}^2 = \sum_{j=1}^{p} r_{X_i V_j}^2$.

The *p* of the *p*-dimensional principal component subspace is chosen such that some threshold of variation in the data is accounted for by these $p \le m$ principal component vectors. For visualization purposes one can choose *p = 2* or *p* = 3. In most of the microarray experiments, almost 98 percent of variation in the data is accounted for by the first two principal components. After choosing *p* and making the projection, the question reduces to *unsupervised learning* of structure within this new space. Clustering itself lies within the realm of unsupervised learning techniques. One may also consider to another such instance of unsupervised learning – that of non-parametrically estimating the density of log-transformed data developed by Silverman. If principal component projection is successful then hopes to distinguish associations among the experiments and whether or not they can be considered as functions of a smaller set of *latent* variables evinced by the modality of the data. Thus, perhaps it is possible to represent the distribution of data by a mixture of simpler densities representing types or classes of observations. To this, the *EM* algorithm approach may be use [Dempster (1977)] to distinguish the possible classes of observations and then draw boundaries, when appropriate, to denote these possible classes.

The EM Algorithm

The *Expectation Maximization* or *EM* algorithm is less an algorithm than a prescription for an algorithm. It is a numerically stable tool for simplifying difficult maximum likelihood problems and can easily be understood in the context of a simple mixture model. Perhaps we have reason to believe, by insight or observation, that the experimental data in consideration is bimodal so that estimating the distribution with a single Gaussian distribution would not be sufficient. Instead we could model the data *X* as a mixture of two Gaussian distributions:

$$X_1 \sim N_p(m_1, \Sigma_1)$$
$$X_2 \sim N_p(m_2, \Sigma_2)$$
$$X = (1-q)X_1 + q X_2$$

for $q_i \in \{0,1\}$, $\Pr(q = 1) = p$ and N_p denotes the *p*-dimensional normal distribution. If $f_{h_i}(x)$ is the *p*-dimensional normal density with parameters $h_i = \{m_i, \Sigma_i\}$ then the density of *X* is $f_X(x) = (1-p)f_{h_1}(x) + p f_{h_2}(x)$ and the parameters to be estimated are $\{p, h_1, h_2\}$. The log-likelihood using the *n* data points is

$l(h; X) = \sum_{i=1}^{n} \log[(1-p)f_{h_1}(x_i) + p f_{h_2}(x_i)]$ and the conventional maximum likelihood method would require maximization of $l(h; X)$ (here *X* is fixed and the parameters $\{p, h_1, h_2\}$ variables). The sum of terms inside the logarithm makes maximization numerically challenging. However, if one considers the *latent* variables q_i such that when q_i =1 X_i comes from f_{h_1} and when q_i =0 X_i comes from f_{h_2}. The new log-likelihood is $l(h; X, q) = \sum_{i=1}^{n} \log[(1-q_i)f_{h_1}(x_i) + q_i f_{h_2}(x_i)]$. The maximum likelihood parameter estimates for this log-likelihood would be the sample mean vector m_1 and sample covariance matrix Σ_1 when q_i = 0 and sample mean vector m_2 and sample covariance matrix when Σ_2 when q_i = 1. The values of q_i are unknown so one must proceed iteratively substituting the conditional expectation $E[q_i | h, X] = \Pr(q_i = 1 | h, X)$ at each observation. Thus in the so-called *Expectation* step $w_i = \Pr(q_i = 1 | h, X)$ and in the *Maximization* step

$$\hat{m}_1 = \frac{\sum_{i=1}^{n}(1-\hat{w}_i)x_i}{\sum_{i=1}^{n}(1-\hat{w}_i)}, \qquad \hat{\Sigma}_1 = \frac{\sum_{i=1}^{n}(x_i-\hat{m}_1)(1-\hat{w}_i)(x_i-\hat{m}_1)'}{\sum_{i=1}^{n}(1-\hat{w}_i)}$$

$$\hat{m}_2 = \frac{\sum_{i=1}^{n}\hat{w}_i x_i}{\sum_{i=1}^{n}\hat{w}_i}, \qquad \hat{\Sigma}_2 = \frac{\sum_{i=1}^{n}(x_i-\hat{m}_2)\hat{w}_i(x_i-\hat{m}_2)'}{\sum_{i=1}^{n}\hat{w}_i}$$

with mixing parameters $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} w_i$. Initial parameter guesses may be chosen at random and it is often advised to use multiple starting points as the *EM* cannot distinguish local from global maximum points.

Interestingly, the ascent properties of the EM can be explained using the ideas of entropy and relative entropy. Say the *EM* is used to classify observations of the random variable $X = x_i$ into one of *M* classes by maximizing the likelihood estimate of the model parameters h and the probability *p* that $X = x_i$ belongs to class z_i. By Baye's formula and using logarithms $\log p(x_i, z_i \mid h) = \log p(z_i \mid x_i, h) + \log p(x_i \mid h)$. For unknown parameters let $p(z_i \mid x, q)$ denote the probability of the class z_i given observation x_i under the unknown parameter set. Multiplying $\log p(x_i, z_i \mid h) = \log p(z_i \mid x_i, h) + \log p(x_i \mid h)$ by $p(z_i \mid x, q)$, summing over z_i and $i = 1,\ldots,n$ and using the fact that $\sum_{z_i} p(z_i \mid x_i, h) = 1$ then the log-likelihood for h is

$$l(h; X) = \sum_i \log p(x_i \mid q)$$
$$= \sum_i \sum_{z_i} p(z_i \mid x_i, q) \log p(x_i, z_i \mid h) - \sum_i \sum_{z_i} p(z_i \mid x_i, q) \log p(z_i \mid x_i, h).$$

The entropy of the hidden parameter q is

$h_q = \sum_i \sum_{z_i} p(z_i \mid x_i, q) \log p(z_i \mid x, q)$. Define

$$f(q, h) = \sum_i \sum_{z_i} p(z_i \mid x_i, q) \log p(x_i, z_i \mid h) - h_q$$
$$= \sum_i \sum_{z_i} p(z_i \mid x_i, q) \log(p(x_i \mid z_i, h)\, p(z_i \mid h)) - h_q$$

and the relative entropy

$$d(\theta,\eta) = \sum_i \sum_{z_i} p(z_i \mid x_i,\theta)\log\left(\frac{p(z_i \mid x_i,\theta)}{p(z_i \mid x_i,\eta)}\right)$$
$$= \sum_i \sum_{z_i} p(z_i \mid x_i,\theta)\log p(z_i \mid x_i,\theta) - p(z_i \mid x_i,\theta)\log p(z_i \mid x_i,\eta)$$
$$= h_\theta - \sum_i \sum_{z_i} p(z_i \mid x_i,\theta)\log p(z_i \mid x_i,\eta)$$

so that the log-likelihood can now be written $l(\eta;X) = f(\theta,\eta) + d(\theta,\eta)$. Now the ascent properties of the EM become manifest. $d(\theta,\eta)$ is nonnegative for all η and zero when $\eta = \theta$ so that $f(\theta,\eta) < l(\eta;X)$ for all η and equal when $\eta = \theta$. Now, if $f(\theta,\eta) > f(\theta,\theta)$ then $l(\eta;X) > l(\theta;X)$. If the relative entropy term $d(\theta,\eta)$ has a minimum at $\eta = \theta$ then $\partial l(\eta)/\partial\eta\,|_{\eta=\theta} = \partial f(\theta,\eta)/\partial\eta\,|_{\eta=\theta}$ so if $\eta = \theta$ is not a critical point for $l(\eta;X)$ then it is not for $f(\theta,\eta)$. Thus, starting with arbitrary parameter $\theta = \theta_0$ and iterating the *E* step and *M* step, $l(\eta;X)$ is guaranteed to converge.

Results Using Data

We use data generated for the purpose of understanding the gene funcions in a plant called *Arabidopsis thaliana* studied by a group of NYU biologists under the direction of Dr Gloria Coruzzi. The data consists of 21 replicated experiments and a total of 10 unreplicated experiments. There are 8297 probes in each experiment making for a gene expression matrix of 8297 X 10. Filtering removes up to 80% - 90% of the gene probes that are considered unchanged across all experiments by using a statistical t-test. The genes that are considered duly changed are clustered together. A correlation threshold of 0.85 is used to pick the clusters.

We have broken the complete list of experiments into one of two types. The first type makes no distinction and uses all 10 experiments. The second type uses only those experiments that are not deprived of light. For each type we use the methods presented in this thesis in order of their appearance. We start by ranking the importance of each experiment in the inducement of each cluster found. The gene expressions for each cluster are projected on to a two-dimensional rank 2 space. We

execute the *EM* algorithm choosing several different modes to describe the distribution of data in this space. Finally We draw posterior boundary conditions when the modes in the projected space are quite distinguishable. In the following charts, Exp*J* represents experiment *J* and C*I* cluster number *I*. Experiment list names the actual experiments used to cluster the data. Each name in the list denotes a treatment condition of either depriving or providing the plant with a light source, nutrient source, nitrogen source or carbon source . It is not important to know which names correspond to what treatment conditions, one need know only that none of the treatment conditions are the same.

Finally, in the principal component projection plots, I have estimated the mixture of bivariate normal distributions corresponding to the data. The black crosses denote the covariance estimates of the distributions while the red ellipses signify one std deviation from the means of the estimated distributions.

All Experiments (Type 1):

Experiment List:

'tAndrew117a' 't138' 't182T' 't188T' 't153-Control' 't0.1XN' 't184T' 't190T' 't154-Sucrose' t275'

Mutual Information matrix:

Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
C1	2.2122	1.6958	1.6551	1.7319	1.7518	1.7586	1.6724	1.6624	1.7330	1.7287
C2	2.1377	1.7966	1.7744	1.8253	1.9108	1.9109	1.9132	1.9044	1.9265	1.9201
C3	2.1116	1.7557	1.7575	1.8183	1.8692	1.8996	1.9005	1.8933	1.9377	1.9194
C4	1.9691	1.8064	1.7392	1.7584	1.9708	1.9681	1.7717	1.7768	1.8511	1.8493
C5	2.1516	1.8026	1.7744	1.8147	1.9230	1.9282	1.8937	1.8810	1.9349	1.9168

Information rank:

	1	2	3	4	5	6	7	8	9	10
C1	3	8	7	2	10	4	9	5	6	1
C2	3	2	4	8	5	6	7	10	9	1
C3	2	3	4	5	8	6	7	10	9	1
C4	3	4	7	8	2	10	9	6	1	5
C5	3	2	4	8	7	10	5	6	9	1

All experiments in the first two rankings of the information rank matrix, with the exception of experiment #8, have been deprived of light. Light or the absesnce of light can be inferred to be an overwhelming condition that has large effects on the gene expressions in each experiment

Principal Component projection

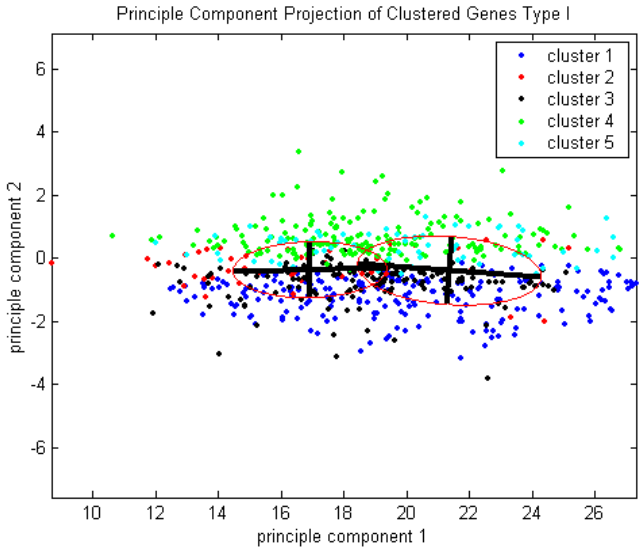

figure 3: PC projection, Type 1 experiment, EM with 2 modes

For experiment type 1, it does not look like there exist any naturally ocurring distinguishable modes in the principal component projection of the data. I have used a mixture of two Gaussians to represent the data but probably only one is necessary. One can note however that clusters 4 and 5 are mostly positive for the 2nd principal component while the others are negative. Correlating the principal components with

the experiments one finds that all those experiments deprived of light have negative correlation to the 2nd principal component while those that were treated with sufficient light correlated positively to the 2nd principal component.

All Experiments with light (Type 2):

Experiment list

't153-Control' 't0.1XN' 't184T' 't190T' 't154-Sucrose' 't275'

Mutual Information matrix:

	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6
C1	1.7213	1.7344	1.7422	1.7522	1.7884	1.7591
C2	1.9643	1.9971	1.9071	1.9073	1.9237	1.9174
C3	1.9556	1.9540	1.9053	1.9067	1.9189	1.9001
C4	2.0179	2.0272	1.9254	1.9256	1.9597	1.9498
C5	1.9349	1.9570	1.8970	1.8949	1.9359	1.9132

Information rank:

	1	2	3	4	5	6
C1	1	2	3	4	6	5
C2	3	4	6	5	1	2
C3	6	3	4	5	2	1
C4	3	4	6	5	1	2
C5	4	3	6	1	5	2

This experiment yields a more varied selection of experiments that are the most important for the inducement of these clusters.

Principal Component projection

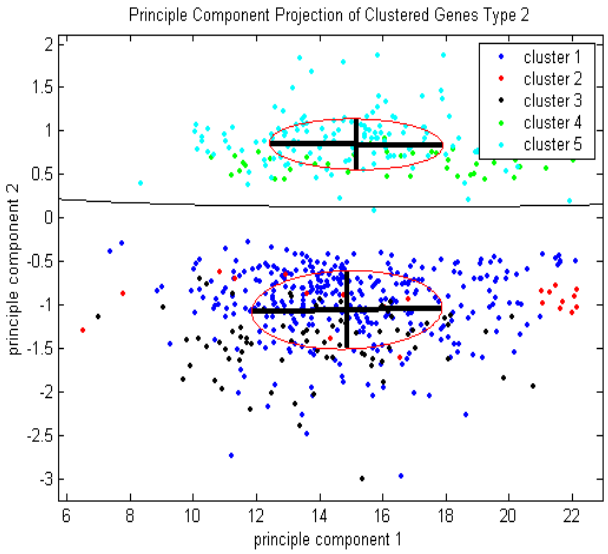

figure 4: PC projection, Type 2 Experiment, EM with 2 modes and Decision Boundary

Figure 4 shows a nice separation between the sets of clusters. The separation relies on the 2nd principal component as clusters 4 and 5 are mostly above zero and the others mostly below. A decision boundary is shown which divides the space into regions where it is most probable that the data is engendered from either one of the two modes. Correlating principal components to experiments one finds that all experiments that included a carbon source correlated positively to principal component two while those without a carbon source correlated negatively to the second principal component. Thus the groupings of data here may have something to do with the inclusion or exclusion of a carbon source.

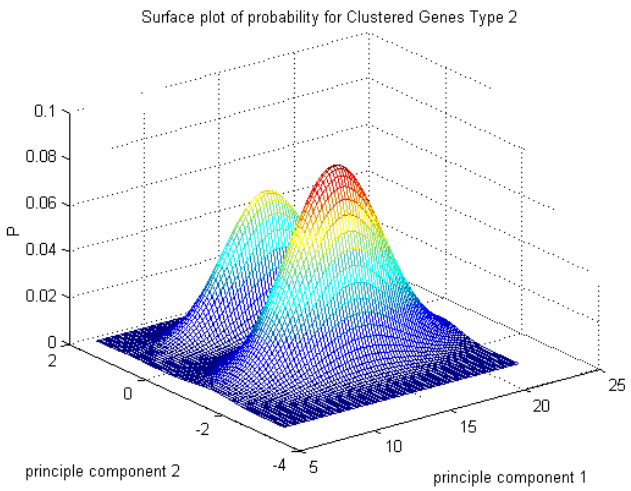

figure 5: Estimated Surface probability plots for figure 4

Finally, figure 5 represents the mixed Gaussian probability distribution of the data in figure 4.

Concluding Remarks

Microarrays and gene expression data present researchers with a powerful tool to examine the function of known genes at the cellular level. Discriminating biochemical pathways, identifying genes responsible for a particular phenotype or regulation of a set of functionally similar genes, classifying healthy cells from tumor cells, and distinguishing different types of tumors - each are applications of microarray technology. Using unsupervised learning procedures such as clustering to identify functional families of genes has been a widely used and putatively successful technique thus far. The goal of this thesis was to identify ways to find the salient experiments of each cluster and the possibility of distinguishing patterns of the data in an orthogonal subspace. Insofar as this is concerned, one needs a way of estimating such distributions which may be multi-modal. The EM method is a good way of doing this. It can also be used in conjunction with more sophisticated methods such as neural networks and projection pursuit to determine patterns in data or the best (or, most interesting insofar as the projections deviation from Gaussian) linear projections of the data, respectively. Such methods may be found in Friedman (2001) and Duda (2000). In the end, the analysis reduces to how quantitively we can state the effect subsets of treatments have on certain groups of genes and the precision of how quantitatively such effects can be determined.

References

Efron, B.and Tibshirani, R. (1993) An Introduction to the Bootstrap. Chapman and Hall.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998) "Cluster Analysis and Display of Genome-wide Expression Patterns". PNAS 95 pp.14863-68.

Dempster, A., Laird, N. and Rubin, D. (1977) "Maximum Likelihood From Incomplete Data via the EM Algorithm", *J.R. Statist. Soc. B*. vol 39 pp. 1-38.

Duda, R. and Hart, P. (2000). *Pattern Recognition,* Wiley.

Friedman, J., Hastie, T., Tibshirani, R. (2001) *The Elements Of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Papoulis, A. (1991) *Probability, Random Variables and Stochastic Processes,* McGraw-Hill.

Tavazoie S., Jason D. Hughes, Michael J. Campbell, Raymond J. Cho, and George M. Church. (1999) Systematic determination of genetic network architecture. *Nature Genetics*, Vol 22, 281-285

Silverman, B. (1982) "Density Estimation", Applied Statistics, vol. 31.

Shannon, C., Weaver, W. (1949) The Mathematical Theory of Communication, University of Illinis Press.

Thomas, J., Cover, T. (1991) *Elements of Information Theory*, Wiley.

Wong, W., Cheng Li, (2001). *Model Based Analysis of Olgonucleotide Arrays: Expression Index Computation and Outlier Detection*. PNAS, January, no 1, pp.31-36.

Wong, W., Cheng Li, "DNA Chip Analyzer", http://www.dchip.org

* All computation outside of clustering was executed with scripts and programs I've written in Matlab 6, C and Matlab-C interface code.