PATTERN DISCOVERY IN GENE EXPRESSION DATA

Gráinne Kerr
Biocomputation Research Lab., The School of Computing, Dublin City University,
Glasnevin, Dublin 9, Ireland

E-mail: GRAINNE.KERR@COMPUTING.DCU.IE
Telephone: +353 1 700 8449
Fax: +353 1 700 5442

Heather Ruskin
Biocomputation Research Lab., The School of Computing, Dublin City University,
Glasnevin, Dublin 9, Ireland

E-mail: HEATHER.RUSKIN@COMPUTING.DCU.IE
Telephone: +353 1 700 5513
Fax: +353 1 700 5442

Martin Crane
Biocomputation Research Lab., The School of Computing, Dublin City University,
Glasnevin, Dublin 9, Ireland

E-mail: MARTIN.CRANE@COMPUTING.DCU.IE
Telephone: +353 1 700 8974
Fax: +353 1 700 5442

PATTERN DISCOVERY IN GENE EXPRESSION DATA

ABSTRACT

Microarray technology[i] provides an opportunity to monitor mRNA levels of expression of thousands of genes simultaneously in a single experiment. The enormous amount of data produced by this high throughput approach presents a challenge for data analysis: to extract meaningful patterns, to evaluate its quality and to interpret the results. The most commonly used method of identifying such patterns is cluster analysis. Common and sufficient approaches to many data-mining problems, e.g. Hierarchical, K-means, do not address well the properties of "typical" gene expression data and fail, in significant ways, to account for its profile. This chapter clarifies some of the issues and provides a framework to evaluate clustering in gene expression analysis. Methods are categorised explicitly in the context of application to data of this type, providing a basis for reverse engineering of gene regulation networks. Finally, areas for possible future development are highlighted.

Keywords: Bi-clustering, Clustering, Gene expression, Microarray, Data Analysis, Data Mining, Data Mining Algorithms.

INTRODUCTION

A fundamental factor of function in a living cell is the abundance of proteins present at a molecular level, i.e. its *proteome*. The variation between proteomes of different cells is often used to explain differences in phenotype and cell function. Crucially, gene expression is the set of reactions that controls the level of messenger RNA (mRNA) in the *transcriptome*, which in turn maintains the proteome of a given cell. The transcriptome is never synthesized *de novo*; instead, it is maintained by gene expression replacing mRNA's that have been degraded, with changes in composition brought about by switching different sets of genes on and off. To understand the mechanisms of cells, involved in a given biological process, it is necessary to measure and compare gene expression levels in different biological phases, body tissues, clinical conditions and organisms. Information on the set of genes expressed, in a particular biological process, can be used to characterise unknown gene function, identify targets for drug treatments, determine effects of treatment on cell function, and understand molecular mechanisms involved.

DNA microarray technology has advanced rapidly over the past decade, although the concept itself is not new (Friemert et al., 1989; Gress et al., 1992). It is now possible to measure the expression of an entire genome simultaneously, (equivalent to the collection and examination of data from thousands of single gene experiments). Components of the system technology can be divided into: (1) Sample preparation, (2) Array generation and sample analysis and (3) Data handling and interpretation. The focus of this chapter is on the third of these.

Microarray technology utilises base-pairing hybridisation properties of nucleic acids, whereby one of the four base nucleotides (A, T, G, C) will bind with only one of the four base ribonucleotides (A, U, G, C: pairing = A – U, T – A, C – G, G - C). Thus, a unique sequence of DNA that characterises a gene will bind to a unique mRNA sequence. Synthesized DNA molecules, complementary to known mRNA, are attached to a solid surface, referred to as *probes*. These are used to measure the quantity of specific mRNA of interest that is present in a sample (the *target*). The molecules in the target are labelled, and a specialised scanner is used to measure the amount of hybridisation (intensity) of the target at each probe. Gene intensity values are recorded for a number of microarray experiments typically carried out for targets derived under various experimental conditions (Figure 1). Secondary variables (covariates) that affect the relationship between the dependent variable (experimental condition) and independent variables of primary interest (gene expression) include e.g. age, disease, and geography among others, and can also be measured.

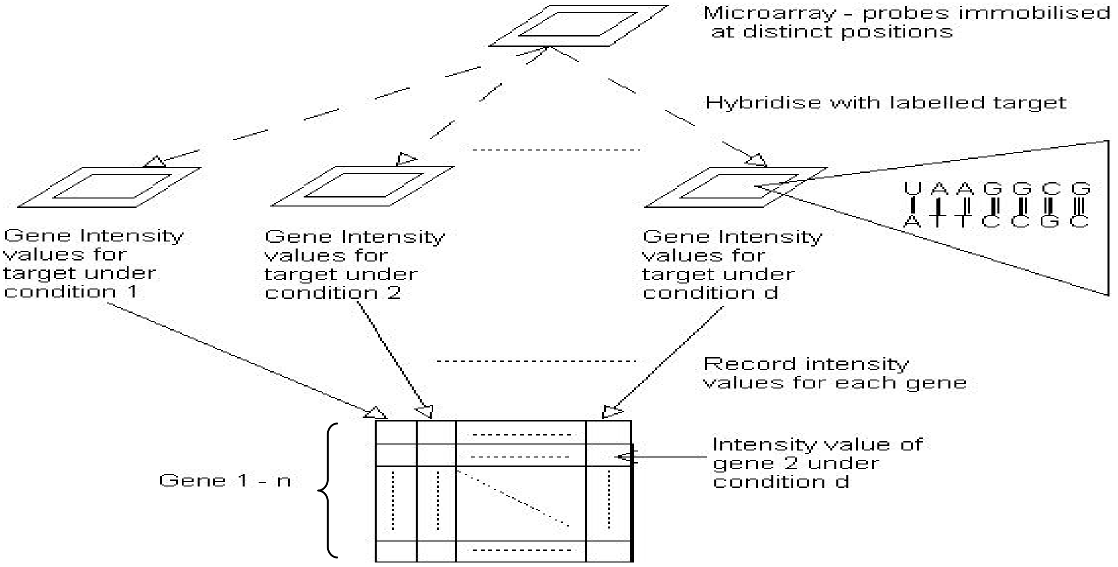

Microarray - probes immobilised at distinct positions
Hybridise with labelled target
U A A G G C G
A T T C C G C
Gene Intensity values for target under condition 1
Gene Intensity values for target under condition 2
Gene Intensity values for target under condition d
Record intensity values for each gene
Gene 1 - n
Intensity value of gene 2 under condition d

Figure 1 - mRNA is extracted from a transcriptome of interest, (derived from cells grown under precise experimental conditions). Each mRNA sample is hybridised to a reference microarray. The gene intensity values for each experiment are then recorded.

An initial cluster analysis step is applied to gene expression data to search for meaningful informative patterns and dependencies among genes. These provide a basis for hypothesis testing - the basic assumption is that genes, showing similar patterns of expression across experimental conditions, may be involved in the same underlying cellular mechanism. For example, Alizadeh et al. (2000) used a hierarchical clustering technique, applied to gene expression data derived from diffuse large B-cell lymphomas (DLBCL), to identify two molecularly distinct subtypes. These had gene expression patterns, indicative of different stages of B-cell differentiation, – germinal centre B-like DLBCL and activated B-like DLBCL. Findings suggested that patients, with germinal centre B-like DLBCL, had a significantly better overall survival rate than those with activated B-like DLBCL. This work indicated a significant methodology shift towards characterisation of cancers *based on gene expression*, rather than morphological, clinical and molecular variables.

BACKGROUND

The Gene Expression Dataset

Data are typically presented as a real-valued matrix, with rows representing the expression of a gene over a number of experiments, and columns representing the pattern of expression of all genes for a given microarray experiment. Each entry x_{ij} is the measured expression of a gene *i* in experiment *j*, (Figure 1). The following terms and notations are used throughout this chapter:

- A gene/gene profile *x* is a single data item (feature vector) used by the clustering algorithm. It consists of *d* measurements, $x = (x_1, x_2, \ldots x_d)$.
- A condition *y* is a single microarray experiment corresponding to a single column in the gene expression matrix, $y = (x_1, x_2, \ldots x_n)^T$, where *n* is the number of genes in the dataset.
- The individual scalar components of each gene vector x_{ij} represent the measured expression of gene *i* under experimental condition *j*.

There are a number of publicly available dataset repositories, which contain a wealth of microarray datasets[ii]: Table 1 provides a sample of these. Typically, these repositories store data using the 'Minimum Information About Microarray Experiment' (MIAME) standard (Brazma et al., 2001), which allow researchers to replicate the experiments. This allows analysts to compare gene expression data from different laboratories effectively, based on information about the microarrays used in experiments, how these were produced, samples obtained and mRNA extracted and labelled. Additional information is also recorded on methods used to hybridise the sample, scan the image and normalise the data.

Database	**Description**	**URL**
ArrayExpress	Gene expression and hybridisation array data repository	http://www.ebi.ac.uk/arrayexpress/#ae-main[0]
CellMiner	Data from 60 cancer cell lines based on Affymetrix and cDNA microarray data	http://discover.nci.nih.gov/cellminer
ExpressDB	Collection of E. Coli and Yeast RNA expression datasets	http://arep.med.harvard.edu/ExpressDB/
GEO	Gene expression and hybridisation array data repository	http://www.ncbi.nlm.gov/geo/
RAD	Gene expression and hybridisation array data repository	http://www.cbil.upenn.edu/RAD/
SMD	Extensive collection of microarray data	http://genome-www.stanford.edu/microarray

Table 1 – Selection of publicly available dataset repositories.

Characteristics of the Gene Expression Dataset:

Choice of the appropriate clustering technique relies on the amount of information on the particular properties of gene expression data available to the analyst, and hence the likely underlying structure. The following data characteristics are typical of the gene expression dataset:

Measurement accuracy of mRNA expression levels depends on the experimental design and rigour. While design of experiments is not a specific focus of this chapter, a good design minimises variation and has a focused objective, (Kerr and Churchill, 2001). *Technical variation* between microarray slides depends on numerous factors including experimental technique, instrument accuracy for detecting signals, and observer bias. *Biological variation* may arise due to differences in the internal states of a population of cells, either from predictable processes, such as cell cycle progression, or from random processes such as partitioning of mitochondria during cell division, variation due to subtle environmental differences, or ongoing genetic mutation, (Raser and O'Shea, 2005). *Pre-processing techniques* attempt to remove technical variation while maintaining interesting biological variation.

Many variables, both random and fixed, (biological and technical), are associated with microarray measurements. Data is thus ***intrinsically noisy*** and outliers in the dataset need to be identified and managed effectively. This usually takes one of two forms, (i) outlier accommodation; uses a variety of statistical estimation or testing procedures, which are robust against outliers, (ii) identification and decision on inclusion/exclusion, used when outliers may contain key information (Liu et al., 2002). *Normalisation procedures* applied to gene expression data (Bolstad et al., 2003), aim at minimising the effect of outliers, (assuming these to be due to experimental variation and thus undesirable). Most manufacturers of microarrays, aware of effects of optical noise and non-specific binding, include features in their arrays to measure these directly: these measurements can be used in the normalisation procedures. Note: Although pre-processing methods attempt to remove all noise these may be only partially successful.

Missing values are common to microarray data, and can be caused by insufficient resolution in image analysis, image corruption, dust or scratches on the slide, robotic method used to create the slide and so on, (Troyanskaya et al., 2001). In general, the number of missing values increases with the number of genes being measured. Many clustering algorithms, used for gene expression data, require a complete matrix of input values. Consequently, imputation or missing data estimation techniques need to be considered in advance of clustering. The effect of missing data on pattern information can be minimised through *pre-processing*.

Commonly, missing values in the gene expression matrix are replaced by zeroes or by an average expression level of the gene, (or "row average"). Such methods do not, however, take into account the correlation structure of the data and more sophisticated options include K-Nearest Neighbour (KNN) and Support Vector Decomposition type methods. Troyanskaya et al. (2001) note that KNN and SVD-based methods are more effective than traditional methods of replacement, with KNN being more robust as the number of missing values increases.

Clustering algorithms that ***permit overlap (probabilistic or fuzzy clusters)*** are typically more applicable to gene expression data since: (i) the impact of noisy data on clusters obtained is a fundamental consideration in algorithm choice. (The assumption is that "noisy genes" are unlikely to belong to any one cluster, but are equally likely to be members of several clusters): (ii) the underlying principal of clustering gene expression data, is that genes with similar change in expression for a set of conditions are involved, together, in a similar biological function. Typically, gene products (mRNA) are involved in several such biological functions and groups need not be co-active under all conditions. This gives rise to high variability in the gene groups and/or some overlap between them. For these reasons, constraining a gene to a single cluster (*hard clustering*) is counter-intuitive with respect to natural behaviour.

Additionally, methods that aim at a ***partial clustering*** tend to be more suited to expression data, with some genes or conditions not members of any cluster (Maderia and Oliveira, 2000). Clustering the microarray dataset can be viewed in two ways: (i) genes can form a group which show similar expression across conditions, (ii) conditions can form a group which show similar gene expression across all genes. It is this interplay of conditions and genes that gives rise to bi-clusters, whereby conditions and genes are simultaneously grouped. Such partial clusterings, (or *bi-clusters*), are defined over a subset of conditions and a subset of genes, thus capturing local structure in the dataset. Clearly, this allows: (i) “noisy genes” to be left out, with correspondingly less impact on the final outcome, (ii) genes belonging to no cluster – omitting a large number of irrelevant contributions, (iii) genes not belonging to well-defined groups. (Microarrays measure expression for the entire genome in one experiment, but genes may change expression, independent of the experimental condition, (e.g. due to stage in cell cycle). *Forced inclusion* of such genes in well-defined but inappropriate groups may impact the final structures found for the data).

Methods of Identifying Groups of Related Genes:

Cluster definition is dependent on clearly defined metrics, which must be chosen to reflect the data basis. Metric categories include:

Similarity-based

The cluster is defined to be the set of objects in which each object is closer, (or more similar), to a prototype that defines that cluster as opposed to any other cluster prototype. A typical gene expression cluster prototype is often the *average* or *centroid* of all gene vectors in the cluster. The *similarity metric* used affects the cluster produced. Common measures include: (i) Euclidean distance, (ii) Manhattan distance, (iii) Squared Pearson correlation distance, (Quakenbush, 2001) with the last being the most popular as it captures gene expression “shape” without regard to the magnitude of the measurements. However, this distance measurement is quite sensitive to outliers, although, correlation, rather than “distance”, is inherently more important for gene expression data. Take, for example, two gene vectors $X_1=(1,2,3,4,5)$ and $X_2=(3,6,9,12,15)$. These two profiles result in a Euclidean distance of 14.8323 and a Manhattan distance of 30. The Pearson correlation distance however is 0, reflecting the fact that the two genes are showing the same patterns of expression.

Density-based

Clusters, in this instance, are based on dense regions of genes, surrounded by less-dense regions. Such methods are often employed when clusters are irregular or intertwined, and when noise and outliers are present (Sander et al., 1998). However, as each cluster is assumed to have a *uniform* density, the method is not readily applicable to gene expression data, as some biological functions involve more gene products than others. The high dimensionality also means that density thresholds can be difficult to define and expensive to compute.

Model-based

Despite the convenience of similarity-based measures, it can be biologically meaningless to characterise a cluster through a cluster prototype, such as the mean or centroid, as these may be poorly representative of the cluster elements as a whole. As a typical gene expression dataset is large, noisy distortion of these prototypes may be considerable, resulting in relatively uninformative structures. In contrast, model-based techniques, applied to expression space, consider the "fit" of genes in a given cluster to the "ideal" cluster. Concentrating on the strengths of the bi-clustering approach, and following notation from Maderia and Oliveira (2004), four types of model can be identified:

i. Bi-clusters with constant values.

   A perfect cluster is a sub-matrix *(I,J)* of the gene expression matrix *(N,D)*, with all values equal, $x_{i,j} = \mu$. The ideal bi-cluster is, of course, rarely found in noisy gene expression data.

ii. Bi-clusters with constant values on rows or columns.

   A subset of the "ideal" or constant bi-cluster model, and one which is more realistic for gene expression data is a sub-matrix with constant rows or columns. For the former, rows have constant value in a sub-matrix *(I,J)* given by $a_{ij} = \mu + \alpha_i$ or $a_{ij} = \mu \times \alpha_i$, where μ is the "typical" bi-cluster value and α_i is the row offset for $i \in I$. Similarly, perfect bi-clusters with constant columns can be obtained for $a_{ij} = \mu + \beta_j$ or $a_{ij} = \mu \times \beta_j$, where $j \in J$.

iii. Bi-clusters with coherent values.

   From *(ii)*, a combined additive model can be derived. In this framework, a bi-cluster is a sub-matrix *(I,J)*, with coherent[iii] values, based on the model:

$$a_{ij} = \mu + \alpha_i + \beta_j \qquad \textbf{Eq. 1}$$

   (where μ, α_i and β_j are as for *(ii)*). Similarly, the multiplicative model assumes that a perfect bi-cluster could be identified using $a_{ij} = \mu' \times \alpha'_i \times \beta'_j$. Note: the additive form clearly follows for $\mu = log(\mu')$, $\alpha_i = log(\alpha'_i)$ and $\beta_j = log(\beta'_j)$.

   The artificial example in Figure 2(a) and (b) illustrates this point. The sub-matrix is an ideal bi-cluster found in a fictional dataset, where $\mu=1$, the offset for row 1 to 3 is $\alpha_1=0$, $\alpha_2=2$, $\alpha_3=4$ respectively, and the offset for columns 1 to 6 is $\beta_1=0$, $\beta_2=1$, $\beta_3=2$, $\beta_4=4$, $\beta_5=1$, $\beta_6=-1$ respectively. The expression levels can be obtained from Eq. 1. Of course, when searching the dataset for a fit to this model, the mean and offset parameters are unknown and must be estimated from the data. The schematic illustrates the coherent expression profile over the six conditions. Similarly for the multiplicative model where $\mu=1$, $\alpha_1=1$, $\alpha_2=2$, $\alpha_3=5$, $\beta_1=1$, $\beta_2=2$, $\beta_3=4$, $\beta_4=6$, $\beta_5=3$, $\beta_6=1.5$.

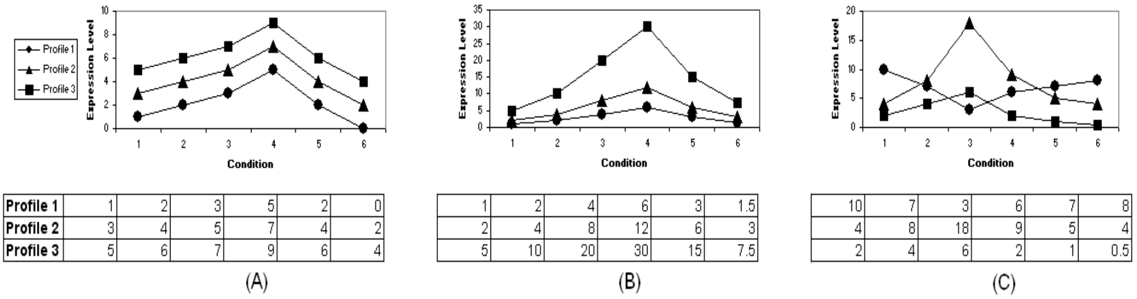
In reality, these "perfect" bi-clusters are, of course, unlikely to occur, so each entry in the sub-matrix can be regarded as having a *residue component* (Cheng and Church, 2000):

$$r_{ij} = \mu + \alpha_i + \beta_j - a_{ij}\,. \qquad \textbf{Eq. 2}$$

Thus, finding bi-clusters is equivalent to finding sub matrices that minimise the average residue.

iv. Bi-clusters with coherent evolution.

Local structures, with *coherent evolution* across a sub-matrix *(I,J)*, can exist in the data regardless of the exact values. This occurs if there is a pattern of co-regulation for a subset of genes and conditions. Expression can occur at different levels, so e.g. if two genes are up-regulated by different degrees, (e.g. due to a specific condition), these are said to experience coherent evolution. Taking Figure 2(c) as an example. Gene 1 and gene 2 are regulated, with similar periodicity, while gene 3 shows alternated periodicity. Although the genes are expressed at different levels, each change in expression level is triggered by the same condition. In a simple form, each gene can be said to be exhibiting three states, down-regulated, up-regulated or no change. Additional states can be used, e.g. strongly up-regulated, weakly up-regulated etc. depending on the detail of the model required. Adding additional states, of course, adds complexity to the model, and cut-off points between states of regulation must be considered carefully. The problem then reduces to finding profiles that show consistent patterns of regulation across all conditions.



Profile 1	1	2	3	5	2	0
Profile 2	3	4	5	7	4	2
Profile 3	5	6	7	9	6	4

1	2	4	6	3	1.5
2	4	8	12	6	3
5	10	20	30	15	7.5

10	7	3	6	7	8
4	8	18	9	5	4
2	4	6	2	1	0.5

Figure 2 - Models in gene expression datasets. The matrix gives clusters found, where rows are gene expression values across 6 experimental conditions (columns). X-axis indicates experimental condition or time point, y-axis indicates gene expression level. Model forms are (a) Additive for rows and columns, (b) Multiplicative for rows and columns and (c) Coherent evolution.

CLUSTER ANALYSIS

Current Methods

With extensive choice of metric, structure, completeness etc. in cluster analysis it is useful to consider a framework (Table 2) for performance comparison. The taxonomy used is due to Jains et al. (1999).

COMMON CLUSTERING TECHNIQUES				
	Gene Membership	Cluster Structure	Cluster Type	Complete/Partial
Hierarchical (Eisen et al., 1998)	Hard	Hierarchical (nested)	Similarity-Based	Complete
K-Means (Tavazoie et al., 1999)	Hard	No structure	Similarity-Based	Complete
FCM (Gasch and Eisen, 1999)	Fuzzy	No structure	Similarity-Based	Complete
SOM (Golub et al., 1999)	Hard	Topological Structure	Similarity and Neighbourhood kernal function-based	Complete
Delta clusters (Cheng and Church, 2000)	Shared	Overlap	Based on Coherent Additive Model	Partial
FLOC (Yang et al., 2003)	Shared	Overlap	Based on Coherent Additive Model	Partial
SAMBA (Tanay et al., 2002)	Shared	Overlap	Based on Coherent Evolution Model	Partial

Table 2 - Popular clustering techniques applied to gene expression data. Partial (overlapping) clusters are more relevant in this context.

Hierarchical methods:

Ever since the landmark paper of Eisen et al. (1998), numerous clustering algorithms have been applied to gene expression data. Predominantly these have been hierarchical methods, (Wen et al., 1998; Khodursky et al., 2000; Higgins et al., 2003; Makretsov et al., 2004), due mainly to ease of implementation, visualisation capability and general availability.

The basic steps of a hierarchical clustering algorithm include: (i) computation of the proximity matrix of distances between each gene, (initially each is in a unique cluster of size one), (ii) searching the proximity matrix for the two closest clusters, (iii) merging these two clusters and updating the proximity matrix, (iv) repeating steps two and three until all genes are in one cluster.

Such *agglomerative* clustering techniques vary with respect to the (i) distance metric used and the decision on cluster merger (i.e. linkage choice as single, complete, average or centroid; see Quackenbush (2001)). Typically output of a hierarchical clustering algorithm is a dendogram, representing nested patterns in the data and the similarity level at which clusters are merged. The choice of parameters affects both structure of, and relationship between the clusters. Hierarchical cluster structure works well for situations where *membership is crisp*, but, despite their popularity these methods may *not be appropriate to capture natural structures in gene expression data.*

Nevertheless, some successes of clustering *conditions* based on gene expression have been reported. For example, Makrestov et al. (2004) used gene expression profiles, to determine whether sub-types of invasive breast cancer could be identified, with a view to improving patient prognosis. Hierarchical clustering successfully identified three cluster groups with significant differences in clinical outcome. Similarly, a study on renal cell carcinoma, Higgins et al. (2003), found that hierarchical clustering led to segregation of "histologically distinct tumour types solely based on their gene expression patterns" (p. 925). These studies indicate that characterisation of tumours is potentially viable from gene expression profiling.

Hierarchical clustering algorithm properties include location of complete clusters, forced membership and large time-space complexity, but inclusion of "noisy genes" in the cluster can affect the final grouping, (depending to a greater or lesser extent on the linkage method and the distance measure used). As algorithms are prototype-based, further iterations exacerbate noise effects. Given the distance metric basis, hierarchical techniques also tend to produce globular structures.

Partitive Methods:

In contrast to hierarchical algorithms, which create clusters in a bottom up fashion resulting in nested levels of clustering, partitive methods optimise a function of given criteria, partitioning the entire dataset and obtaining one cluster structure.

Partitive K-Means clustering (MacQueen, 1967) produces *hard clustering* with no structural relationship between the individual clusters. The main steps of the K-means algorithm are: (i) Identification *K* prototype vectors for *K* clusters in the dataset. (ii) Assignment of each gene to a cluster based on its similarity to the cluster prototype, (iii) computation of cluster prototypes based on current genes in the cluster, (iv) repeating steps two and three until *convergence criteria* are satisfied. These may be e.g. no (or minimal) reassignment of genes to new clusters or e.g. minimal improvement in optimisation of the criteria function. A typical optimisation approach is to minimise the squared error within a cluster:

$$C = \sum_{j=1}^{k} \sum_{i=1}^{n} y_{ij} d(x_i, q_j) \qquad \textbf{Eq. 3}$$

where q_j is the vector representing the mean of the cluster, x_i is the vector representing the gene, $d(x_i,q_j)$ is a distance measure and y_{ij} is a partition element. Here $y_{ij} \in \{0,1\}$, and $y_{ij}=1$, indicates that gene *i* is assigned to cluster *j*.

An example of use of the K-means method is discussed in Tavazoie et al. (1999), and is based on a yeast time-course gene expression dataset, containing profiles for more than 6000 genes, with 15 time points (at 10 minute intervals - over nearly two cell cycles), (Cho et al., 1998). (This work succeeded in identifying transcriptional co-regulated genes in yeast). Unfortunately, initial prototype vectors in K-Means usually have a large impact on the data structures found. Prototype vectors are often genes selected at random from the dataset. Alternatively, Principal Component Analysis can be used to project the data to a lower dimensional sub-space and K-means is then applied to the subspace (Zha et al. 2002). Whichever method is used in practice to select prototype vectors, it is usually the case that different initial prototypes are investigated to assess stability of the results, with the best configuration, (according to the optimisation criteria), used as output clusters.

Fuzzy Methods:

As observed, (Section *Characteristics of the Gene Expression Dataset*), multiple cluster membership is more appropriate for gene expression data. The Fuzzy C-Means (FCM) algorithm extends the standard K-means algorithm, to the case where each gene has a membership degree indicating its "fuzzy" or percentage association with the centroid of a given cluster. Typically, each gene has a total membership value of 1, which is divided proportionally between clusters according to its similarity with the cluster means. A *fuzzy partition matrix Y,* (of dimension *NK*, where *K* is the number of clusters and *N* is the number of genes), is created, where each element y_{ij} is the membership grade of gene *i* in cluster *j* and a weighted version of Eq. 3 applies. At each iteration, the membership value, y_{ij}, and the cluster center, k_j, is updated by:

$$y_{ij} = 1 \Big/ \sum_{c=1}^{k} \left(\frac{d(x_i - k_j)}{d(x_i - k_c)} \right)^{\frac{2}{m-1}} \qquad \textbf{Eq. 4}$$

$$k_j = \sum_{i=1}^{N} y_{ij}^{m} x_i \Big/ \sum_{i=1}^{N} y_{ij}^{m} \qquad \textbf{Eq. 5}$$

where m>1 denotes the degree of fuzziness, (everything else is as for Eq. 3). The iterations stop when $max| j_{ij}^{k+1} - j_{ij}^{k} |< \varepsilon$, where ε is a *termination criterion* with value between 0 and 1, and *k* is the number of iterations.

Given the usual constraint that membership values of a gene must sum to unity, these values should be interpreted with care. A large "membership value" does not indicate "strength of expression" but rather reduced co-membership across several clusters, (Krishnapuram and Keller, 1993). Table 3 illustrates this idea for three clusters. FCM was carried out on published yeast genomic expression data, (Gasch and Eisen, 2002), (results available at http://rana.lbl.gov/FuzzyK/data.html). The membership values for gene B and gene D are very different for cluster 21, although they are approximately equidistant from the centroid of the cluster. Similarly, gene C and gene D have comparable membership values for cluster 4. However, gene C is more "typical" than gene D. With similar centroid distance measures, membership value for gene B in cluster 21 is smaller than membership value of gene A in cluster 46. These values arise from the constraint that membership values must sum to unity across all clusters, forcing a gene to give up some of its membership in one cluster to increase it in another. Listing the genes of a cluster, based on membership values alone is somewhat non-intuitive as it is not a measure of their compatibility with the cluster. However, if interpretation of the list in terms of degree of sharing between clusters is of value.

GID	Cluster 4		Cluster 21		Cluster 46	
	Centroid Dist.	Mem.	Centroid Dist.	Mem.	Centroid Dist.	Mem.
GENE1649X	10.691	0.002575	8.476	0.002002	3.864	0.482479
GENE6076X	6.723	0.009766	3.855	0.009341	6.33	0.007381
GENE5290X	6.719	0.007653	5.29	0.00515	8.024	0.005724
GENE2382X	7.725	0.007609	3.869	0.01782	6.279	0.010249

Table 3 - Difficulties interpreting membership values for FCM. GENE1649X (Gene A), GENE6076X (Gene B), GENE5290X (Gene C) and GENE2382X (Gene D). The table highlights distance to cluster centroid, in terms of Euclidean distance, and the associated membership values of the gene.

The work of Gasch and Eisen (2002) on the use of FCM in analysing microarray data looked at clustered responses of yeast genes to environmental changes. Groups of known functionally co-regulated genes, and novel groups of co-regulated genes, were found by this method, although missed by both hierarchical and K-means methods.

Artificial Neural Networks

Artificial neural networks (ANN) mimic the idea of biological neural networks, where links between various neurons (nodes) can be strengthened or weakened through learning. A number of ANN types have been explored, with Self-Organising Maps (SOM) (Kohonen, 1990) proving popular for the analysis of gene expression, as these provide a fast method of visualising and interpreting high dimensional data. The network *maps the high-dimension input gene vector into a lower dimensional space.* A SOM is formed by an input layer of *D* nodes, (where *D* is the gene vector dimension), and an output layer of neurons arranged in a regular grid (usually of 1 or 2 dimensions). A vector, of the same dimension as the input gene, references each node in the output layer. Briefly, the mechanism involves: (i) Initialisation of the prototype vectors of the output nodes, (ii) Training the network to find clusters in the data. (Genes are selected at random from the dataset and the closest output neuron is identified by its prototype vector. Once an output neuron is identified, its topological neighbours are updated to reflect this. Training continues until the reference vectors satisfy a stopping criterion). (iii) Sequential application of all gene vectors to the SOM, where only one output neuron 'fires' upon receiving an input gene vector. "Members" of a cluster, represented by output neuron *i*, are the set of genes, applied to the input neurons, causing output neuron *i* to fire.

ANN techniques have been used in a number of gene expression studies, including Tamayo et al. (1999) (to analyse haematopoietic differentiation), Toronen et al. (1999) (to analyse yeast gene expression data) and Golub et al. (1999) (to cluster acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML)). The stopping criterion of the SOM is crucial, since over-fitting to the training dataset is a risk. A further disadvantage is the amount of prior information needed. SOM requires input parameters such as learning rate, neighbourhood size and kernel function, as well as topology of the map, (typically hexagonal or square). The stability of results is also an issue, as a particular gene vector can be found to cause different output nodes to fire at different iterations, (Jains et al. 1999). Furthermore, clusters produced by the SOM are sensitive to choice of initial vectors for the output neurons, and a *sub-optimal structure* may result from a poor selection.

Search Based Methods:

While methods considered so far focus on finding *global structures* in the data, local structures are frequently of great interest. Cheng and Church (2000) adapted work of Hartigan (1972) for gene expression data, producing simultaneous clusters of genes and conditions and an *overall partial clustering* of the data. From Eq. 1 each value a_{ij} of a sub-matrix can be defined from the typical value within the bi-cluster a_{IJ}, plus the offsets for the row mean, a_{iJ}-a_{IJ} and column mean a_{Ij}-a_{IJ}. Thus, each value in the sub-matrix should (ideally) be:

$$a_{ij} = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ} \qquad \textbf{Eq. 6}$$

The Cheng and Church technique defines the *Mean Residue Score (H)* of a sub-matrix, based on Eq. 2 and Eq. 6, such that:

$$H(I,J) = \frac{1}{|I||J|}\sum_{i \in I, j \in J}(r_{ij})^2 \qquad \textbf{Eq. 7}$$

The algorithm carries out greedy iterative searches for sub-matrices *(I,J)*, which minimise this function (Eq. 7), generating a large time cost, as each row and column of the dataset must be tested for deletion. (A sub-matrix is considered a bi-cluster if its Mean Residue Score falls below a user specified threshold). A further overhead is the *masking of a bi-cluster with random numbers* once it is found, to prevent finding the same clusters repeatedly on successive iterations. There is, nevertheless, high probability that this replacement with random numbers affects the discovery of further bi-clusters. To overcome this random "interference" Yang et al. (2003) developed **Flexible Overlapped bi-Clustering** (FLOC), generalising the model of Cheng and Church to incorporate null values.

For both the Cheng and Church (2000) algorithm and the FLOC generalisation, *K* can be specified to be much larger than the desired number of groups, without affecting the outcome, as each row and column in the expression matrix can belong to more than one bi-cluster, (Cheng and Church, 2000; Yang et al., 2003). Selecting *K* then reduces to selecting the percentage of the bi-clusters with the best Mean Residue-score (Eq. 7). The cost is increased computation time – the Cheng and Church algorithm finds one bi-cluster at a time while FLOC finds all simultaneously. However, a major additional strength of the bi-clustering techniques is *the minimal requirement for domain knowledge*. Also, as FLOC accounts for null values in the dataset, the preliminary imputation of missing values is not necessary.

Graph theoretic Methods:

A further approach, which is proving useful in the analysis of large complex biological datasets, is that of graph theory, (Aiello et al., 2000; Maslov et al., 2004; Aittokallio and Schwitowski, 2006; Guillaume and Latapy, 2006). A given gene expression dataset can be viewed as a *weighted bipartite graph, G= (V, U, E)*, where *V* is the set of gene vertices, *U* is the set of condition vertices, with $V \cap U = \Phi$, and *E* is the set of edges, with $(u,v) \in E$ having a weight a_{uv} proportional to the strength of the expression of gene *v* under condition *u* (Figure 3).

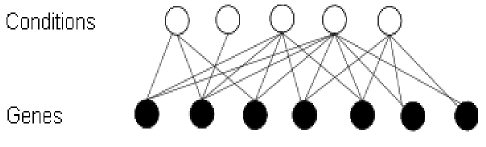


Figure 3 - Bipartite graph representing expression for 7 genes under 5 conditions - edges indicate a change in expression.

Analysis of the models involved focuses on identification of similarly or densely connected sub graphs of nodes and, of course, relies greatly on the method used to define the *edge weights*. Clustering by graph network leads to similar issues as before;

(i) results are highly sensitive to data quality and input parameters, (ii) predicted clusters can vary from one method of graph clustering to another. Clusters that share nodes and edges of a graph networks, clearly "overlap" and as noted in Section *Characteristics of the Gene Expression Dataset*, are desirable for gene expression interpretation.

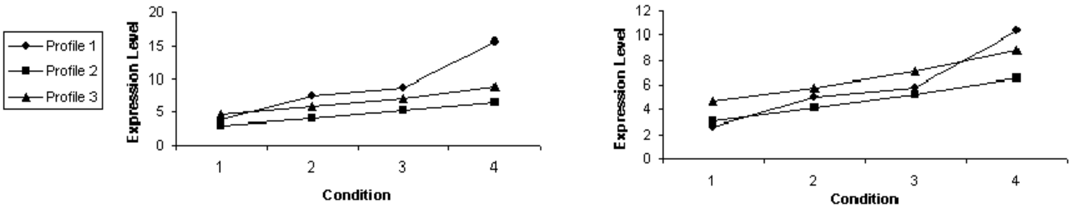
The "Statistical Algorithmic Method for Bi-cluster Analysis" (SAMBA) (Tanay et al., 2000), uses a graphical approach and, unlike previous bi-clustering techniques, finds coherent evolution in the data. An edge is defined to exist between a gene node *u* and a condition node *v* if there is significant change in expression of gene *u* under condition *v,* relative to the genes normal level; (a non-edge exists if *u* does not change expression under *v*). Each edge and non-edge is then weighted, based on a log likelihood model, with weights:

$$\log \frac{p_c}{P_{(u,v)}} > 0 \quad \text{for edges, and} \quad \log \frac{(1-P_c)}{(1-P_{(u,v)})} < 0 \quad \text{for non-edges.} \qquad \textbf{Eq. 8}$$

Here, $P_{(u,v)}$ is the fraction of random bipartite graphs, with degree sequence identical to *G*, that contain edge *(u,v)*, and P_c is a constant probability assigned by the user. For $P_c > \max_{(u,v)\in U\times V} P_{(u,v)}$, edges are taken to occur in a bi-cluster with equal probability. Weights as determined by Eq. 8 are assigned to the edges and non-edges in the graph. A major strength of this method is that statistical significance of any sub graph is then simply determined by its weight.

Cluster Evaluation and Comparison

Evaluation is not particularly well developed for clustering analyses applied to gene expression data, as very little may be known about the dataset beforehand. Many clustering algorithms are designed to be exploratory; producing different clusters according to given classification criteria and will discover a structure, meaningful in that context, which may yet fail to be optimal or even biologically realistic. For example, for K-Means the "best" structure is one that minimises the sum of squared errors (MacQueen, 1967) while, for the Cheng and Church algorithm (Cheng and Church, 2000), it is that which minimises of the Mean Residue-Score (Eq. 7). The two may not be directly comparable, as the former highlights *global patterns* in the data and the latter *local patterns*. While larger deviations from the mean may also correspond to large residue scores this will not always be the case. For example, Figure 4 highlights a simple situation with three genes in a cluster. According to the K-means criterion, the within cluster distance is approximately 11.02, based on Euclidean distance and centroid of the cluster. The Mean Residue Score is 0. Reducing the scale of profile 1 by one third, (Figure 4(b)), decreases the within-cluster distance to 7.91, while increasing the Mean Residue Score slightly to 0.0168. Both (a) and (b) are roughly equivalent. Consequently, interpretation of cluster results relies on some level of subjectivity as well as independent validation and integration of findings. Subjective evaluation, even for low dimensional data, is non-trivial at best, but becomes increasingly difficult for high dimensional gene expression data. Clearly, each technique *will* find patterns even if these are not meaningful in a biological context.



Profile 1	3.9	7.5	8.7	15.6
Profile 2	3.1	4.1	5.2	6.5
Profile 3	4.7	5.8	7.1	8.8

(A)

2.6	5	5.8	10.4
3.1	4.1	5.2	6.5
4.7	5.8	7.1	8.8

(B)

Figure 4 – Gene expression profiles for two equivalent clusters; cluster in (B) has Profile 1 scaled down by one third.

The benefits of the individual techniques, as applied to gene expression data, were highlighted in the last section. This section aims at providing navigational guidelines for some degree of objective evaluation and comparison.

Determining the correct number of clusters:

Cluster number, in the absence of prior knowledge, is determined by whether a non-random structure exists in the data. Limiting to a specific number of groups will bias the search, as patterns tend to be well-defined only for the strongest signals. Commonly, statistical tests for spatial randomness test if a non-random structure exists, but identification of small, biologically meaningful clusters remains non-trivial. This is particularly true of standard methods, which find global structure, but lack fine-tuning to distinguish local structures. Selection of the correct *number* of clusters *(K)* is thus inherently iterative. Near optimal *K* should clearly minimise heterogeneity between groups, while maximising homogeneity within groups, but determining the number of significant clusters relies, not only on direct extraction (or assessment) but also on appropriate hypothesis testing. Direct methods are based on various of criteria[iv]. Nevertheless, improvement in terms of identification of local clusters is slight. Specific tests include Gap Statistic (Tibshirani et al., 2001), Weighted Discrepant Pairs (WADP) (Bittner et al., 2000) and a variety of permutation methods (Bittner et al., 2000; Fridlyand and Dudoit, 2001). Since most involve bootstrapping, these methods can be computationally very expensive. Comparison of methods for selecting the number of groups is discussed by Milligan and Cooper (1985) and, more recently, by Fridlyand and Dudoit (2001), who note that *no existing tests are optimal for gene expression data.*

Comparing Results from clustering algorithms:

Numerical measures of cluster "goodness" include **cluster cohesion** (compactness or tightness), i.e. how closely related genes in a cluster are, while measures of **cluster separation** (isolation), determine how distinct each cluster is. The *Group Homogeneity Function*, is often used to measure the association (distinctiveness) of genes within and between groups.

Comparison with Metadata

Including biological function information in the gene list for each cluster inevitably provides a more complete picture of the dataset and of the success of the technique.

This information can be used to validate the clusters produced, and a number of functional annotation databases are available. The Gene Ontology database (Ashburner et al., 2000) for example, provides a structured vocabulary that describes the role of genes and proteins in all organisms. The database is organised into three ontologies: biological process, molecular function and cellular component. Several tools[v] have been developed for batch retrieval of GO annotations for a list of genes. Statistically relevant GO terms can be used to investigate the properties shared by a set of genes. Such tools facilitate the transition from data collection to biological meaning by providing a template of relevant biological patterns in gene lists.

FUTURE TRENDS

Despite the shortcomings, the application of clustering methods to gene expression data has proven to be of immense value, providing insight on cell regulation, as well as on disease characterisation. Nevertheless, not all clustering methods are equally valuable in the context of high dimensional gene expression. Recognition that well-known, simple clustering techniques, such as K-Means and Hierarchical clustering, do not capture more complex local structures in the data, has led to bi-clustering methods, in particular, gaining considerable recent popularity, (Califano et al., 2000; Cheng and Church, 2000; Getz et al., 2000; Ben-Dor et al., 2002; Busygin et al., 2002; Lazzeroni and Owen, 2002; Tanay et al., 2002; Kluger et al., 2003; Liu and Wang, 2003; Segal et al., 2003; Sheng et al., 2003; Yang et al., 2003;). Indications to date are that these methods provide increased sensitivity at local structure level for discovery of meaningful biological patterns.

Achieving full potential of clustering methods is constrained at present by the lack of robust validation techniques, based on external resources, such as the GO database. *Standardisation of gene annotation methods* across publicly available databases is needed before validation techniques can be successfully integrated with clustering information found from datasets.

The "Central Dogma" that "DNA makes mRNA makes proteins" that comprise the proteome is overly simple. A single gene does not translate into one protein and protein abundance depends not only on transcription rates of genes but also on additional control mechanisms, such as mRNA stability[vi], regulation of the translation of mRNA to proteins[vii] and protein degradation[viii]. Proteins can also be modified by *post-translation activity*[ix] (Brown, 2002(a)). The study of proteomic and transcription data investigates the way in which changes connect gene expression to the physical chemistry of the cell. Integration and merger of proteomic and transcription data sources across platforms is needed, together with development of automated high-throughput comparisons methods if detailed understanding of cell mechanisms is to be achieved. To this end, a standard method of gene and protein annotation across databases is overdue, (Waters, 2006). The development of Bioinformatics/data-mining tools that span different levels of *"omics"* is a necessary next step in the investigation of cell function.

CONCLUSION

Clustering gene expression data is non-trivial and selection of appropriate algorithms is vital if meaningful interpretation of the data is to be achieved. Successful analysis has profound implications for knowledge of gene function, diagnosis, and for targeted drug development amongst others. The evidence to date is that *methods, which* *determine global structure, are insufficiently powerful given the complexity of the data*. Bi-clustering methods offer interpretability of data features and structure to a degree not possible with standard methods. However, even though less sophisticated algorithms such as K-means are achieving some success and while bi-clustering methods seem promising, these are the first steps only to analysing cellular mechanisms and obstacles remain substantial. A significant barrier to the integration of genomic and proteomic platforms and understanding cellular mechanisms is the lack of standardisation. Integration of heterogeneous datasets must be addressed before analysis of gene expression data comes of age.

REFERENCES:

1. Aiello, W, Chun, F. and Lu, L. (2000) A random graph model for massive graphs. In *Proceedings of the 32nd Annual ACM symposium on Theory of computing,* Portland, Oregon, USA, 171 – 180, ACM Press.
2. Aittokallio, T. and Schwikowski, B. (2006) Graph-based methods for analyzing networks in cell biology. *Briefings in Bioinformatics,* 7(3), 243 – 255.
3. Al-Shahrour, F., Díaz-Uriarte, R. and Dopazo, J. (2003) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics,* 20(4), 578 – 580.
4. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C.Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature* 403(6769), 503-511.
5. Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J. M., Davies A. P., Dolinski K., Dwight S. S., Epping J. T., Harris M. A., Hill D. P. Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J. E., Ringwald M., Rubin G. M., Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 25(1), 25-29.
6. Ben-Dor, A., Chor, B., Karp, R. and Yakhini, Z. (2002) Discovering local structure in gene expression data: The order-preserving submatrix problem. In *Proceedings of the 6th International Conference on Computational Biology (RECOMB '02),* Washington DC, USA, 49 – 57, ACM Press.
7. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D. and Sondak, V. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature* 406(6795), 536-540.
8. Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics* 19(2), 185-93.
9. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, *Nature Genetics,* 29(4), 365-371.
10. Brown, T. A. (2002(a)) Transcriptomes and Proteomes. In *Genomes, 2 edition* (pp 70 – 91), Manchester, UK: Wiley-Liss.
11. Brown, T. A. (2002(b)) Synthesis and Processing of the Proteome. In *Genomes, 2 edition* (pp 314 – 344), Manchester, UK: Wiley-Liss.
12. Busygin, S., Jacobsen, G. and Kramer, E. (2002) Double conjugated clustering applied to leukaemia microarray data. In *Proceedings of the 2nd SIAM International Conference on Data Mining, Workshop on Clustering High Dimensional data,* Arlington, Virgina, USA, 420 – 436, Soc for Industrial & Applied Math.

13. Califano, A., Stolovitzky, G. and Tu, Y. (2000) Analysis of gene expression microarrays for phenotype classification. In *Proceedings of the International Conference on Computational Molecular Biology*, Tokyo, Japan, 75 – 85, ACM Press.
14. Cheng, Y. and Church, G. M. (2000) Biclustering of expression data. *Proceedings International Conference on Intelligent Systems for Molecular Biology*; ISMB.International Conference on Intelligent Systems for Molecular Biology 8, 93-103.
15. Cho, R.J., Campbell, M.J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and Davis, R. W. (1998) A genome-wide transcriptional analysis of mitotic cell cycle. *Molecular Cell 2,* 1, 65-73.
16. Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. (2003) DAVID: Database for Annotation, Visualization and Integrated Discovery. *Genome Biology,* 4:R60.
17. Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. and Krawetz, S. A. (2003) Global functional profiling of gene expression. *Genomics,* 81(2), 98 – 104.
18. Eisen, M B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome wide expression patterns, *PNAS,* 95(25), 14863-14868.
19. Fridlyand, J. and Dudoit, S. (2001) *Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method*. (Technical Report 600), Berkeley, California: University of California, Department of Statistics.
20. Friemert, C., Erfle, V. and Strauss, G. (1989) Preparation of radiolabeled cDNA probes with high specific activity for rapid screening of gene expression. *Methods Molecular Cell Biology*, 1, 143-153.
21. Gasch A. P. and M. B. Eisen. M.B. (2002) Exploring the conditional coregulation of yeast in gene expression through fuzzy K-Means clustering. *Genome Biology,* 3(11) RESEARCH0059.1 – RESEARCH0059.22.
22. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286 (5439), 531-537.
23. Getz, G., Levine, E. and Domany, E. (2000) Coupled two-way clustering analysis of gene microarray data. *PNAS,* 97(22), 12079 – 12084.
24. Gress, T.M., Hoheisel, J.D., Sehetner, G. and Leahrach, H. (1992). Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mammalian. Genome,* 3, 609-619.
25. Guillaume, J. L. and Latapy, M. (2006) Bipartite graphs as models of complex networks. *Physica A,* 317, 795 – 813.
26. Hartigan, J.A. (1972) Direct clustering of a data matrix. *Journal of the American Statistical Association,* 67(337), 123-129.
27. Higgins, J. P., Shinghal, R., Gill, H., Reese, J. H., Terris, M., Cohen, R. J., Fero, M., Pollack, J. R. van de Rijn, M. and Brooks**,** J. D. (2003) Gene Expression Patterns in Renal Cell Carcinoma Assessed by Complementary DNA Microarray, *The American Journal of Pathology* 162(3), 925 – 932.
28. Hosack, D. A., Dennis, G. Jr., Sherman, B. T., Lane, H. C. and Lempicki, R. A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biology,* 4:R70.
29. Jain, A.K., Murty M.N., and Flynn P.J. (1999) Data Clustering: A Review, *ACM Computing Surveys,* 31(3), 264-323.
30. Kaufmann, L. and Rousseeuw, P. J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons Inc., Chinchester, New York, Weinheim, 1990.
31. Kerr, M.K. and Churchill, G.A., (2001) Experimental design for gene expression microarrays, *Biostatistics,* 2(2), 183 – 201.
32. Khodursky, A. B., Peter, B. J., Cozzarelli, N. R., Botstein, D., Brown, P. O. and Yanofsky, C. (2000) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli, *PNAS*, 97(22), 12170 - 12175
33. Kishnapuram R. and Keller J.M. (1993) A possibilistic approach to clustering. *Fuzzy Systems, IEEE Transactions on,* 1(2), 98-110.
34. Kluger, Y., Basri, R., Chang, J. T. and Gerstein, M. (2003) Spectral biclustering of microarray data: coclustering genes and conditions, *Genome research,* 13(4), 703-716.
35. Kohonen, T. (1990) The self-organizing map. *Proceeding of the IEEE,* 78(9), 1464-1480.
36. Lazzeroni, L. and Owen, A. (2002) Plaid models for gene expression data. Statistica Sinica, 12, 61 – 86.

37. Liu, J. and Wang, W. (2003) Op-cluster: Clustering by tendancy in high dimensional space. In Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, Florida, USA, 187 – 194, IEEE Computer Society Press
38. Liu, X., Cheng, G., and Wu, J. X. (2002) Analyzing outliers cautiously. *Knowledge and Data Engineering, IEEE Transactions on,* 14(2), 432-437.
39. MacQueen, J.B. (1967) *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1:281-297, University of California Press
40. Maderia S. C. and Oliveira, A. L. (2004) Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE Transactions on Computational Biology and Bioinformatics,* 1(1), 24-45
41. Makretsov, N. A., Huntsman, D. G., Nielsen, T. O., Yorida, E., Peacock, M., Cheang, M. C. U., Dunn, S. E., Hayes, M., van de Rijn, M., Bajdik, C. and Gilks, C. B. (2004) Hierarchical Clustering Analysis of Tissue Microarray Immunostaining Data Identifies Prognostically Significant Groups of Breast Carcinoma, *Clinical Cancer Research,* 18(10), 6143 – 6151.
42. Maslov, S., Sneppen, K. and Zaliznyak, A. (2004) Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A,* 333, 529 – 540.
43. Milligan, G. W. and Cooper M. C. (1985) An examination of procedures for determining the number of clusters in a dataset. *Psychometrika,* 50, 159-179.
44. Quakenbush, J. (2001) Computational Analysis of Microarray Data, *Nature Review Genetics*, 2 (6), 418 – 427
45. Raser, J. M. and O' Shea E. K. (2005) Noise in Gene Expression Data: Origins and Control. *Science,* 309, 2010-2013.
46. Sander, J., Ester, M., Kriegel, K. P. and Xu, X. (1998) Density-Based Clustering in Spatial Databases: The Algorithmic GDBSCAN and its Applications. *Data Mining and Knowledge Discovery,* 2(2), 169 – 194.
47. Schulze, A. and Downward, J. Navigating gene expression using microarrays – a technology review. *Nature Cell Biology*, 3(8), E190 - 195.
48. Segal, E. Taskar, B., Gasch, A., Friedman, N. and Koller, D. (2003) Decomposing gene expression into cellular processes. In *Proceedings of the Pacific Symposium on Biocomputing,* Lihue, Hawaii, USA, 89 – 100, World Scientific Press.
49. Scott A. J. and Symons, M. J. (1971) Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2), 387-397.
50. Sheng, Q., Moreau, Y. and De Moor, B. (2003) Biclustering microarray data by Gibbs sampling. *Bioinformaics,* 19(Supp. 2), ii196 – ii205.
51. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lan-der, and Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proceedings of the National Academy of Sciences of the United States of America* 96(6), 2907-2912.
52. Tanay, A. Sharan, R., and Shamir, R. Discovering statistically significant biclusters in gene expression data, *Bioinformatics*, 18(1), S136-44.
53. Tavazoie, S., Hughes, J. D., Campbell, M. J. Cho, R. J. and Church, G. M. (1999) Systematic determination of genetic network architecture, *Nature genetics,* 22(3), 281-285.
54. Tibshirani, R., Walther, G. and Hastie, T. (2001) Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 63(2), 411-423.
55. Toronen, P., Kolehmainen, M., Wong, G. and Castren, E. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Letters,* 451(2), 142 – 146.
56. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520 – 525.
57. van der Laan, M. J. and Pollard, K. S. (2001) *Hybrid clustering of gene expression data with visualization and the bootstrap.* (Technical Report 93), U.C. Berkeley Division of Biostatistics Working Paper Series, Berkeley, California, University of California, School of Public Health, Division of Biostatisitcs.
58. Waters, K. M., Pounds, J. G. and Thrall B. D. (2006) Data merging for integrated microarray and proteomics analysis *Briefings in Functional Genomics and Proteomics,* 5(4), 261 – 272.
59. Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. and Somogyi**,** R. (1998) Large Scale temporal gene expression mapping of central nervous system development. *PNAS,* 95(1), 334-339

60. Yang, J., Wang, H., Wang, W. and Yu, P. (2003) Enhanced biclustering on expression data, *Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering (* BIBE '03), IEEE Computer Society, 321 - 327.

61. Zeeberg, B.R., Feng, W., Wang, Geoffrey, W., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C. and Weinstein, J. N. (2003) GoMiner: a resource for biological interpretation of genomic and proteomics data. *Genome Biology,* 4:R28.

62. Zha, H., Ding, C., Gu, M., He, X. and Simon, H.D. (2002) Spectral relaxation for k-means clustering. *Proceedings Neural Information. Processing Systems.* 14, 1057 – 1064.

[i] Microarray development timeline: 1989 – development of world's first microarray; 1991 – Photolithographic printing technique developed by Affymetrix; 1993 – Microarray containing over 1 million DNA sequences developed; 1994 – First cDNA collections developed by Stanford; 1995 – Quantitative monitoring of gene expression patterns with cDNA microarray; 1996 – Commercialisation of arrays (Affymetrix); 1997 – Genome-wide expression monitoring in Yeast; 2000 – Portraits/Signatures of gene expression in cancer identified; 2002 - Genechip® Human Genome two array set developed for analysis of over 33,000 genes from public databases; 2003 – Microarray technology introduced to clinical practices; 2004 – Whole human genome on one microarray.

[ii] The two most popular array platforms are complementary DNA (cDNA) and oligonucleotide microarrays. The former contains cDNA probes that are products synthesized from polymerase chain reactions generated from cDNA and clone libraries, the latter contain shorter synthesized oligonucleotide probes (prefect match and mismatch) generated directly from sequence data. A key difference between the two platforms is the manner in which the data is presented for analysis. Intensity measurements for cDNA arrays are the result of competitive hybridisation, (where two transcription samples of interest (labelled with two different dyes) are hybridised to the same array), resulting in a measurement of the ratio of transcript levels for each gene, (usually reported as a log ratio). Oligonucleotide arrays, on the other hand, results from non-competitive hybridisation (where one transcription sample is hybridised to a array and difference in expression levels between two samples are compared across arrays). Here, measurement level for a gene is presented as the average measurement of all probes representing the gene (depending on pre-processing technique this may have mismatch probes subtracted first). See Schulze and Downward (2001) for a review.

[iii] Gene expression patterns with similar frequency and phase.

[iv] These include likelihood ratios (Scott and Symons, 1971), cluster sums of squares (Milligan and Cooper, 1985), average silhouette (Kaufmann and Rousseeuw, 1990) or mean split silhouette (van der Laan and Pollard, 2001).

[v] Tools important for the management and understanding of large scale gene expression data: FatiGo (Al-Shahrour et al., 2003), GoMiner (Zeeberg et al., 2003), OntoExpress (Draghici et al., 2003), EASE (Hosack et al., 2003), DAVID Gene classification tool (Dennis et al., 2003).

[vi] Sequences of mRNA may vary considerably in stability. The balance between mRNA degradation and mRNA synthesis determines the level of mRNA in the cell.

[vii] The mechanisms, including regulatory proteins, which dictates which genes are expressed and at what level.

[viii] The method and rate at which protein is broken down in the body.

[ix] Before taking on a functional role in the cell an amino acid sequence must fold into its correct tertiary structure. Additional post-processing events may occur, such as proteolytic cleavage, chemical modifications, intein splicing. (Brown, 2002(b)).