

Data Acquisition

The data acquisition software ArraySuite (Chen Y, Dougherty E, Bittner M. Ratio-Based Decisions and the Quantitative Analysis of cDNA Micro-array Images. *Journal of Biomedical Optics* 2: 364 (1997)), used by NIEHS to process pixel images of hybridized microarrays, computes the mean pixel intensity for each spot - The average of all pixel gray-levels within the target area, after trimming top and bottom 5%. A histogram of the pixel intensities from a region surrounding the target is constructed, where the mode represents where the fluorescent background is. The mean pixel intensity of the background region is used for local background adjustment.

Array Normalization

After pixel intensity determination and background subtraction, the ratio of the intensity of the 532nm and 635nm scans is calculated for each spot. The ratio intensity data from all the targets on the array (or a set of control genes if spotted on the array) is used to fit a probability distribution to the ratio intensity values and to estimate the normalization constant (m) that this distribution provides. The m provides a measure of the intensity gain between the two scans. Ratio intensity values are normalized to account for imbalance between the two scans by multiplying the ratio values by m . ArraySuite software is used for normalization (Chen, *et al.* 1997).

Determination of Altered Genes

Differentially expressed genes are detected using the Chen, *et al.* (1997) ratio confidence level approach implemented in ArraySuite software. The probability distribution (see above) is fit to the ratio intensity data and used to construct confidence intervals. Genes having normalized ratio intensity values outside of the limits set at a given confidence level are considered significantly differentially expressed.

Statistical Validation of Altered Genes

In the **MAPS** database system (Bushel, *et al.* MAPS: a microarray project system for gene expression experiment information and data validation, *Bioinformatics*. 2001 Jun;17(6):564-5) a binomial distribution is used to model the occurrence of stored differentially expressed genes at a given confidence level across multiple, independent hybridization trials.

$$P_{(k \text{ out of } n)} = \frac{n!}{k!(n-k)!} (p^k)(q^{n-k})$$

where:

n = the number of replicate array hybridizations;

k = the number of times that a gene is detected as altered across replicate hybridizations;

p = the confidence level used (divided by 100) to detect altered genes;

q = 1 - p.

The ratio intensity values from each hybridization are analyzed using a modified Z-score computation to detect outlying ratio intensity values. A coefficient of variation (CV) is also calculated using the log based 2 ratio intensity value of each gene detected as differentially expressed from hybridization trials. Genes with a CV greater than 0.4 and/or with ratio intensity values flagged as outliers using the modified Z-score are removed from the gene list which is used for subsequent higher order analyses.

$$M_i = \frac{0.6745(x_i - x_{\text{median}})}{\text{MAD}}$$

where:

$$\text{MAD} = \{ |x_i - x_{\text{median}}| \};$$

M_i = Modified Z-Score for the i th ratio value of the j th gene;

$$E(\text{MAD}) = 0.6745 \text{ std. dev.};$$

x_i = the i th ratio value for of the j th gene;

x_{median} = sample median of the ratio values of the j th gene.

Finding Groups in Data

Hierarchical Clustering

Grouping samples based on gene expression data is performed by agglomerative hierarchical clustering as has been implemented in **Cluster** (Eisen M, Spellman P, Brown P, and Botstein D. Cluster Analysis and Display of Genome-wide Expression Patterns PNAS 95:14863 (1998)) and visualized through **TreeView** (Eisen et al. 1998) in a heat map of the expression profiles and dendrogram (tree) to graphically represent the hierarchy of the similar groups of samples.

Basically, a clustering algorithm accepts data in a 2-dimensional matrix of ratio intensity values for each gene (attribute) of interest in the rows of the matrix and the samples/experiments (objects) of interest representing the columns of the matrix. Specific rows and columns of the matrix can be designated for normalizing genes and/or arrays as well as for arranging the presentation of clusters according a computational or user-defined ordering convention. Standardization of the genes and/or arrays can be performed by mean or median centering of the data or other statistical methods.

Depending on the type of similarity (or dissimilar) coefficient selected, for example:

Pearson Correlation coefficient:

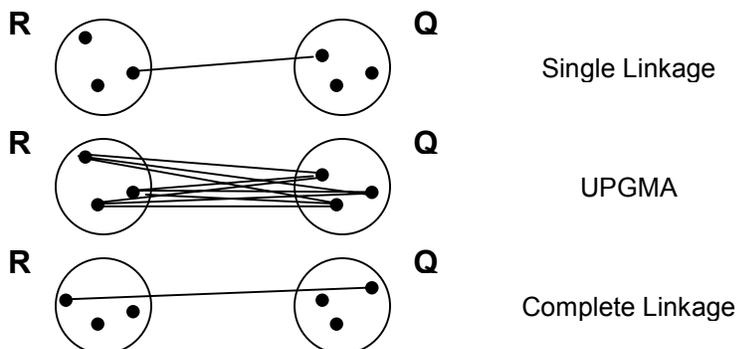
$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N}) (\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

or

Euclidean Distance coefficient:

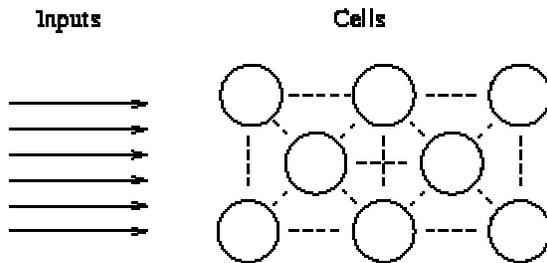
$$c = (a^2 + b^2)^{1/2}$$

a resemblance matrix is generated measuring the overall degree of similarity between each pair of objects. The correlation coefficient is generally used to cluster arrays or genes based on similar expression profiles whereas the distance coefficient is often used to cluster arrays or genes according to similar expression levels. Finally, an agglomerative nesting algorithm is used to iterate through the resemblance matrix for incremental merging of the most similar clusters until there is only one group left. The single linkage (nearest neighbor) method defines the distance between two clusters as the smallest dissimilarity between an object in cluster *R* and an object in cluster *Q*, the unweighted pair-group method using arithmetic averages (UPGMA) takes the average of all dissimilarities between object *i* in cluster *R* and object *j* in cluster *Q*, and the complete linkage (furthest neighbor) method uses the largest dissimilarity between an object in cluster *R* and an object in cluster *Q*.



Self Organizing Maps

Self Organizing Maps (Kohonen T, Self Organizing Maps Springer (2001)) is a pattern recognition and dimension reduction algorithm developed to project and visualize high-dimensional data in 2-dimensional space. A SOM can be thought of as a mapping of multi-dimensional input data onto elements of a 2 dimensional array of interconnected nodes or cells to aid in the exploration of data.



The network of nodes are tuned to the input signals or classes of patterns in an orderly fashion. SOMs were first used by Tamayo et al. (1999) in a software called **GeneCluster** to partition gene expression data. Simply, given an input vector: $x(t)$ element of R^n and model vector: $m_i(t)$ element of R^n for node i in the neural network where t is an integer-value index, the representation of an input vector on the map is defined as the array element m_c (winner node) that best matches with x defined by:

$$c = \arg \min_i \{d(x, m_i)\}.$$

The distance measure $d(x, m_i)$ is defined over all x items and a large set of models m_i . Before recursive partitioning, m_i is initialized by randomly selecting values from the input samples. During the learning process, the nodes that are topographically close in the array up to a certain geometric distance, will activate each other to learn from the same input data of the winning node. This leads to a smoothing/updating of the weight vectors of the nodes in the vicinity of the winner so that they are more similar to the input vector. Performing 1,000 to 100,000 epochs of this process for all input vectors results in overall ordering of patterns.

Principal Component Analysis

Principal component analysis (PCA) is concerned with capturing the variance within data using linear combinations of the system random variables for data reduction and interpretation purposes. Although p principal components (PCs) explain the total variability in the data, a small number of (k) PCs can account for much of the variability in the system. Therefore, k PCs can substitute for the original p random variables permitting the data set consisting of n measurements on p random variables to be reduced to one containing n measurements on k PCs.

Algebraically, PCs are linear combinations (weights) of the random variables X_1, X_2, \dots, X_p and geometrically, these PCs represent the selection of new orthogonal axes (obtained by rotating the data with X_1, X_2, \dots, X_p as the coordinate axes) that represent the direction through the data with the maximum variation.

$$PC_1 = w_{11}X_1 + w_{12}X_2 + \dots + w_{1p}X_p$$

$$PC_2 = w_{21}X_1 + w_{22}X_2 + \dots + w_{2p}X_p$$

.

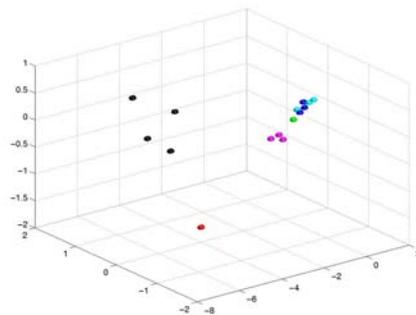
.

.

$$PC_p = w_{p1}X_1 + w_{p2}X_2 + \dots + w_{pp}X_p$$

The first PC (PC_1) accounts for the maximum variation in the data, the second PC (PC_2) accounts for the maximum variance that has not been accounted for by PC_1 and so on.

Plotting PC scores of the observations in 3-dimensions is useful for visual examination and interpretation of clustered groups within the data.



Finally, the PC scores can also be utilized as input variables in other multivariate techniques such as clustering algorithms, regression and discriminant analysis.

Discriminant Analysis (LDA and QDA)