THE USE AND ANALYSIS OF MICROARRAY DATA

Atul Butte

Functional genomics is the study of gene function through the parallel expression measurements of genomes, most commonly using the technologies of microarrays and serial analysis of gene expression. Microarray usage in drug discovery is expanding, and its applications include basic research and target discovery, biomarker determination, pharmacology, toxicogenomics, target selectivity, development of prognostic tests and disease-subclass determination. This article reviews the different ways to analyse large sets of microarray data, including the questions that can be asked and the challenges in interpreting the measurements.

At the time of writing this article, the genomes of more than 800 organisms have been sequenced, and well over 3.5 million genetic sequences have been deposited in international repositories. However, the biological functions of most of these genes remain unknown, or have been predicted only through homology to genes with functions that are better known. One way to determine the functions of these genes is through repeated measurements of their RNA transcripts; for example, knowing that a particular gene is expressed only in cardiac muscle and only under particular conditions implicitly gives us functional knowledge about that gene. Functional genomics is the study of gene function through parallel expression measurements of a genome. The most common tools used to carry out these measurements include complementary DNA microarrays[1], oligonucleotide microarrays[2] or serial analysis of gene expression (SAGE)[3]. This article will focus on microarrays, which are artificially constructed grids of DNA, such that each element of the grid probes for a specific RNA sequence — that is, each holds a DNA sequence that is a reverse complement to the target RNA sequence.

Although there are many protocols and types of system available, the basic technique involves extraction of RNA from biological samples in either normal or interventional states. The RNA (or, in some protocols, isolated messenger RNA) is then copied, while incorporating either fluorescent nucleotides or a tag that is later stained with fluorescence. The labelled RNA is then hybridized to a microarray for a period of time, after which the excess is washed off and the microarray is scanned under laser light. This process is schematized in FIG. 1. With oligonucleotide microarrays, for which all probes have been designed to be theoretically similar with regard to hybridization temperature and binding affinity, each microarray measures a single sample and provides an absolute measurement level for each RNA molecule, although this absolute measurement might not correlate exactly with concentration in terms of micrograms per unit volume. With cDNA microarrays, for which each probe has its own hybridization characteristic, each microarray measures two samples, and provides a relative measurement level for each RNA molecule. Regardless of the technique, the end result is 4,000–50,000 measurements of gene expression per biological sample. As a complete experiment might involve anywhere up to hundreds of microarrays, the resultant RNA-expression data sets can vary greatly in size.

As the cost of microarrays continues to drop, it is clear that microarrays are becoming more integral to the drug discovery process. In addition to the obvious use of functional genomics in basic research and target discovery, such as finding genes expressed in significantly different patterns across samples, there are many other specific uses in this domain. These include: biomarker determination, to find genes that correlate with and presage disease progression, but are easier to measure and follow in clinical trials; pharmacology, to

Children's Hospital Informatics Program and Division of Endocrinology, Children's Hospital, 300 Longwood Avenue, Boston, Massachusetts 02115, USA.
e-mail: atul_butte@harvard.edu

doi:10.1038/nrd961

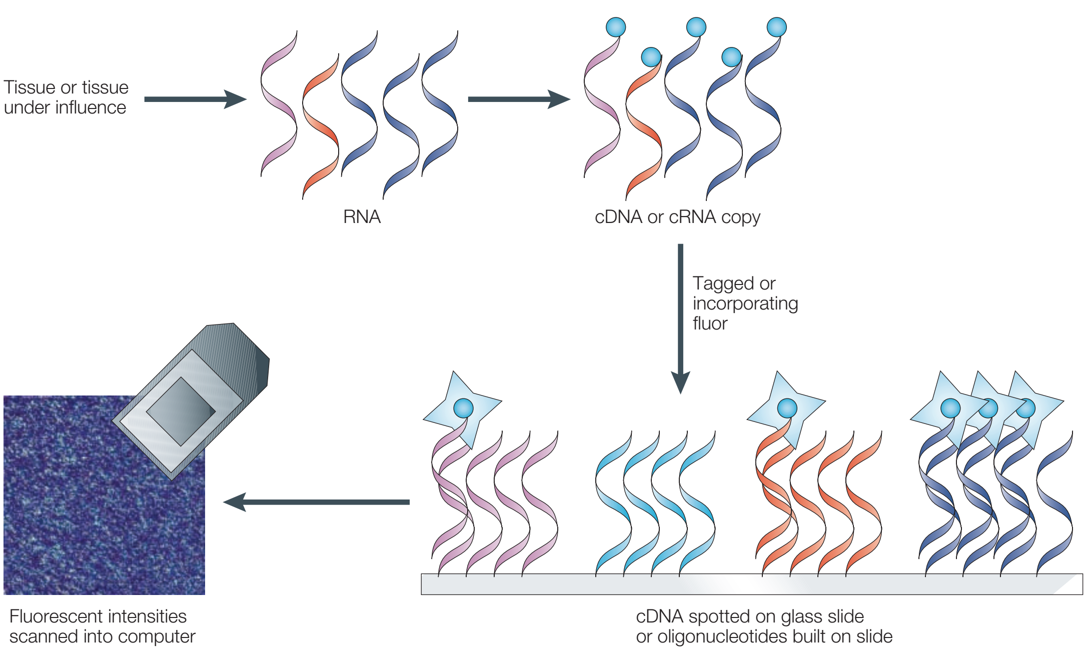
Tissue or tissue under influence
RNA
cDNA or cRNA copy
Tagged or incorporating fluor
Fluorescent intensities scanned into computer
cDNA spotted on glass slide or oligonucleotides built on slide

Figure 1 | **Schematized experimental process using a microarray.** Although the specific protocols differ, the microarray approach first involves isolating RNA or messenger RNA from appropriate biological samples, making the RNA (or a copy of it) fluorescent, hybridizing it to the microarray, washing off the excess and scanning the microarray under laser light. cDNA, complementary DNA; cRNA, complementary RNA.

determine differences in gene expression in tissues exposed to various doses of compounds; toxicogenomics, to find gene-expression patterns in a model tissue or organism exposed to a compound and their use as early predictors of adverse events in humans; target selectivity, to define a compound by the gene-expression pattern it provokes in a target tissue and then compare it with other compounds using these patterns; prognostic tests, to find a set of genes that accurately distinguishes one disease from another; and disease-subclass determination, to find multiple subcategories of tumours in a single clinical diagnosis.

Many free (BOX 1) and commercial software packages are now available to analyse microarray data sets, although it is still difficult to find a single off-the-shelf software package that answers all functional-genomics questions. As the field is still young, when developing a bioinformatics analysis pipeline, it is more important to have a good understanding of both the biology involved and the analytical techniques rather than having the right software. This article reviews the different ways to analyse microarray data, and will concentrate on choosing the appropriate method for the given hypothesis.

Normalization and noise

Before multiple microarray measurements can be integrated into a single analysis, the reported measurements need to be normalized, or modified (possibly corrected) to make them comparable. When microarrays are used to collect gene-expression data in an experiment in which the measurements are made at the same time, with homogeneous populations of similar cells and using a single microarray technology, normalization might simply be a matter of adjusting the overall brightness of each scanned microarray image, assuming that the quantity of RNA is equal[4]. Other normalization methods include: using expression levels of 'housekeeping' genes[5]; using assumptions that most genes do not change across experiments[6]; using SPLINES[7]; or other nonlinear techniques[8,9].

Typically, however, functional-genomics experiments are more complicated. Recently, increasing efforts have been invested in characterizing the 'noise' in microarray technology. Studies addressing the reproducibility of microarray data analysed replicated data[10], compared microarray measurements with NORTHERN BLOTS[11,12] and SAGE[13], and evaluated strategies for REVERSE TRANSCRIPTION[14] and *in vitro* transcription amplification[15]. As a result, it has become increasingly clear that there are several substantial sources of noise in microarray data. Intra- and inter-microarray variations can markedly skew the interpretation of such expression data.

First, improving the reliability of expression measurements starts with proper experimental design. For example, microarrays can measure across the genome, including genes with expression that is controlled by hormones, such as growth hormone or cortisol. So, if organ samples are acquired at various times during the day, genes that appear to be differentially expressed might only be reflecting normal circadian physiology. Pooling samples before hybridization might control for this biological 'noise.'

In addition, scanned hybridization images need to be inspected for artefacts, such as scratches and bubbles[16,17]. Measuring replicate microarrays for each biological sample allows the modelling of this technical noise.

SPLINES
Instead of fitting a complex polynomial curve to data, splines allow the fitting of data by putting together smaller, less complex curves.

NORTHERN BLOT
Different RNA molecules are separated by mass on a gel, then radioactively labelled complementary DNA or RNA molecules are used to quantify specific RNA amounts.

REVERSE TRANSCRIPTION
The synthesis of a strand of DNA from RNA, which is used to make a complementary DNA copy of sample RNA.

Most reported expression data have been obtained on relatively homogeneous cell populations. However, when RNA is extracted from whole organs or from tumour biopsies, the sources of variation increase. There is substantial heterogeneity of expression in cell subpopulations in most organs and in many tumours. Failure to account for such variation could lead to overinterpretation or spurious functional gene association. Microdissection of cell subpopulations (for example, with laser capture[18]) is possible only in a minority of the systems of interest. If microarray-based gene-expression measurements are to be reliable and economical, both at the level of basic biology and clinical assays, then all of these further sources of noise/variation must be incorporated directly into the analytical tools that interpret these data.

A further issue that needs to be addressed is the difference between the two most commonly used microarray technologies: spotted cDNA microarrays, which report differences in gene expression between two samples, and oligonucleotide microarrays, which report absolute expression levels. Normalization techniques for one microarray technology might not apply to another, owing to differences in assumptions and the distributions of the output measurements. For example, if we assume that in any given experiment, most genes in a cell do not change in expression and an equal number of genes are up- and downregulated (not always a valid assumption), then differential-expression measurements from spotted arrays might be found to be normally distributed, whereas measurements from oligonucleotide microarrays will not have the same distribution.

Expression measurements made across microarray technologies are not directly comparable. For example, the published microarray measurements from the National Cancer Institute (NCI) 60-cancer-cell-line panel, from spotted arrays[19] and from oligonucleotide arrays[20], show poor correlation between measurements[21]. This can be explained by the differences in low-level hybridization and analysis between the two techniques. To be precise, Affymetrix microarrays contain between 11 and 20 pairs of oligonucleotide probes for a target RNA, for which one of the pair is the reverse complement to an ideally unique 25-mer in the RNA and the other contains a mutated middle base pair and serves as a measure of stray signal. Using the differences between these intensities, the Affymetrix quantitative software judges the reliability of each probe pair and calculates a qualitative and quantitative measurement (see Affymetrix web site). Other quantitative methods are available in addition to the Affymetrix software[7].

Compared with this, cDNA microarrays contain a single probe for each target RNA, and the two biological samples are different colours, so that after hybridization, the two colours are scanned separately and relative expression is determined by comparing intensities. There is striking non-correlation between Affymetrix quantitative measurements and ratios of intensities from a cDNA microarray, because these two technologies clearly measure expression differently.

Supervised or unsupervised

Current methodologies to analyse RNA-expression data sets can be divided into two categories: supervised approaches, or analysis to determine genes that fit a predetermined pattern; and unsupervised approaches, or analysis to characterize the components of a data set without the a priori input or knowledge of a training signal.

Supervised methods. Supervised methods are generally used for two purposes: finding genes with expression levels that are significantly different between groups

Box 1 | Some freely available software for microarray analysis

Although many bioinformatics companies sell software that assists in microarray analysis, there are several freely available software packages that can be used to perform the six analytical techniques described in this article. Only a few are listed here.

Cluster and TreeView. . . . . http://rana.lbl.gov/EisenSoftware.htm
- Although it is the standard for hierarchical clustering and viewing dendrograms, this software also creates self-organizing maps and performs principal-components analysis.

GeneCluster 2.0 . . . . http://www-genome.wi.mit.edu/cancer/software/genecluster2/gc2.html
- Initially used for constructing self-organizing maps, the latest version now also finds nearest neighbours and performs other supervised methods. Written in Java, this program can essentially run under any computer operating system

MultiExpression Viewer. . . . . http://www.tigr.org/software/
- Software that creates self-organizing maps and performs hierarchical clustering, as well as finding principal components. This package also includes a component for support vector machines, but at present offers little for documentation. The software is written in Java, and a license for the source code of the software is also available.

MAExplorer. . . . . http://maexplorer.sourceforge.net/
- This tool performs many aspects of microarray processing, including the raw image analysis. It contains a few analytical techniques, including hierarchical clustering. The software is written in Java, and the source code is freely available for modification.

RELNET . . . . http://www.chip.org/relnet
- Software to create relevance networks. The software is written in Java, and a license for the source code is also available.

of samples, and finding genes that accurately predict a characteristic of the sample. Most functional-genomics experiments still typically use only a handful of microarrays (or equivalent technology), with samples measured under two or three conditions, and the application has the goal of finding those genes that are significantly differentially expressed. Significance has been evaluated in many different ways, including parametric[22] and nonparametric tests[23,24], analysis of variance[25] and many others. Although it would be an understatement to call the analyses of these smaller sets of microarray data trivial, there are several published techniques that have been used to find genes that have the most different expression levels between a few samples. When determining whether a particular gene is differentially expressed between two samples, there are four characteristics that need to be considered: absolute expression level, or whether the gene is expressed at a high or low level; subtractive degree of change between groups, or the difference in expression levels across samples (calculated using subtraction); fold change between groups, or the ratio of expression levels across samples (calculated by division); and reproducibility of the measurement, or whether samples with similar characteristics have similar amounts of the gene transcript.

These four characteristics are related; for example, genes measured at low amounts of expression often have measurements that are less reliable, which leads to poor reproducibility across samples and high-fold changes that do not adequately describe the actual degree of change. All of the available techniques for comparing two sets of microarray measurements essentially evaluate these four characteristics for each gene in various ways to rank genes that are most significantly different.

For larger data sets, comparing microarrays one pair at a time misses trends that might exist between measurements. There are several published supervised methods that find genes or sets of genes that accurately predict sample characteristics, such as distinguishing one type of cancer from another, or a metastatic tumour from a non-metastatic one. These methods might find individual genes, such as the nearest-neighbour approach[26], and/or multiple genes, such as decision trees[27], neural networks[28] and support vector machines[29–31]. This article will focus on the two more popular supervised techniques: nearest-neighbour analysis and support vector machines.

Unsupervised methods. Users of unsupervised methods try to find internal structure or relationships in a data set instead of trying to determine how best to predict a 'correct answer'. Within unsupervised learning, there are three classes of technique: feature determination, or determining genes with interesting properties without specifically looking for a particular a priori pattern, such as principal-components analysis[32–36]; cluster determination, or determining groups of genes or samples with similar patterns of gene expression, such as nearest-neighbour clustering[26,37], self-organizing maps[38,39], *k*-means clustering and one- and two-dimensional hierarchical clustering[19,40]; and network determination, or determining graphs representing gene–gene or gene–phenotype interactions using Boolean networks[41–43], BAYESIAN NETWORKS[44] and relevance networks[20,45,46]. This article will focus on the four most common unsupervised techniques of principal-components analysis, hierarchical clustering, self-organizing maps and relevance networks.

Like, but how different?

It is crucial to make a distinction between dissimilarity measures (also known as 'metrics') and clustering methods. A dissimilarity measure indicates the degree of similarity between two genes. A clustering method builds on these dissimilarity measures to create groups of features with similar patterns.

A commonly used dissimilarity measure is Euclidean distance, for which each gene is treated as a point in multidimensional space, each axis is a separate biological sample and the coordinate on each axis is the amount of gene expression in that sample[35,38]. A schematic of this method is shown in FIG. 2a. One disadvantage of Euclidean distance is that if measurements are not normalized, correlation of measurements can be missed, the focus being instead on the overall amount of expression. A second disadvantage is that genes that are negatively associated with each other will be missed. Negative associations include gene interactions, such as those of tumour-suppressor genes. As an example, the tumour-repressor protein p53 acts as a transrepressor of several other genes. This means that the higher the level of p53, the less expression of other genes is expected. The concept of negative interaction is clearly different than the concept of no interaction.

Another dissimilarity measure that is commonly used is the PEARSON CORRELATION COEFFICIENT, which is measured between two genes that are treated as vectors of measurements[40] (FIG. 2b). The disadvantages in using this measure with gene-expression measurements are: first, it assumes that the measurements are normally distributed, which might not be the case for oligonucleotide-microarray measurements; and second, it assumes that genes interact in the assumed linear model, when in biology, a particular gene might best regulate other genes when in the middle of its own range of expression. Operationally, this measure is sensitive to outliers, and although techniques such as the RANK CORRELATION COEFFICIENT deal with these by replacing the measurements with ranks, it is not clear whether eliminating the outliers is ideal — many past discoveries have been found by focusing on the outliers in biology.

A third dissimilarity measure is mutual information (FIG. 2c), which allows for any possible model of interaction between genes and uses each expression-level measurement equally regardless of the actual value, and is therefore not biased by outliers[46]. However, calculating the mutual information requires using discrete expression measurements (for example, representing the gene as 'high' and 'low', or 'high', 'medium' and 'low', and so on), and the mutual information depends on the number of 'bins' used. Ideally, this would be performed in a gene-specific manner, but sufficient information about

BAYESIAN NETWORK
A graphical representation in which variables (that is, genes) are represented as nodes. Arrows between nodes represent conditional dependence, which is interpretable as causal associations.

PEARSON CORRELATION COEFFICIENT
A measurement of the degree of fit of a linear-regression line to data points, calculated as the average distance of points from the regression line normalized to the standard deviations of the individual coordinates.

RANK CORRELATION COEFFICIENT
Points are restated in terms of their ordinal rank (for example, first, second, third) before calculation of the correlation coefficient.

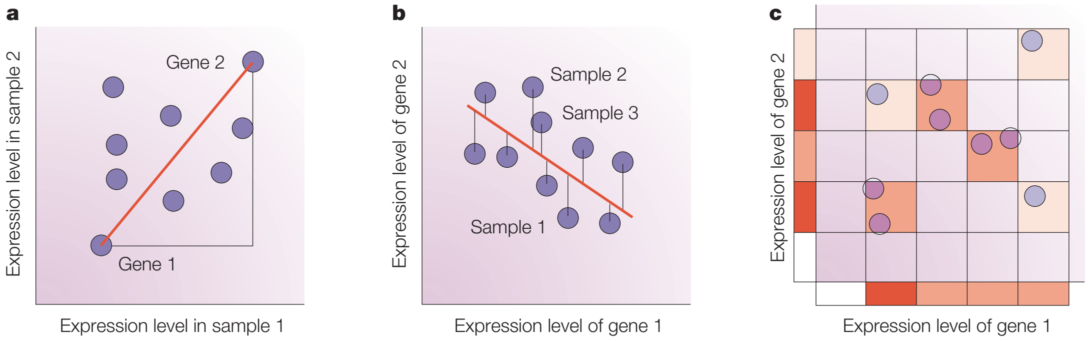
a
Expression level in sample 2
Gene 2
Gene 1
Expression level in sample 1
b
Expression level of gene 2
Sample 2
Sample 3
Sample 1
Expression level of gene 1
c
Expression level of gene 2
Expression level of gene 1

Figure 2 | **Dissimilarity measures.** Almost all clustering techniques require some method of comparing one gene with another and determining similarity. **a** | Euclidean distance can be used as a measurement of the degree of similarity between two genes, and is calculated using the Pythagorean theorem. **b** | The Pearson correlation coefficient is calculated from the distances of each point from the linear-regression line (known as residuals). **c** | Mutual information is a model-free measurement of the degree of information content in one gene known from another gene, and is highest when genes are randomly distributed separately, but show a non-random joint distribution.

the range of expression of each gene in all tissues has not yet been collected. Furthermore, gene–gene associations with high mutual information might not even be functions in the mathematical sense, and might be difficult to explain biologically.

Analytical methods

Once a dissimilarity measure has been chosen, the appropriate analytical technique can be applied. This section describes the four commonly used unsupervised techniques — hierarchical clustering, self-organizing maps, relevance networks and principal-components analysis — and two commonly used supervised techniques — nearest neighbours and support vector machines.

Hierarchical clustering. Hierarchical clustering is a commonly used unsupervised technique that builds clusters of genes with similar patterns of expression. This is done by iteratively grouping together genes that are highly correlated in terms of their expression measurements, then continuing the process on the groups themselves. DENDROGRAMS (FIG. 3a) are used to visualize the resultant hierarchical clustering. A dendrogram represents all genes as leaves of a large, branching tree. Each branch of the tree links two genes, two branches or one of each. Although construction of the tree is initiated by connecting genes that are most similar to each other, genes added later are connected to the branches that they most resemble. Although each branch links two elements, the overall shape of the tree can sometimes be asymmetric. In visually interpreting dendrograms, it is important to pay attention to the length of the branches. Branches connecting genes or other branches that are similar are drawn with shorter branch lengths. Longer branches represent increasing dissimilarity.

Hierarchical clustering is particularly advantageous in visualizing overall similarities in expression patterns observed in an experiment, and because of this, the technique has been used in many publications[40,47]. The number and size of expression patterns within a data set can be estimated quickly, although the division of the tree into actual clusters is often performed visually.

DENDROGRAM
A visual representation of hierarchical clusters.

It is important to note the few disadvantages in their use. First, hierarchical clustering ignores negative associations, even when the underlying dissimilarity measure supports them. Negative correlations might be crucial in a particular experiment, as described above, and might be missed. Furthermore, hierarchical clustering does not result in clusters that are globally optimal, in that early incorrect choices in linking genes with a branch are not later reversible as the rest of the tree is constructed. So, this method falls into a category known as 'greedy algorithms,' which provide good answers, but for which finding the most globally optimal set of clusters is computationally intractable. Despite these disadvantages, hierarchical clustering is a popular technique in surveying microarray expression patterns in an experiment.

Self-organizing maps. Self-organizing maps are similar to hierarchical clustering, in that they also provide a survey of expression patterns within a data set, but the approach is quite different[38,39]. As shown in FIG. 3b, genes are first represented as points in multidimensional space. In other words, each biological sample is considered a separate dimension or axis of this space, and after the axes are defined, genes are plotted using expression levels as coordinates. This is easiest to visualize with three or less microarrays, but extends to a larger number of experiments/dimensions. Nearness can be defined using any of the dissimilarity measures described above, although Euclidean distance is most commonly used.

The process starts with the answer, in that the number of clusters is actually set as an input parameter. A map is set with the centres of each cluster-to-be (known as centroids) arranged in an initial arbitrary configuration, such as a grid. As the method iterates, the centroids move towards randomly chosen genes at a decreasing rate. The method continues until there is no further significant movement of these centroids.

The advantages of self-organizing maps include easy two-dimensional visualization of expression patterns and reduced computational requirements compared with methods that require comprehensive pairwise

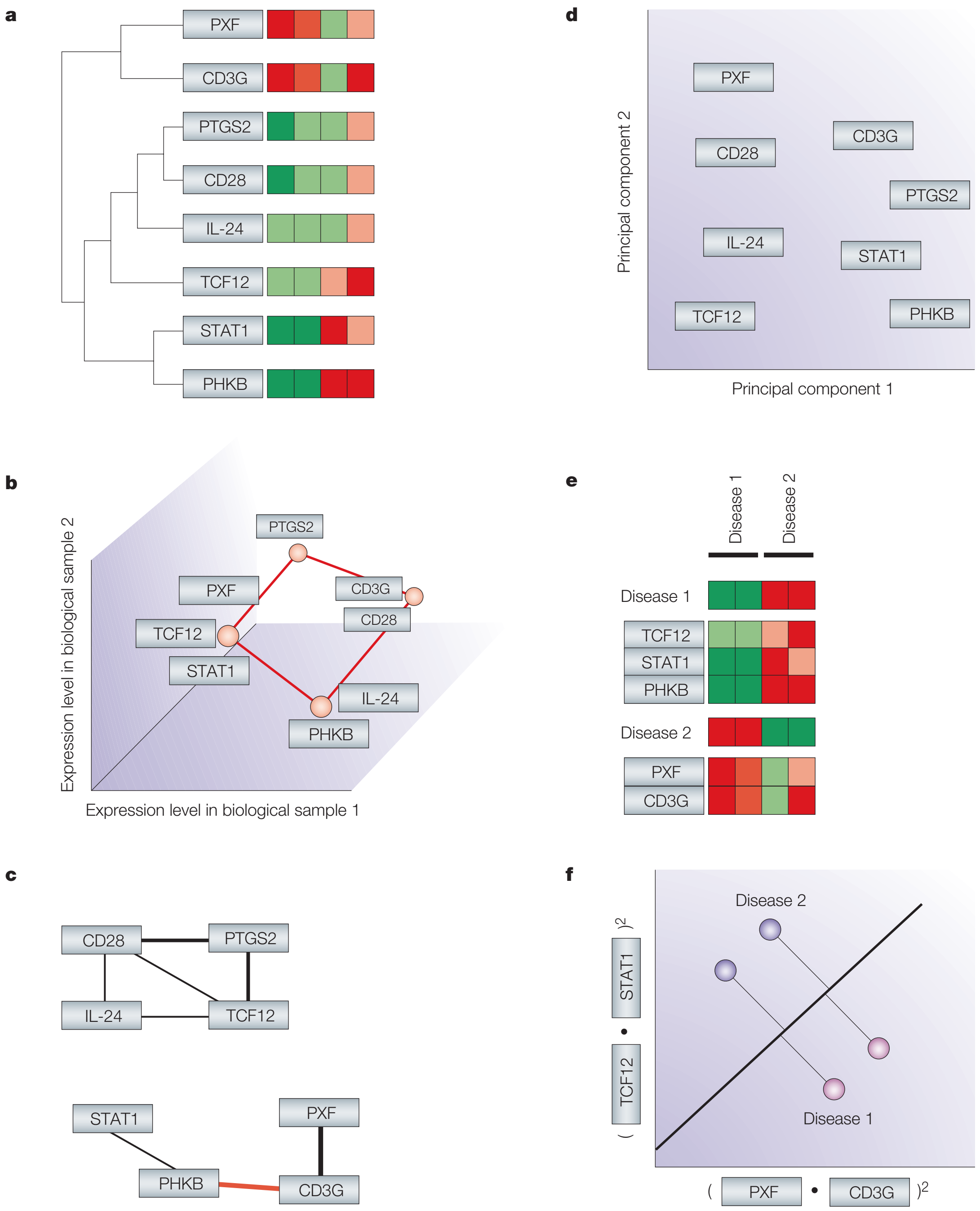

Figure 3 | **Clustering and network-determination methods used in microarray analysis.** The choice of the proper method and the results obtained clearly depend on the starting hypothesis. This figure shows the results of six analytical methods applied to the same hypothetical data set **a** | Hierarchical clustering sorts all genes (or samples), such that similar genes appear near each other. The length of the branch is inversely proportional to the degree of similarity. Shades of red indicate increased relative expression; shades of green indicate decreased relative expression. **b** | Self-organizing maps find variable-sized clusters of genes that are similar to each other, given the input number of clusters to find. **c** | Relevance networks find and display pairs of genes with strong positive and negative correlations, then construct networks from these gene pairs; typically, the strength of correlation is proportional to the thickness of the lines between genes, and red indicates a negative correlation. **d** | Principal-components analysis is typically used as a visualization technique, showing the clustering or scatter of genes (or samples) when viewed along two or three principal components. In the figure, a principal component can be thought of as a 'meta-biological sample', which combines all the biological samples so as to capture the most variation in gene expression. **e** | The nearest-neighbour supervised method first involves the construction of hypothetical genes that best fit the desired patterns (for example, a gene with high expression in disease 1 and low expression in disease 2, or vice versa). The technique then finds individual genes that are most similar to the hypothetical genes. **f** | Instead of restricting to individual genes, support vector machines efficiently try several mathematical combinations of genes to find the line (or plane) that best separates groups of biological samples. CD3G, CD3G antigen, γ-polypeptide; CD28, CD28 antigen; IL-24, interleukin-24; PHKB, phosphorylase kinase-β; PTGS2, prostaglandin-endoperoxidase synthase 2; PXF, peroxisomal farnesylated protein; TCF12, transcription factor 12; STAT1, signal transducer and activator of transcription 1.

comparisons, such as dendrograms. However, there are several disadvantages. First, the initial topology of a self-organizing map is arbitrary and the movement of the centroids is random, so the final configuration of centroids might not be reproducible. Second, similar to dendrograms, negative associations are not easily found. Third, even after the centroids reach the centres of each cluster, further techniques are needed to delineate the boundaries of each cluster. Finally, genes can belong to only a single cluster at a time.

Relevance networks. Continuing through the set of unsupervised techniques, relevance networks allow networks of features to be built, whether they represent genes, phenotypic or clinical measurements[20]. The technique works by first comparing all features with each other in a pairwise manner, similar to the initial steps of hierarchical clustering. Typically, two genes are compared with each other by plotting all the samples on a scatterplot, using expression levels of the two genes as coordinates. A correlation coefficient is then calculated, although any dissimilarity measure can be used. A threshold value is then chosen, and only those pairs of features with a measure greater than the threshold are kept. These are displayed in a graph similar to FIG. 3c, in which genes and phenotypic measurements are nodes, and associations are edges between nodes. Although the threshold is chosen using permutation analysis, it can actually be used as a dial, increasing and decreasing the number of connections shown.

There are several advantages in using relevance networks. First, they allow features of more than one data type to be represented together; for example, if strong enough, a link between systolic blood pressure and expression of a particular gene could be visualized. Second, features can have a variable number of associations; theoretically, a transcription factor might be associated with more genes than a downstream component. Finally, negative associations can be visualized as well as positive ones.

One disadvantage to this method is the degree of complexity seen at lower thresholds, at which many links are found associating many genes in a single network. Completely connected subcomponents of these complex graphs (known as 'cliques') are not easy to find computationally.

Principal-components analysis. Principal-components analysis is more useful as a visualization technique than as an analytical method[33,36]. It can be applied to either genes or samples, which are represented as points in multidimensional space, similar to self-organizing maps. Principal components are a set of vectors in this space that decreasingly capture the variation seen in the points. The first principal component captures more variation than the second, and so on. The first two or three principal components are used to visualize the gene on screen or on a page, as shown in FIG. 3d.

Because each principal component exists in the same multidimensional space, they are linear combinations of the genes. For example, the greatest variation of biological samples might be described as 3 times the expression level of the first gene, plus –2.1 times the expression level of the second gene, and so on. The principal components are linear combinations that include every gene or sample, and the biological significance of these combinations is not directly intuitive.

There are other caveats in using principal components. First, it is important to note whether genes have been centred before analysis; if not, then the first principal component might serve to centre the genes. Second, it is crucial to note that although principal components might best describe the variation seen in an expression data set, they do not describe how to best separate groups of genes or samples. For example, if microarrays are measured on samples from two conditions, principal components will best describe the variation of those samples, but will not always be the best way to split samples from those two conditions[48].

Nearest neighbours. Although the nearest-neighbour technique can be used in an unsupervised manner, it is commonly used in a supervised fashion to find genes directly with patterns that best match a designated query pattern. For example, an ideal gene pattern might be one that is highly expressed in one condition and expressed at a low level in another condition. All the genes that have been measured can then be compared to this ideal gene pattern and ranked by their similarity, as shown in FIG. 3e. For example, acute lymphocytic leukaemia was distinguished from acute myelogenous leukaemia using this method[26].

Although this technique results in genes that might individually split two sets of microarrays, it does not necessarily find the smallest set of genes that most accurately splits the two sets. In other words, a combination of the expression levels of two genes might split two conditions perfectly, but these two genes might not necessarily be the top two genes that are most similar to the idealized pattern.

This technique can be modified for specific cases, such as toxicogenomics. Tissue exposed to various compounds known to cause different types of toxicity can be subjected to microarray measurements, as well as measurements from normal tissues. This could make up a training set, in that these data points implicitly make up a model of toxicity. Newer compounds can then be tested on these tissues (the test set), and the 'distance' of these expression patterns from the training set can be calculated, and a decision made as to similarity of mechanism of toxicity.

Support vector machines. Support vector machines address the problem of finding combinations of genes that better split sets of biological samples[30]. Although it is easy to find individual genes that split two sets with reasonable accuracy owing to the large number of genes (also known as features) measured on microarrays, occasionally it is impossible to split sets perfectly using individual genes. The support vector machines technique actually further expands the number of features available by combining genes using mathematical

Box 2 | Downloadable large data sets of microarray measurements

Although Perou and others have called for the release of raw microarray data after the publication of manuscripts, this still rarely occurs[52]. The following web sites are the exception, and contain large amounts of microarray data that are of good quality and are freely available for academic use:

Stanford Microarray Database. http://genome-www5.stanford.edu/MicroArray/SMD/
- **3,290 Microarrays measured across 11 species, from 80 publications.**

National Center for Biotechnology Information Gene Expression Omnibus. . . . . http://www.ncbi.nlm.nih.gov/geo/
- **2,354 Microarrays from 105 types of microarray, measured across 78 experiments.**

TREX Program in Genomic Applications . . . . http://pga.tigr.org/data.shtml
- **565 Microarrays from mouse and rat models of sleep, infection, hypertension and pulmonary disease.**

Children's National Medical Center (HopGenes Program in Genomic Applications) . . . . http://microarray.cnmcresearch.org/pgadatatable.asp
- **More than 500 microarrays from many human diseases, including muscular dystrophy, dermatomyositis and heart failure, as well as mouse, rat and dog models of spinal-cord injury, pulmonary disease and heart failure.**

CardioGenomics Program in Genomic Applications . . . . http://cardiogenomics.med.harvard.edu/public-data.html
- **142 Microarrays involving mouse models of cardiac development and signal transduction, including measurements made in time-series.**

Human Gene Expression Index . . . . http://www.hugeindex.org/databases/index.html
- **121 Microarrays from 19 normal human tissues.**

Whitehead Institute Center for Genome Research . . . . http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi
- **Microarrays from 12 publications involving many types of cancer, including some clinical measurements associated with each sample.**

operations (called kernel functions). For example, in addition to using the expression levels of two individual genes A and B to separate two sets of biological samples, the combination features $A \times B$, A/B, $(A \times B)^2$ and others, can also be generated and used. To make this clear, it is possible that even if genes A and B individually could not be used to separate the two sets of biological samples, together with the proper kernel function, they might successfully separate the two.

This can be visualized graphically as well, as shown in FIG. 3f. Similar to the principal-components analysis method above, consider each biological sample as a point in multidimensional space, in which each dimension is a gene and the coordinate of each point is the expression level of that gene in the sample. Using support vector machines, this high-dimensional space gains even more dimensions, representing the mathematical combinations of genes. The goal for support vector machines is to find a plane in this high-dimensional space that perfectly splits two or more sets of biological samples. Using this technique, the resulting plane has the largest possible margin from samples in the two conditions, therefore avoiding data over-fitting.

It is clear that within this high-dimensional space, it is easier to separate samples from two or more conditions, but one problem is that the separating plane is defined as a function using all the dimensions available. For example, the most accurate plane to separate one disease from another might be $(A \times B)^2 < 20$, where A and B are expression levels of genes. Although this might not be the most mathematically accurate way to separate two diseases, the biological significance of such functions is not always intuitive.

Relating hypotheses to techniques

As microarrays now cover the measurable genome, a user does not need to have a set of candidate genes in mind before use. However, this does not mean that a microarray user operates in a hypothesis-free manner. With all the supervised and unsupervised methods that are available for analysis, the challenge is in translating hypotheses into an appropriate bioinformatics technique.

In the domains of drug discovery and diagnostic testing, it is easy to find uses for supervised methods. Two examples illustrate this point. First, hypotheses in toxicogenomics, such as 'some genes in the genome influence liver metabolism of a particular compound', could be answered with a technique designed to find those genes for which expression measurements are most significantly different between liver cells with and without the compound. If many samples are present from each group, a nearest-neighbour approach could be used, in which an idealized gene is created that is expressed at a high level in samples with the compound and at a low level in samples without the compound. This is then used as a query to find genes that are most similar to the idealized pattern.

Second, hypotheses in the development of diagnostic tests, such as 'a combination of gene expression measurements accurately distinguishes metastatic from non-metastatic disease', can be answered by searching for such patterns using any of the supervised methods, including support vector machines. Certainly, any specific question can be answered directly in a supervised manner.

The need for unsupervised methods is less intuitive, because these start with a less direct question. For example, questions about the number and type of expression responses in a period of time after application of a compound cannot be found in a supervised manner. Unsupervised techniques, such as hierarchical clustering and self-organizing maps, survey all genes and cluster them together on the basis of their expression patterns. A search for the pairs of genes that are most likely to be co-expressed can be accomplished using a technique such as relevance networks. Finally, true genetic regulatory networks (that is, hypotheses that the expression of one gene correlates with the expression of another) might be found using methods such as constructing Bayesian networks. What is common in these three examples is that there is no ideal answer that is being sought, and that it is difficult to ascertain when the method is correct. Nonetheless, unsupervised methods could be instrumental in the early discovery process.

There are a few special cases worth noting. When attempting to find connections between dissimilar items, such as a response to a small molecule (that is, a phenotypic measurement) and expression measurements, two-dimensional hierarchical clustering can be used for each measurement separately (for example, clustering the small molecules separately from the genes, as done by Ross *et al.*[19]), or building a relevance network (as in Butte *et al.*[20]). Finally, when searching for subtypes of a condition or disease that might influence survival or a disease-free state, unsupervised hierarchical clustering can be paired with Kaplan–Meier survival statistics, as shown by Alizadeh and colleagues[49].

Challenges after analysis

After several microarray analyses, it quickly becomes obvious that the rate-limiting step in functional-genomics experiments is neither the handling of the biological samples nor the actual analysis, but instead the post-analytical work in determining what the results actually mean. First, detailed names and information might not yet be available for genes that have been found to be significant, even though these genes might have been measured on microarrays for years. This complicates the interpretation of results. The official gene name, predicted protein domains or gene-ontology[50] classification might become available as early as tomorrow, or as late as decades from now.

There are further post-analysis challenges. Occasionally, microarray probes are designed against chromosomal regions instead of expressed products, and when these probe sets are positive in an analysis, it is usually not clear which genes are being detected. It is worth finding these probe sets before analysis begins and eliminating them. Occasionally, probe sets are incorrectly designed against the wrong strand or wrong species. Oligonucleotide sequences that were once thought to be unique for a particular gene might not remain unique as more genomic data are collected. Finally, in using spotted cDNA arrays, particularly those for which the probe sequences have not been validated, the findings might be incorrect.

Operationally, this means that one is never done analysing a set of microarray data. The infrastructure has to be developed to re-investigate constantly genes and gene information from microarray analyses performed in the past. It could be next month, for example, that new information about a gene that was positive in an analysis performed three months ago finally leads to a new and important hypothesis.

Conclusion

The challenge in determining the proper analytical methods to use is usually only a short-term difficulty, and typically, after the 'functional-genomics pipeline' has been established, the rate-limiting step shifts to the post-analytical challenges[51]. In the future, truly showing a 'return on investment' from functional genomics will depend on taking findings beyond the microarray stage and integrating them with the rest of the discovery pipeline. The 'list of genes' resulting from a microarray analysis should not be viewed as an end in itself; its real value increases only as that list moves through biological validation, ranging from the numerical verification of expression levels with alternative techniques, to ascertaining the meaning of the results, such as finding common promoter regions or biological relationships between the genes. However, tools that link these genes back to known biological pathways, as well as discovering new pathways, are in their infancy. Tools that can automatically indicate the importance of particular findings have yet to be invented. Until they come into being, the analysis of microarray data sets in a vacuum devoid of biological knowledge will be less rewarding.

Finally, the use of microarrays in basic and applied research in drug discovery is only going to increase, but as these data sets grow in size, it is important to recognize that untapped information and potential discoveries might still be present in existing data sets (BOX 2). It should be clear that any set of microarray measurements could be analysed and re-analysed in many different ways. In the application of functional genomics to drug discovery, to extract the most information from microarrays, an open mind always needs to be kept with regard to the choices of analytical methods, using supervised and unsupervised techniques, and methods yet to come.

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
2. Lockhart, D. J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
3. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
4. Wu, T. D. Analysing gene expression data from DNA microarrays to identify candidate genes. *J. Pathol.* **195**, 53–65 (2001).
5. Eickhoff, B., Korn, B., Schick, M., Poustka, A. & van der Bosch, J. Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Res.* **27**, 33 (1999).
6. Zien, A., Aigner, T., Zimmer, R. & Lengauer, T. Centralization: a new method for the normalization of gene expression data. *Bioinformatics* **17** (Suppl. 1), S323–S331 (2001).
7. Li, C. & Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* **2**, research0032.1–0032.11 (2001).
This article describes normalization techniques, as well as a popular alternative quantification method for Affymetrix microarrays.
8. Ramdas, L. *et al.* Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biol.* **2**, research0047.1–0047.7 (2001).

9. Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C. & Wong, W. H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29**, 2549–2557 (2001).
10. Livesey, F. J., Furukawa, T., Steffen, M. A., Church, G. M. & Cepko, C. L. Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene *Crx. Curr. Biol.* **10**, 301–310 (2000).
11. Jelinsky, S. A. & Samson, L. D. Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl Acad. Sci. USA* **96**, 1486–1491 (1999).
12. Chen, J. J. *et al.* Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* **51**, 313–324 (1998).
13. Ishii, M. *et al.* Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* **68**, 136–143 (2000).
14. Vernon, S. D. *et al.* Reproducibility of alternative probe synthesis approaches for gene expression profiling with arrays. *J. Mol. Diagn.* **2**, 124–127 (2000).
15. Baugh, L. R., Hill, A. A., Brown, E. L. & Hunter, C. P. Quantitative analysis of mRNA amplification by *in vitro* transcription. *Nucleic Acids Res.* **29**, E29 (2001).
16. Schadt, E. E., Li, C., Su, C. & Wong, W. H. Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* **80**, 192–202 (2000).
17. Yang, Y. H., Buckley, M. J., Dudoit, S. & Speed, T. P. *Comparison of Methods for Image Analysis on cDNA Microarray Data* (Univ. California, Berkeley, 2000).
18. Emmert-Buck, M. R. *et al.* Laser capture microdissection. *Science* **274**, 998–1001 (1996).
19. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24**, 227–235 (2000).
Using dendrograms, Ross and colleagues found clusters of genes measured across the various cancer cell lines in the NCI-60 panel.
20. Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R. & Kohane, I. S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. USA* **97**, 12182–12186 (2000).
21. Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L. & Kohane, I. S. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412 (2002).
One of the first studies to compare published measurements of, in theory, the same cancer cell lines on cDNA and oligonucleotide microarrays. Shows that these measurements are not directly comparable.
22. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
23. Butte, A. J. *et al.* Determining significant fold differences in gene expression analysis. *Pac. Symp. Biocomput.* 6–17 (2001).
24. Park, P. J., Pagano, M. & Bonetti, M. A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac. Symp. Biocomput.* 52–63 (2001).
25. Pavlidis, P. & Noble, W. S. Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biol.* **2**, research0042.10–0042.15 (2001).
26. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
One of the first publications to show how microarrays can assist in difficult clinical diagnosis; in this case, determining acute lymphocytic leukaemia from acute myelogenous leukaemia using a nearest-neighbour approach.
27. Quinlan, J. *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, California, 1992).
28. Rumelhart, D., McClelland, J. & The Parallel Distributed Processing Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press, Cambridge, Massachusetts, 1986).
29. Furey, T. S. *et al.* Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914 (2000).
30. Brown, M. P. *et al.* Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA* **97**, 262–267 (2000).
31. Chow, M. L., Moler, E. J. & Mian, I. S. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol. Genomics* **5**, 99–111 (2001).
32. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA* **97**, 10101–10106 (2000).
33. Raychaudhuri, S., Stuart, J. M. & Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.* 455–466 (2000).
34. Fiehn, O. *et al.* Metabolite profiling for plant functional genomics. *Nature Biotechnol.* **18**, 1157–1161 (2000).
35. Wen, X. *et al.* Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA* **95**, 334–339 (1998).
One of the first large microarray publications, with 112 genes measured in 9 conditions, analysed using dendograms created using Euclidean distance.
36. Hilsenbeck, S. G. *et al.* Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Natl Cancer Inst.* **91**, 453–459 (1999).
37. Ben-Dor, A. *et al.* Tissue classification with gene expression profiles. *J. Comput. Biol.* **7**, 559–583 (2000).
38. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
Tamayo and colleagues were the first to use self-organizing maps to show clusters of genes measured across time from differentiating hematopoetic cells.
39. Toronen, P., Kolehmainen, M., Wong, G. & Castren, E. Analysis of gene expression data using self-organizing maps. *FEBS Lett.* **451**, 142–146 (1999).
40. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
The first group to show the now-standard Eisen-style dendrogram.
41. Liang, S., Fuhrman, S. & Somogyi, R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* 18–29 (1998).
42. Wuensche, A. Genomic regulation modeled as a network with basins of attraction. *Pac. Symp. Biocomput.* 89–102 (1998).
43. Szallasi, Z. & Liang, S. Modeling the normal and neoplastic cell cycle with 'realistic Boolean genetic networks': their application for understanding carcinogenesis and assessing therapeutic strategies. *Pac. Symp. Biocomput.* 66–76 (1998).
44. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620 (2000).
45. Butte, A. & Kohane, I. in *Fall Symposium, American Medical Informatics Association* (ed. Lorenzi, N.) 711–715 (Hanley and Belfus, Washington DC, 1999).
46. Butte, A. J. & Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput*. 418–429 (2000).
47. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
The first publication to merge several microarray experiments, to show clusters using dendrograms constructed using correlation coefficients, and to analyse the time-series pattern of genes using Fourier analysis.
48. Yeung, K. Y. & Ruzzo, W. L. *An Empirical Study of Principal-Components Analysis for Clustering Gene Expression Data* Technical Report UW-CSE-2000-11-03. (Univ. Washington, Washington DC, 2000).
49. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
Alizadeh and colleagues were the first to use microarrays to find subtypes of a single disease that could be defined only by their gene-expression patterns, and which showed significant differences in patient mortality.
50. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
51. Kohane, I. S., Kho, A. T. & Butte, A. J. *Microarrays for an Integrative Genomics* (MIT Press, Cambridge, Massachusetts, 2002).
52. Perou, C. M. Show me the data! *Nature Genet.* **29**, 373 (2001).

Acknowlegements

The author wishes to thank T. Deshpande, A. Kho, M. Ramoni and I. Kohane for critical comments and interesting discussions on the manuscript. During the writing of this work, the author has been funded by and wishes to thank the Endocrine Fellows Foundation, the Genentech Centre for Clinical Research and Education, the Lawson Wilkins Paediatric Endocrinology Society, the Harvard Centre for Neurodegenerative Research and the Merck–Massachusetts Institute of Technology partnership. The author was also supported in part by grants from the National Heart, Lung and Blood Institute, the National Institute of Diabetes and Digestive and Kidney Diseases, and the National Institute of Neurological Disorders and Stroke.

Online links

DATABASES

Cancer.gov: http://www.cancer.gov/cancer_information/
acute lymphocytic leukaemia | acute myelogenous leukaemia
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/
p53

FURTHER INFORMATION

CardioGenomics:
http://cardiogenomics.med.harvard.edu/public-data.html
Eisen's laboratory: http://rana.lbl.gov/EisenSoftware.htm
GeneCluster 2.0:
http://www-genome.wi.mit.edu/cancer/software/genecluster2/gc2.html
National Cancer Institute: http://www.cancer.gov/
RELNET: http://www.chip.org/relnet
Access to this interactive links box is free online