Primer on Medical Genomics
Part III: Microarray Experiments and Data Analysis

AYALEW TEFFERI, MD; MARK E. BOLANDER, MD; STEPHEN M. ANSELL, MD, PHD; ERIC D. WIEBEN, PHD; AND THOMAS C. SPELSBERG, PHD

Genomics has been defined as the comprehensive study of whole sets of genes, gene products, and their interactions as opposed to the study of single genes or proteins. Microarray technology is one of many novel tools that are allowing global and high-throughput analysis of genes and gene products. In addition to an introduction on underlying principles, the current review focuses on the use of both complementary DNA and oligodeoxynucleotide microarrays in gene expression analysis. Genome-wide experiments generate a massive amount of data points that require systematic methods of analysis to extract biologically useful information. Accordingly, the current educational communication discusses different methods of data analysis, including supervised and unsupervised clustering algorithms. Illustrative clinical examples show clinical applications, including (1) identification of candidate genes or pathological pathways (ie, elucidation of pathogenesis); (2) identification of "new" molecular classes of diseases that may be relevant in disease reclassification, prognostication, and treatment selection (ie, class discovery); and (3) use of expression profiles of known disease classes to predict diagnosis and classification of unknown samples (ie, class prediction). The current review should serve as an introduction to the subject for clinician investigators, physicians and medical scientists in training, practicing clinicians, and other students of medicine.

***Mayo Clin Proc*. 2002;77:927-940**

ALL = acute lymphocytic leukemia; AML = acute myeloid leukemia; bp = base pair; cDNA = complementary DNA; CLL = chronic lymphocytic leukemia; cRNA = complementary RNA; DLBCL = diffuse large B-cell lymphoma; EST = expressed sequence tag; FL = follicular lymphoma; mRNA = messenger RNA; PCR = polymerase chain reaction; SOM = self-organizing map

Microarray analysis embodies the essence of genomics by allowing simultaneous (in a single assay) study of many genes and gene products.[1] A microarray system may be applied to global analysis of genomic DNA,[2,3] expressed DNA (RNA),[4,5] or translated DNA (protein).[6-10] All this is made possible by the current availability of a powerful set of technologies that includes miniaturized interrogating (ie, probe linked) of solid substrates and computer-assisted scanning and imaging devices.[11] Robust methods of computational analysis are being developed to interpret the large amount of data generated from microarray experiments and to extract biologically useful information.[12-14]

In DNA microarray analysis, an altered gene or biochemical pathway associated with a particular disease may be revealed by the identification of a consistently up-regulated or down-regulated gene across a cohort of patients with the same disease. Once identified, the aberrant functional pathway may be targeted in the development of novel therapeutics (drug discovery).[15] Similarly, analyzing gene expression patterns across individual patients with the "same" disease may reveal molecular-level differences that may allow refinement of current disease classification, prognostication, and treatment selection.[5,16-19] Other applications of DNA microarray analysis include gene identification,[20,21] screening for DNA mutations[22,23] or polymorphisms,[24] and comparative genomic hybridization.[25] The current article focuses on gene expression profiling of diseased tissue as a prelude toward gaining pathogenetic insight into the disease and identifying molecularly distinct classes of individual disease categories. The methods described in this communication are based on a review of the literature.

BASICS IN DNA MICROARRAY TECHNOLOGY

A microarray or macroarray is an orderly arrangement (ie, the coordinates are known) of a usually rectangular grid of "spots" (cells, features). In DNA microarray analysis, data are usually presented by columns and rows; columns represent different samples, patients, or experiments, and rows represent genes (Figure 1).[26] Therefore, a cell (feature) represents the quantification of a gene expression for a given gene in a given sample.[26] Thousands of microarray

From the Division of Hematology and Internal Medicine (A.T., S.M.A.), Department of Orthopedic Surgery (M.E.B.), and Department of Biochemistry and Molecular Biology (E.D.W., T.C.S.), Mayo Clinic, Rochester, Minn. The authors are members of the Mayo Clinic Genomics Education Steering Committee.

Individual reprints of this article are not available. The entire Primer on Medical Genomics will be available for purchase from the *Proceedings* Editorial Office at a later date.

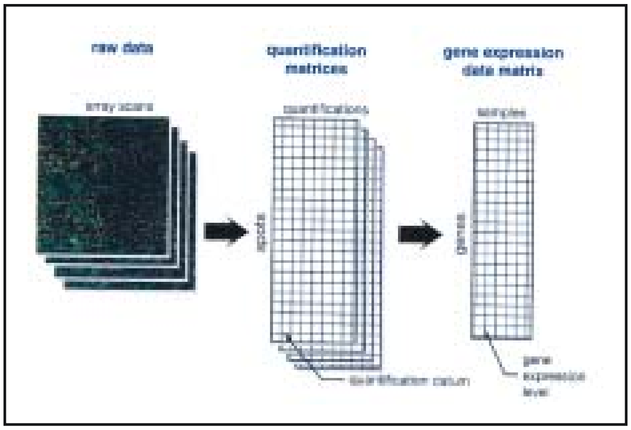

Figure 1. Three levels of microarray gene expression data processing. The spots on the actual microarray are imaged after DNA-DNA hybridization procedures, and the raw data are subsequently quantified for spot intensity and processed onto a quantification matrix with rows (representing genes) and columns (representing samples, patients, or experiments). Reprinted with permission from Brazma et al.[26]

spots may be ordered by printing small fragments of DNA (probes) onto a small solid matrix (chemically coated glass, nylon membrane, silicon) using modern technology. The term *chip* is arbitrarily used for solid substrate units with small dimensions (usually <4 × 4 cm). The number of spots on a single microarray chip depends on the size of the chip and the spotting technique used. Each spot on a DNA microarray is less than 250 μm in diameter (compared with >300 μm in a macroarray) and carries millions of copies of a specific probe.

The DNA molecule that is tethered (embedded, immobilized) to each spot on a microarray chip is called a *probe*. A coating of polylysine or silane on a glass substrate facilitates adhesion of the DNA probe. The test samples (usually complementary DNA [cDNA] representing messenger RNA [mRNA] from study samples) are called *targets*. The interaction (hybridization by base pairing) between probes and targets defines the experiment.[27]

Various sources of DNA may be used as probes. Although genomic DNA may be an acceptable source in prokaryotes, the presence of introns and intergenic regions in eukaryotes makes genomic DNA difficult to use in higher animals. Regardless, the choice of probes also depends on the objective of the experiment. For the study of single nucleotide polymorphisms, for example, the use of genomic DNA is essential. On the other hand, microarray probes for the study of gene expression patterns are usually prepared from expressed DNA sequences (cDNA clones, expressed sequence tags [ESTs]). To ensure genome-wide representation, the information to prepare probes is derived from comprehensive public databases (UniGene, GenBank, dbEST). For human experiments, the UniGene database is preferred because of a low redundancy of representation and its superb organization, which allows systematic evaluation of the results of the experiment by providing full information on the expressed genes.

At present, there are 2 distinct methods of spotting DNA probes on chips. In one, polymerase chain reaction (PCR)–amplified ESTs (or whole cDNA clones extracted from plasmid-containing bacterial cultures) (500-5000 base pairs [bp] long) are deposited (nanoliter quantities per spot) on the microarray spots by using high-speed robotics (Figure 2).[28] This is the method used in comparative cDNA microarrays. In the second method, the probes are oligodeoxynucleotide sequences (20-50 bp long) that are synthesized in situ on the chip itself. The aforementioned databases are used to obtain the information on the gene-specific sequences that are used to construct the oligonucleotide probes. This is the method used in oligonucleotide microarrays (oligo chips), which is described in detail in the next section.

PHOTOLITHOGRAPHY AND THE OLIGODEOXYNUCLEOTIDE CHIP

Affymetrix Corporation (Santa Clara, Calif) produces a high-density oligo chip (GeneChip) that is less than 1.5 × 1.5 cm and carries several hundred thousand spots, representing more than 15,000 human genes. The probes are synthesized in situ with use of photolithography and DNA chemical analysis. The process of photolithography requires a light source and a filmlike apparatus (a mask) that transmits light in a pattern that follows a specific design. This technology is borrowed from the computer chip industry. When a computer chip is being made, the mask is first prepared in a larger scale by designing a circuitry pattern on an opaque chromium film that rests on a glass. A UV beam of light is then transmitted through the mask, and the passing light is focused onto a photosensitive polymer (photoresist) that rests on a silicon wafer.[29] Thus, the original pattern of the mask is captured and replicated in a miniaturized form on the silicon by removing the light-exposed parts of the photoresist.

In making the oligo chip, a glass slide coated first with a covalent linker molecule that is covered by a photolabile protector is selectively activated at different spots on the chip by UV light shining through a mask with a predesigned pattern (Figure 3).[30] A set of photomasks with different pattern designs is used to expose specific spots for specific nucleotide attachments. The base nucleotides are themselves photoprotected (photosensitive hydroxyl-protected deoxynucleotides that are tethered at the 5′ end and ready to be light activated at the 3′ end) for subsequent light-directed, mask-determined nucleotide attachment

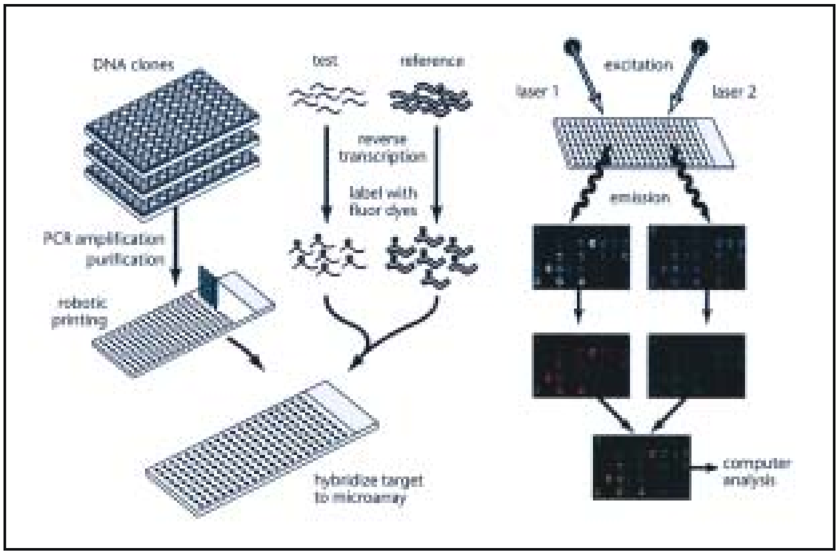

Figure 2. Complementary DNA microarray experiment. See text for detailed description. PCR = polymerase chain reaction. Reprinted with permission from Duggan et al.[28]

(Figure 3). The whole process of selective light activation using different photomasks and coupling is repeated to lay nucleotides 1 at a time on a growing chain (in situ oligonucleotide synthesis).[30] Therefore, such chips are called oligo chips, in comparison to cDNA chips in which the probes are prepared separately by a PCR amplification process of ESTs or cDNA clones (cDNA chips).

The Affymetrix Hum6000 chip (1 type of oligo chip) is 1.28 × 1.28 cm and contains approximately 65,000 spots (features), each spot measuring 24 × 24 µm and containing approximately 10 million 25 mer oligodeoxynucleotide probes. A set of 4 such chips contains approximately 6817 genes or ESTs. One gene or EST is represented by a probe set that is made up of 20 different spots (features) with probes that differ in their sequences but are all complementary to different domains of the same specific target gene or EST (ie, the 20 different probes represent different segments of the same gene or EST) (Figure 4).[30] This is important because the oligo probes are small enough to cross-hybridize, and the representation of a gene with multiple oligomers (ie, the probe set) increases the accuracy of gene detection and quantification. Furthermore, each feature in a probe set is accompanied by a neighbor feature (making a pair), with probes having a single nucleotide mismatch at the center distinguishing the members of each pair (the so-called perfect match and mismatch features) (Figure 4).[30] In other words, there are 20 feature pairs per gene or EST. This multiplication of features per gene or EST helps in internal quality control of the hybridization process described in the next section.

MICROARRAY EXPERIMENT FOR GENE EXPRESSION ANALYSIS

In a DNA microarray experiment, the basic requirements are a microarray chip and an RNA sample. Either total RNA or enriched mRNA samples may be used. However, only 3% of total RNA is represented by mRNA, and therefore extracting mRNA from patient samples in an amount that is adequate for microarray analysis (approximately 5-100 µg per sample) may be difficult. In general, cDNA microarray experiments require more than 50 µg of total RNA from target tissues, whereas 5 µg of RNA may be adequate for an oligo chip microarray experiment. Recently described target amplification methods that use in vitro transcription (sample RNA–cDNA–complementary RNA [cRNA]) may allow the use of even smaller amounts of test sample (approximately 1-50 ng of total RNA).[11]

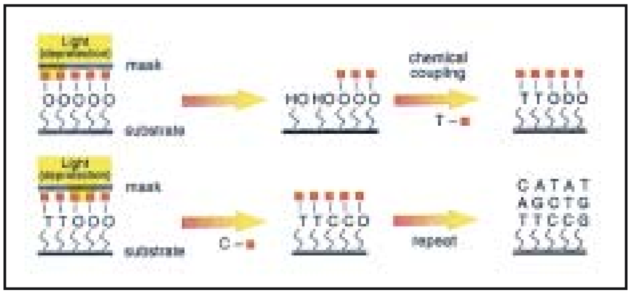

Figure 3. Light-directed oligonucleotide synthesis on a glass slide. See text for detailed description. Reprinted with permission from Lipshutz et al.[30]

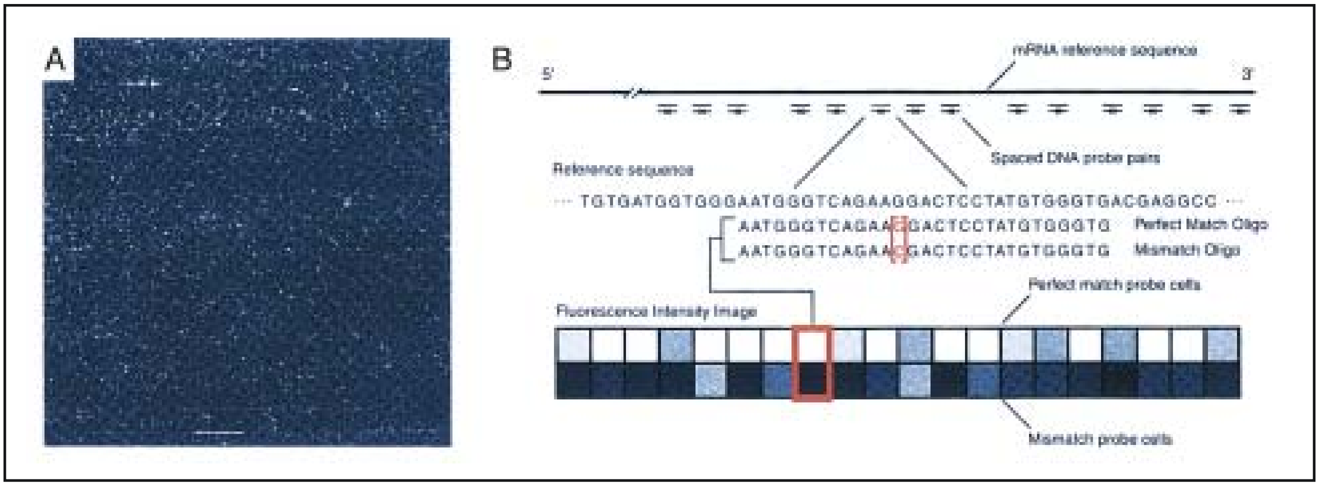
A
B
5′
3′
mRNA reference sequence
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
AATGGGTCAGAAGGACTCCTATGTGGGTG
Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTGGGTG
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 4. Gene expression monitoring with oligonucleotide arrays. A, Single 1.28 × 1.28-cm array containing probe sets for approximately 40,000 human genes and expressed sequence tags (ESTs). This array contains features (cells) smaller than 22 × 22 μm and only 4 probe pairs per gene or EST. B, Expression probe and array design. Oligonucleotide probes are chosen based on uniqueness criteria and composition design rules. For eukaryotic organisms, probes are chosen typically from the 3′ end of the gene or transcript, near the poly-A tail, to reduce problems that may arise from the use of partially degraded messenger RNA (mRNA). The use of perfect match minus mismatch differences averaged across a set of probes greatly reduces the contribution of background and cross-hybridization and increases the quantitative accuracy and reproducibility for the measurements. See text for further discussion. Reprinted with permission from Lipshutz et al.[30]

In cDNA microarray, a control sample is usually required for simultaneous analysis with the test sample (ie, 2-sample analysis on a single chip). Thus, gene expression is estimated by comparing the amount of mRNA content in 2 different cell populations (a test sample vs a control sample), and the measurement is given as a ratio. In contrast, in oligo chip analysis, the usual procedure is to analyze the test sample on 1 oligo chip and the control or reference sample on another oligo chip. In other words, the 2 samples on 1 cDNA chip can be viewed as comparable to 2 samples on 2 oligo chips. Therefore, interpretation of gene expression with oligo chips requires comparison of levels from the test sample to those of the control or reference sample. Regardless, both methods of a microarray experiment (cDNA microarray vs oligo chips) consist of several steps, including (1) target preparation (extracting nucleic acids from biological samples and labeling them with either fluorescent dyes for glass array or radioactive isotopes for nylon filters), (2) hybridization (incubation of the labeled targets with cDNA or oligodeoxynucleotide probes on the surface of the chip), (3) scanning (computer-assisted reading of signal intensity that is emitted from labeled targets that are hybridized to probes on the chip surface; laser scanner for fluorescence or PhosphorImager for radioactive isotopes), and (4) computational analysis to extract biologically useful information from the vast quantity of data generated.

Target Preparation: cDNA Microarray vs Oligo Chips

In cDNA microarray experiments, the initial step is to synthesize cDNA (by reverse transcription) from total RNA or mRNA extracted from both the test and the reference samples in the presence of nucleotides that are differentially labeled (test vs control sample) with reporter molecules (fluorochromes for chip arrays and radioactive isotopes for membrane arrays) (Figure 5).[31] The labeled cDNAs therefore represent portions or complete segments of the original mRNA or total RNA from the samples and are comparatively more stable. This method requires relatively more RNA per sample (approximately 100 μg).

Differential labeling is not necessary for microarray experiments using oligo chips. Sample mRNA (approximately 5 μg) is first reverse transcribed into single-stranded cDNA in the absence of labeled nucleotides. The single-stranded cDNA is then converted to a double-stranded cDNA that is in turn transcribed to cRNA (ie, in vitro transcription) in the presence of biotinylated nucleotides. This additional step (compared with cDNA microarray) amplifies the target molecules (original mRNA) by approximately 50-fold.[31]

Hybridization: cDNA Microarray vs Oligo Chips

In cDNA microarray, the test and reference cDNA samples, which are differentially labeled with fluorochromes, are mixed in equal amounts and incubated with a

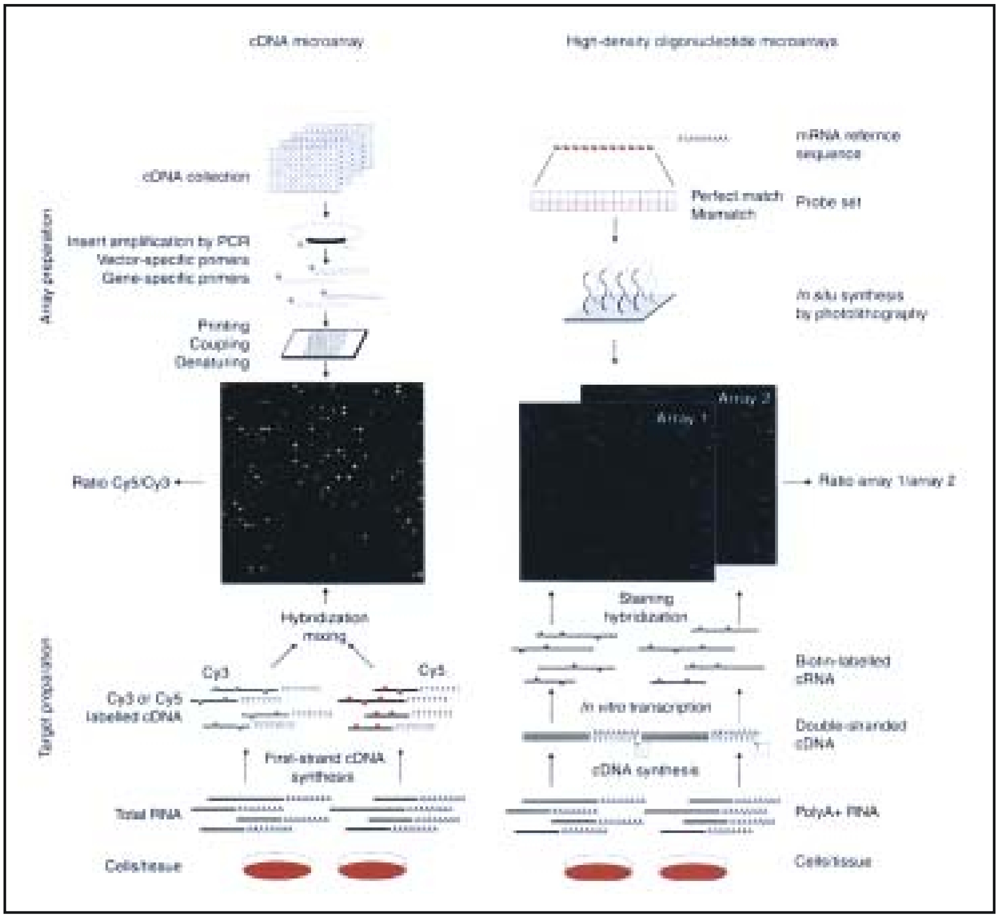

Figure 5. Comparison of the steps in complementary DNA (cDNA) microarray experiment vs oligonucleotide chip analysis. See text for a detailed explanation. cRNA = complementary RNA; mRNA = messenger RNA; PCR = polymerase chain reaction. Reprinted from Schulze and Downward[31] with permission from MacMillan Magazines, Ltd.

microarray slide for hybridization to occur (Figure 2).[28] Each spot on the microarray chip contains enough (millions) DNA copies to allow probe hybridization from both samples without interference. The abundance of specific targets (and therefore the amount of mRNA from the original samples) in the test vs the control sample will dictate the amount of binding to specific probes, and the difference in target (mRNA) content between test and control samples at a specific spot will be determined by the difference in the content of the corresponding reporter molecules.

In oligo chip microarray experiments, the biotinylated cRNA is first hybridized to the oligodeoxynucleotide probes on the glass slide, followed by binding to an avidin-conjugated fluorophore. The abundance of the target molecule is estimated by measuring raw intensities of the fluorescence emitted by the linked reporter molecules (Figure 5).[31]

Scanning: cDNA Microarray vs Oligo Chips

After hybridization is completed, fluorescence from array spots with successful probe-target linkage is detected and digitally imaged. A laser beam is used to excite the fluorescent markers linked to the hybridized target molecules, and the degree of fluorescence correlates with the abundance of target molecules at a specific spot. Fluorescent emission is monitored by a scanner, which also processes the image. The raw intensity of fluorescence from each spot is quantified by either a charge-coupled device camera or a photomultiplier tube, which converts light energy into an electrical signal. High-resolution imaging is optimized by use of confocal microscopy. The recorded intensity is saved as a tagged image file format (the intensity of a given pixel is proportional to the amount of signal coming from the corresponding point on the glass chip).

In cDNA microarray, the test sample is labeled with the fluorochrome Cy3 (rhodamine, with a fluorescence emis-

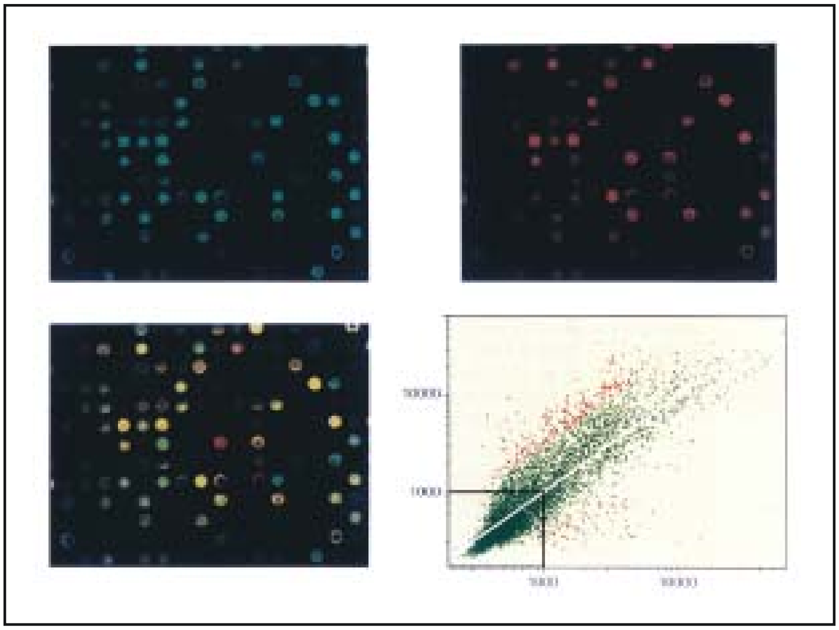

Figure 6. Microarray images in complementary DNA microarray experiment. Reference and test samples are differentially labeled (green and red), and an integrated composite image displays the relative abundance of targets at specific spots displayed by pseudocolor images. Reprinted with permission from Brockman and Tamminga.[32]

sion wavelength of 565 nm) and the control sample with a different fluorochrome, Cy5 (fluorescein, with an emission wavelength of 670 nm). Such double labeling allows assessment of expression in the test sample in relation to expression in the control sample (Figure 5). Independent images (one for Cy3 and the other for Cy5), using the respective excitation wavelengths for generating fluorescence with characteristic emission wavelengths, are then generated and subsequently digitally integrated to produce a composite image that allows measurement of the ratio of the target molecules in the 2 samples (test and control). These ratios can be presented as either Microsoft Excel files or pseudocolor images that assign red for genes that are expressed higher in the test sample than that of the control sample and green for genes that are expressed lower in the test sample compared with those in the control sample (Figure 6).[32]

In the case of oligo chips, the absolute intensity of fluorescence from each spot on the glass chip constitutes the raw data that are subject to further analysis. Scanning devices are often equipped with imaging software that captures spot-specific signal intensity and adjusts for both background intensity and variance of pixel intensity within a specific spot. Additional measures to ensure data quality include the use of replicate pairs of genes in a single array or replicate arrays.[33,34]

Normalization: cDNA Microarray vs Oligo Chips

Raw signal intensity, either from cDNA or oligo chip, must be adjusted to a common standard (normalized) to correct for differences in overall array intensity that include background noise as well as differences in efficiency in detection and data acquisition. In other words, individual chips from many patients or experiments must be comparable in other aspects before a legitimate comparison of gene expression is made. After normalization, the raw gene expression levels are presented as an expression ratio of test vs control sample, or the gene expression profiles from several samples may be compared with a clustering algorithm. Ratios are typically log-transformed to simplify presentation of bidirectional fold differences (ie, produce a normal distribution).[13]

There are several methods of normalization. Background noise subtraction is uniformly applied to all methods.[30] In addition, some investigators assume minimal cell-to-cell variation in the expression of housekeeping genes, and therefore the spot intensities in each array are rescaled accordingly.[35] Another method rescales spot intensities based on the average spot intensity of either the entire chip or a set of probes representing the same gene.[30,33] This particular method assumes that the total number of targets that hybridize to probes is the same for both the test and the reference sample, which, therefore, should have the same total integrated intensity. Alternatively, each array experiment may be coupled with another experiment with a common reference sample to act as a control for normalization purposes.

In GeneChip, each spot (containing a perfect match probe) is accompanied by another neighboring cell that contains a mismatch probe (an oligonucleotide probe that differs by 1 bp at a central position). The difference in spot intensity between the 2 samples is used to account for background noise and nonspecific cross-hybridization (Figure 4).[30]

MICROARRAY DATA ANALYSIS

After normalization, spot intensities or ratios are converted to a table with a numerical value that is suitable for further statistical analysis. The table, also called a matrix, is made up of rows (representing genes) and columns (representing patients or experiments). Therefore, each cell represents an expression value (absolute spot intensity or ratio) that can be color coded as red (relatively overexpressed) or green (relatively underexpressed). As is, these raw data, whether in the form of numbers or colors, are completely unintelligible (Figure 1).[26] The primary purpose of a microarray data analysis, based on various statistical techniques, is to extract order based on similarities and differences in expression.

The first and most important step in microarray data analysis is to define the purpose of the exercise. Current use of microarray data analysis in medical genomics has focused on 3 separate objectives: (1) identification of candidate genes or pathological pathways (ie, elucidation of pathogenesis); (2) identification of "new" molecular classes of diseases that may be relevant in disease reclassification, prognostication, and treatment selection (ie, class discovery); and (3) use of expression profiles of known disease classes to predict diagnosis and classification of unknown samples (ie, class prediction).

Searching for Candidate Genes

The gene expression profile of normal and disease tissue can be compared to identify genes that are differentially displayed. The information regarding differential display is obtained from a single microarray in the case of cDNA microarray analysis as long as the reference sample used represents the normal tissue counterpart of the disease tissue under study.

With oligo chips, raw data from 2 microarrays (one from diseased tissue and another from normal tissue) are compared to identify genes that are differentially expressed. Usually, the initial step involves some degree of data filtration. First, a threshold pixel intensity (usually 1500-3000 pixels) is set, based on background intensities and factors that relate to other experimental conditions, and genes that show expression values below this threshold level are discarded. This step assumes that the genes of interest are adequately expressed at least in some patients or reference samples. Next, either an arbitrary fold-based difference (>2-, 5-, or 10-fold difference in spot intensity) or a statistical test (eg, *t* or F test) is used to select genes that show statistically significant variation in expression among patients or experiments. The *t* test measures the difference in mean expression values between 2 samples and allows the identification of genes with a significant difference. One can also use a paired *t* test for samples from the same origin (eg, normal and diseased tissue from the same patient) or a nonparametric test (eg, Mann-Whitney test) if one assumes a non-Gaussian distribution of data. The difference in mean expression values among 3 or more samples is measured by the F test (Kruskall-Wallis test for nonparametric data).

Subsequently, genes that show little variation in expression across patient samples or experiments are discarded. The remaining genes become the genes of interest for further analysis that includes confirmation of overexpression by real-time PCR. The next step is to group (cluster) the selected genes of interest based on similarity of expression. Such grouping may identify genes that may be coregulated and therefore functionally related. Such information may lead to elucidation of pathologic pathways and drug targets.

Defining Gene Expression Profiles to Facilitate Class Discovery and Class Prediction

The clustering of similarly expressed genes may generate a pattern (profile) that may be useful in the separation of distinct phenotypes, classes, or stages of disease.[14] Both gene and sample (patients or experiments) clusters are based on similarity of gene expression.

The underlying principle is for each measure of gene expression (spot intensity in a matrix cell) to be represented by an expression vector in an expression space with "n" dimensions, in which "n" is the number of patients or experiments.[13] This will allow each gene to be represented by a point in expression space where the geometric coordinates are defined by the expression vectors from each patient or experiment. Accordingly, the similarity between 2 genes, and therefore 2 points in expression space, is deduced from their proximity to each other.[14,36]

One method of measuring the distance between 2 data points (known as the euclidean metric distance) is to calculate the square root of the sum of the squared differences in the expression vectors of each patient or experiment [euclidean distance = $\sqrt{\ }\, {}^{n}\sum_{i = 1} (x_i - y_i)^2$, in which "$x_i$" and "$y_i$" are the measured expression values for genes X and Y in experiment "i" and "n" is the number of patients or experiments].[13,36] The euclidean distance metrics are best suited for data that are normalized for degree of expression despite different absolute levels. Such data transformation may be unnecessary in an alternative method called the Pearson coefficient of correlation (*r*). The latter method compares genes according to similarity in shape (ie, the direction in expression) rather than absolute magnitude of expression. In contrast, euclidean metrics measure the absolute distance between 2 expression vectors (points) in space. Closely residing genes in the expression space (ie, genes with similar expression vectors) can then be grouped (clustered) according to various multivariate statistical methods.

Several statistical algorithms have been used in generating patterns (ie, expression profiles) of gene expression that may be used in identifying coregulated genes and different disease classes (or sample categories). Such gene or sample organization (clustering) may be either unsupervised (not based on prior information) or supervised (based on prior information).[37]

Unsupervised Cluster Analysis

In unsupervised cluster analysis, the statistical algorithm is not trained to recognize a specific gene expression pattern from a previously known class of genes or group of patients (or experiments) that may be used to classify new members (ie, unknown samples). In other words, unsupervised clustering is designed to discover clusters of similar genes or similar samples (ie, facilitates class discovery). Unsupervised clustering may be hierarchical (classification with nested classes resembling a phylogenic tree) or nonhierarchical (classification into clusters without specifying the relationship between individual members of a class). Hierarchical clustering may be agglomerative (a clustering mechanism that starts from single members and their relationship with each other and grows into bigger classes) or divisive (a clustering technique that starts from grouping all members in 1 class first and subsequently breaking the class into smaller groups).

Hierarchical Clustering

Hierarchical agglomerative (aggregative) clustering (not divisive hierarchical clustering) is the most commonly used clustering method in gene expression analysis. The particular method starts with identifying a gene pair with the most similar expression across patients or experiments (Figure 7).[38] The euclidean distance (or alternative distance metrics) between 2 genes in expression space is used for this purpose, and the identified gene pair undergoes fusion (with a node connecting the 2 genes), and the composite expression vector (represented by the mean of the 2 expression levels) is considered as a single element toward further similar analyses (Figure 7).[38] The process is repeated until all elements are included in the analysis and a single dendrogram (a tree) is assembled (Figure 8).[13,14,39] The result is a single hierarchical tree (a dendrogram) in which intercluster distance (branch length of the dendrogram) directly correlates with dissimilarity in gene expression. Of note, the distance between 2 clusters may be calculated by several other previously reported methods.[13] Also, member positions in neighboring clusters are not necessarily fixed and may vary depending on other alternative permutations.

Similarly expressed genes are then ordered next to each other and visualized with color-coded rows (red for overexpressed and green for underexpressed genes) (Figure 8). Such hierarchical clustering can also be applied to samples instead of genes (ie, sample clustering) based on similarity of expression across different genes (ie, 2-dimensional clustering; Figure 9).[40] Therefore, hierarchical agglomerative clustering results in organizing genes into functional categories[12] as well as in the classification of disease types and subtypes.[5] Such an approach may lead to the revelation of clinically important and previously unknown molecular classes within 1 disease group (ie, class discovery).[5]

In divisive hierarchical clustering, the opposite of the aggregative method is performed. The entire set of genes or samples is considered a single cluster and is broken down first into 2 groups, then into 4, and so forth until all elements have been separated into single elements (Figure 7).[38] The process involves random selection of defined vectors for the initial clusters and assignment of genes to specific clusters based on their similarity to the reference vectors. With each assigned gene, the reference vectors are redefined to establish a new average, thus ensuring similarity of genes in a given cluster. The result is a binary tree with information similar to that obtained by aggregative clustering.[41,42]

Nonhierarchical (Partitional) Clustering Algorithms

Partitional techniques require predetermination of the total number of clusters (classes) and amass data elements into these predetermined clusters rather than organize the genes or samples into a dendrogram. One such example is k-means clustering.[13,43] First, all members (genes in this case) are randomly assigned to one of a fixed number of classes. Second, an average value (expression vector) is calculated for each cluster and is used to calculate intercluster distance. Third, individual members are moved from cluster to cluster to group closely related members together (ie, minimize the overall within-cluster dispersion). Thus, k-means clustering is a divisive and unsupervised clustering algorithm.

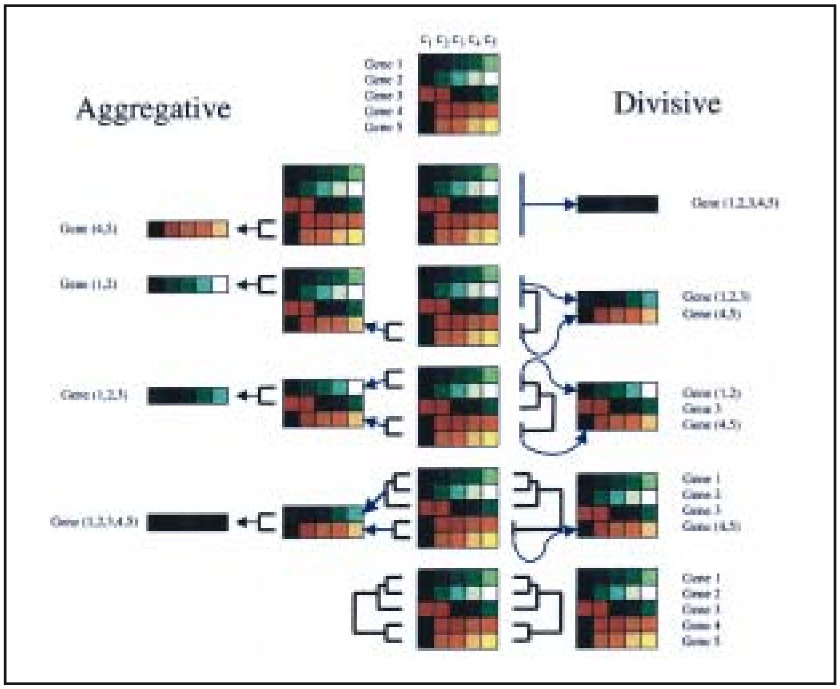

Figure 7. Aggregative and divisive methods of cluster analysis. The color codes represent the measured fluorescence ratios. Genes with unchanged expression are colored black. Red represents relatively high and green represents low gene expression. The intensity of the colors reflects different degrees of expression. The figure shows the expression levels of 5 different genes (genes 1 through 5) under 5 different conditions (c_1-c_5). In the aggregative method of cluster analysis, genes with similar expression (eg, genes 4 and 5 or genes 1 and 2) are fused and subsequently represented by an average row designation. The opposite is performed in a divisive cluster analysis. See text for a more detailed explanation. Reprinted from Dopazo et al[38] with permission from Elsevier Science.

A similar divisive and unsupervised method known as a self-organizing map (SOM) assigns genes to a predetermined number of classes based on the similarity of their expression vectors to reference vectors that are previously defined for each class.[44] In SOM, the reference vectors are recalculated with the assignment of a new member to the class so that the new reference vector is even more similar to the expression vector of the new member of the class. Such adjustment of reference vectors also affects nearby classes that are accordingly moved to new positions, and the result is a group of clusters that contain the most similar genes (or samples) with the least intercluster dispersion (Figure 10).[39] Another unsupervised method of clustering, principal component analysis, is a mathematical technique that takes into account redundancy in gene expression across patients or experiments and thus diminishes the dimensions of expression space without significant loss of information.[45] Principal component analysis allows optimal visual separation of clusters and works best when combined with other clustering algorithms, including k-means clustering or SOM (Figure 11).[13]

Supervised Cluster Analysis

Unsupervised clustering techniques *generate* classes in terms of both genes and samples. In contrast, supervised statistical algorithms *predict* classes based on a priori knowledge. The procedure starts with training a classification machine (ie, a specialized statistical algorithm such as a support vector machine) to recognize specific gene expression patterns of known gene classes or patient groups (eg, normal vs diseased tissue, different disease types). The machine learns to distinguish between members and nonmembers of a specific class based on expression data. The machine then constructs a classifier (discriminator) that accurately assigns a new member (an unknown sample) to a predefined class (Figure 12).[46] Such a classifier may then be used in assigning genes to functional classes and diseases into predefined categories that might predict diagnosis, stage, prognosis, or appropriate therapy.[16,47-49]

The following section provides 3 examples that illustrate the use of gene expression profiling in clinical medicine, including disease classification and prognostication at the molecular level.

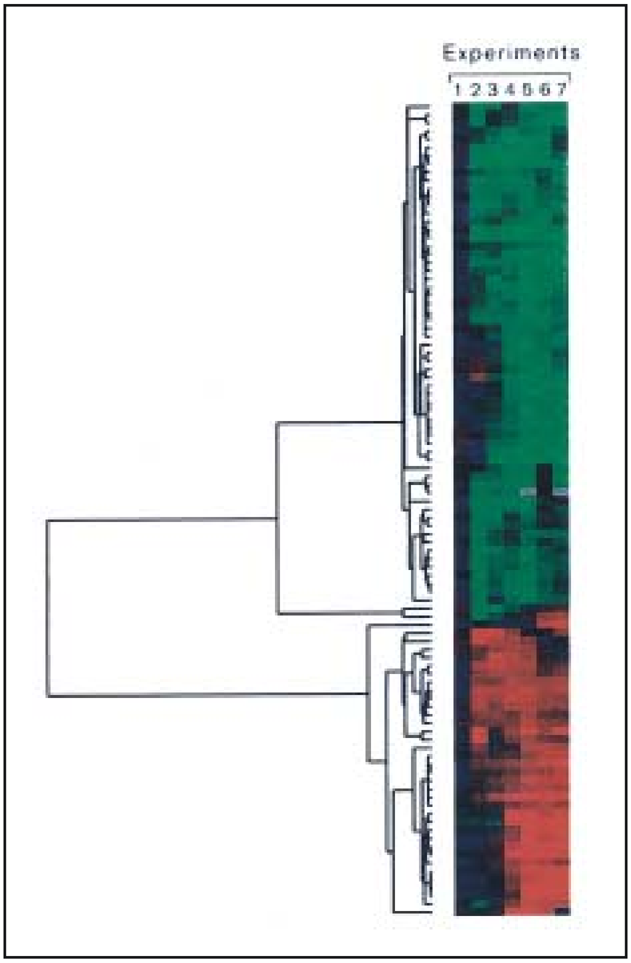

Figure 8. Aggregative cluster. Reprinted from Sherlock[39] with permission from Elsevier Science.

ILLUSTRATIVE CLINICAL EXAMPLES

Example 1[5]

Disease.—B-cell non-Hodgkin lymphoma.

Background.—B-cell non-Hodgkin lymphoma is clinically heterogeneous with some variants, including chronic lymphocytic leukemia (CLL) and follicular lymphoma (FL), which have an indolent but incurable disease phenotype and other variants, including diffuse large B-cell lymphoma (DLBCL), which has a more aggressive but sometimes curable disease phenotype. With modern combination chemotherapy, DLBCL has a 5-year relapse-free survival rate of less than 50%.

Purpose.—To test whether gene expression profiles can accurately distinguish among previously established B-cell non-Hodgkin lymphoma classes and reveal additional molecular classes of DLBCL with different clinical courses (ie, assist in class discovery).

Methods.—A cDNA microarray analysis was performed by using a special chip (Lymphochip) that contained gene probes that are selected from cDNA libraries prepared from normal and malignant lymphocytes at different stages of differentiation and activation (a total of 17,856 cDNA clones). Ninety-six mRNA samples were analyzed, including 62 patient samples (42 DLBCL, 11 CLL, 9 FL) and 34 control samples derived from either normal lymphocytes or tumor cell lines. The reference mRNA sample that was mixed with each one of the 96 test samples was prepared from a pool of mRNA from 9 different lymphoma cell lines.

Results.—Based on global similarity in gene expression patterns, a hierarchical agglomerative clustering algorithm accurately segregated the morphologically recognized classes of lymphoma (DLBCL vs FL vs CLL) (Figure 10).[5] In addition, restriction of the clustering algorithm to genes that define germinal center B cells revealed 2 distinct gene expression patterns among patients with DLBCL. This molecular subclassification of DLBCL provided independent prognostic information. The gene expression patterns of morphologically different lymphoma subclasses mimicked those from normal lymphocytes at different stages of differentiation and activation.[40]

Conclusion.—Morphologically distinct disease categories may display equally distinct gene expression profiles. The establishment of such expression profiles may help in accurately classifying new cases (class prediction). Similarly, clinically distinct phenotypes of an individual disease class may be revealed by gene expression profiling (class discovery). Furthermore, considerable pathogenetic insight may be obtained by comparing gene expression patterns of diseased and normal tissue.

Example 2[16]

Disease.—Acute leukemia.

Background.—Acute leukemia is a deadly disease (10%-20% 5-year survival rate in adults) that can be classified into acute lymphoid leukemia (ALL) and acute myeloid leukemia (AML). Both prognosis and treatment differ in ALL vs AML.

Purpose.—To test whether gene expression profiles correlate with morphologically recognized subclasses of acute leukemia (ie, class discovery) and whether generation of gene expression profiles from morphologically distinct ALL and AML cases allows class prediction of unknown samples.

Methods.—An oligo chip array (Affymetrix) containing 6817 genes was used to array total RNA from bone marrow samples of 27 cases of ALL (all children) and 11 cases of AML (all adults). For the purpose of class discovery, a nonhierarchical clustering technique (SOM) was

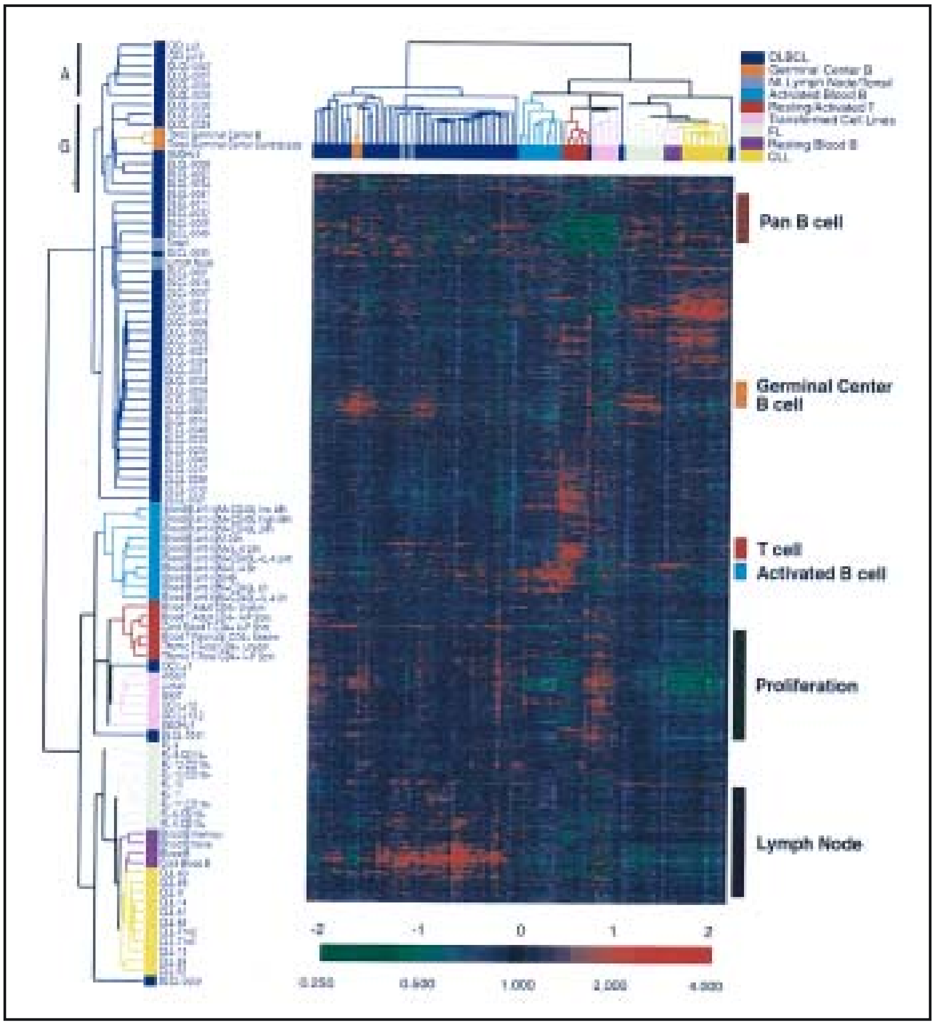

Figure 9. Hierarchical clustering schema depicting relationships between 96 samples of normal and malignant lymphocytes. The dendrogram on the left lists the samples studied and provides a measure of the relatedness of gene expression in each sample. The dendrogram is color coded according to the category of messenger RNA sample studied (see upper right key). Of note, the dendrogram along the top of the color map is the same as the one along the side but rotated and inverted. CLL = chronic lymphocytic lymphoma; DLBCL = diffuse large B-cell lymphoma; FL = follicular lymphoma; Nl = normal. Reprinted from Alizadeh et al[40] with permission from John Wiley & Sons, Ltd.

used with a priori specification of either 2 or 4 clusters and application of data filtration that discarded genes that showed less than a 5-fold difference in expression among the test samples. For the purpose of class prediction, specific genes (1100 genes) that are differentially expressed in ALL vs AML were identified by a modified clustering technique called *neighborhood analysis*. Of these, the 50 stronger discriminators (informative genes) were used to predict the class of an unknown sample based on the similarity in gene expression (of the informative genes) of the new sample to those of the ALL or AML classes. In other words, the expression level of a gene from the new sample is given a value that predicts ALL or AML class categorization, and the total of all such values from all the informative genes is used to cast the final vote (weighted-voting algorithm).

Results.—The unsupervised and nonhierarchical SOM technique generated either 2 (ALL vs AML) or 3 (AML vs B-cell ALL vs T-cell ALL) distinct expression patterns that highly correlated with morphologically and immunophenotypically distinct acute leukemia categories. The class prediction based on the distinct expression profiles generated from known cases was successfully applied to 34 new leukemia samples, with 29 correctly predicted.

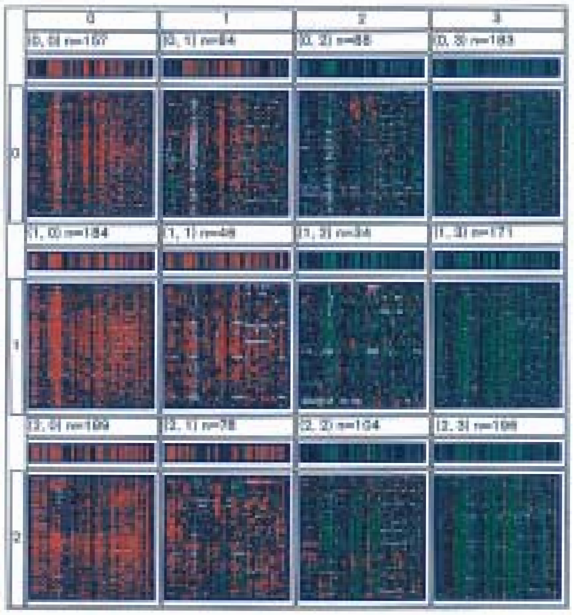

Figure 10. Self-organizing map. Each partition may contain a different number of genes, although the image for each partition is the same size for display purposes. The contents of each partition have been rearranged by clustering. See text for further explanation. Reprinted from Sherlock[39] with permission from Elsevier Science.

Conclusion.—Both class discovery and class prediction of disease categories may be possible with use of gene expression patterns, and such application may complement disease diagnosis and classification. Of note, the investigators used the unsupervised and nonhierarchical (SOM) technique that accurately segregated cases of AML from cases of ALL. They also used a supervised learning classification method based on known classes to develop a classifier (molecular signature) for class prediction. A similar approach was used in a more recent article[50] on lymphoma with results similar to Example 1.

Example 3[51]

Disease.—Breast cancer.

Background.—Breast cancer susceptibility genes 1 and 2 (*BRCA1* and *BRCA2*) are 2 mutations that confer a lifetime risk of breast and ovarian cancer of 50% to 85% and 15% to 45%, respectively.

Purpose.—To test whether gene expression profiles distinguish these 2 types of hereditary breast cancer (*BRCA1* vs *BRCA2*) from each other and from sporadic cases.

Patients and Samples.—Breast cancer tissue from 7 patients with *BRCA1*-related cancer, 7 patients with *BRCA2*-related cancer, and 7 patients with sporadic breast cancer.

Methods.—A cDNA microarray with 6512 cDNA probes representing 5361 genes was used. Total RNA was extracted from frozen tumor tissue. The reference sample was a standard breast cancer cell line. Only genes with an average spot intensity of more than 2500 pixels in any of the samples were included in the analysis (3226 genes). Fluorescence intensity ratios (tumor vs reference sample) were calculated, and a statistical method (modified F test) was used to identify 51 genes (genes of interest) among the 3226 genes analyzed whose expression was significantly different among the disease categories (*BRCA1* vs *BRCA2* vs sporadic cases).

Results.—An agglomerative hierarchical clustering algorithm was then applied to the 51 genes of interest (discriminator genes) to generate gene expression patterns (profiles). The resultant patterns were different for *BRCA1* vs *BRCA2* vs sporadic cases in the study population. Furthermore, accurate class prediction was possible for *BRCA1*-positive vs *BRCA1*-negative tumors.

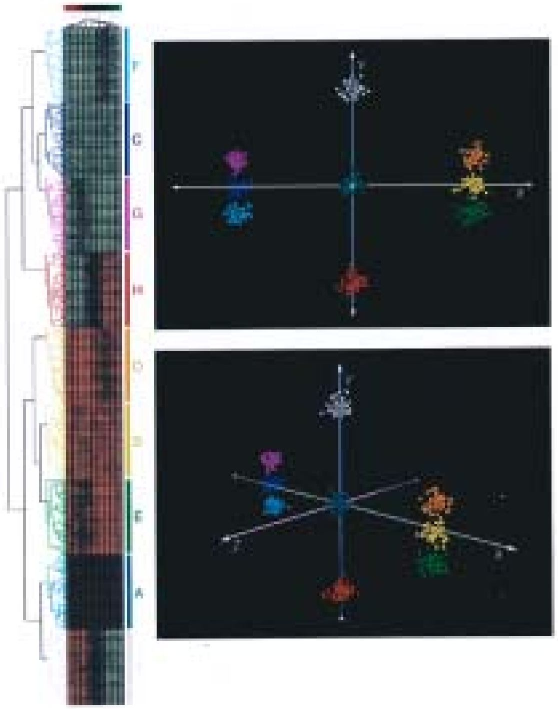

Figure 11. Principle component analysis (PCA). The same demonstration data set was analyzed by using either hierarchical cluster analysis (top) or PCA (bottom). See text for further explanation. Reprinted with permission from Quackenbush.[13]

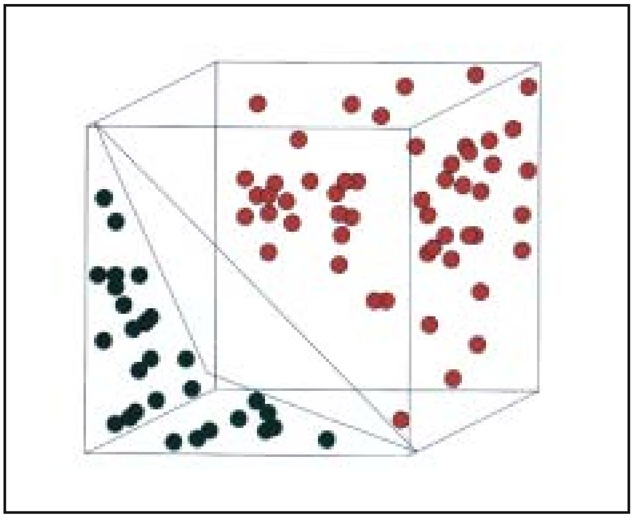

Figure 12. A supervised cluster analysis method. Support vector machine is a computational entity that accepts positive and negative training examples (samples with previously known classes) and uses the knowledge to classify new samples (unknown) into a class membership, shown as red and green dots separated by a hyperplane. Reprinted with permission from Gaasterland and Bekiranov.[46]

Conclusion.—Gene expression profiles can generate molecular signatures that may complement current methods of disease diagnosis and classification. Distinct gene expression profiles suggest different pathways of disease pathogenesis and provide clues to further understanding the cause of disease.

CONCLUSION

The validity of the computational observations in a microarray experiment using current statistical algorithms is often open to methodologic criticisms.[52] Nevertheless, genome-wide assessment of gene structure and function is a science in development that requires more patience and less cynicism. Ultimately, the microarray system needs to be complemented by structural genomics and other innovative platforms to maximize insight into disease pathogenesis and behavior.[53]

REFERENCES

1. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* 1998;16:301-306.
2. Raitio M, Lindroos K, Laukkanen M, et al. Y-chromosomal SNPs in Finno-Ugric-speaking populations analyzed by minisequencing on microarrays. *Genome Res.* 2001;11:471-482.
3. Mei R, Galipeau PC, Prass C, et al. Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.* 2000;10:1126-1137.
4. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270:467-470.
5. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403:503-511.
6. Huang RP. Detection of multiple proteins in an antibody-based protein microarray system. *J Immunol Methods.* 2001;255:1-13.
7. Haab BB. Advances in protein microarray technology for protein expression and interaction profiling. *Curr Opin Drug Discov Devel.* 2001;4:116-123.
8. Borrebaeck CA, Ekstrom S, Hager AC, Nilsson J, Laurell T, Marko-Varga G. Protein chips based on recombinant antibody fragments: a highly sensitive approach as detected by mass spectrometry. *Biotechniques*. 2001;30:1126-1130, 1132.
9. Zhu H, Bilgin M, Bangham R, et al. Global analysis of protein activities using proteome chips. *Science*. 2001;293:2101-2105.
10. Fields S. Proteomics: proteomics in genomeland. *Science*. 2001; 291:1221-1224.
11. Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. *Nature*. 2000;405:827-836.
12. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A.* 1998;95:14863-14868.
13. Quackenbush J. Computational analysis of microarray data. *Nat Rev Genet.* 2001;2:418-427.
14. Planet PJ, DeSalle R, Siddall M, Bael T, Sarkar IN, Stanley SE. Systematic analysis of DNA microarray data: ordering and interpreting patterns of gene expression. *Genome Res.* 2001;11:1149-1155.
15. Bumol TF, Watanabe AM. Genetic information, genomic technologies, and the future of drug discovery. *JAMA*. 2001;285:551-555.
16. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531-537.
17. Yeang CH, Ramaswamy S, Tamayo P, et al. Molecular classification of multiple tumor types. *Bioinformatics*. 2001;17(suppl 1): S316-S322.
18. Bittner M, Meltzer P, Chen Y, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*. 2000;406:536-540.
19. Rickman DS, Bobek MP, Misek DE, et al. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res.* 2001;61:6885-6891.
20. Hughes TR, Marton MJ, Jones AR, et al. Functional discovery via a compendium of expression profiles. *Cell*. 2000;102:109-126.
21. Fuller GN, Rhee CH, Hess KR, et al. Reactivation of insulin-like growth factor binding protein 2 expression in glioblastoma multiforme: a revelation by parallel gene expression profiling. *Cancer Res.* 1999;59:4228-4232.
22. Favis R, Barany F. Mutation detection in K-*ras*, BRCA1, BRCA2, and p53 using PCR/LDR and a universal DNA microarray. *Ann N Y Acad Sci*. 2000;906:39-43.
23. Lomri A, Lemonnier J, Delannoy P, Marie PJ. Increased expression of protein kinase Calpha, interleukin-1alpha, and RhoA guanosine 5′-triphosphatase in osteoblasts expressing the Ser252Trp fibroblast growth factor 2 receptor Apert mutation: identification by analysis of complementary DNA microarray. *J Bone Miner Res.* 2001;16:705-712.
24. Buetow KH, Edmonson M, MacDonald R, et al. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci U S A.* 2001;98:581-584.
25. Dziejman M, Balon E, Boyd D, Fraser CM, Heidelberg JF, Mekalanos JJ. Comparative genomic analysis of Vibrio cholerae: genes that correlate with cholera endemic and pandemic disease. *Proc Natl Acad Sci U S A.* 2002;99:1556-1561.
26. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet.* 2001;29:365-371.

27. Southern E, Mir K, Shchepinov M. Molecular interactions on microarrays. *Nat Genet.* 1999;21(1, suppl):5-9.
28. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. *Nat Genet.* 1999;21(1, suppl):10-14.
29. Whitesides GM, Christopher Love J. The art of building small [published correction appears in *Sci Am*. 2002;286:10]. *Sci Am.* 2001;285:38-47.
30. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat Genet.* 1999;21(1, suppl):20-24.
31. Schulze A, Downward J. Navigating gene expression using microarrays—a technology review. *Nat Cell Biol.* 2001;3:E190-E195.
32. Brockman JA, Tamminga CA. The human genome: microarray expression analysis. *Am J Psychiatry.* 2001;158:1199.
33. Hess KR, Zhang W, Baggerly KA, Stivers DN, Coombes KR. Microarrays: handling the deluge of data and extracting reliable information. *Trends Biotechnol.* 2001;19:463-468.
34. Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A.* 2000;97:9834-9839.
35. Chen H, Liu J, Merrick BA, Waalkes MP. Genetic events associated with arsenic-induced malignant transformation: applications of cDNA microarray technology. *Mol Carcinog.* 2001;30:79-87.
36. Gilbert DR, Schroeder M, van Helden J. Interactive visualization and exploration of relationships between biological objects. *Trends Biotechnol.* 2000;18:487-494.
37. Brazma A, Vilo J. Gene expression data analysis. *Microbes Infect.* 2001;3:823-829.
38. Dopazo J, Zanders E, Dragoni I, Amphlett G, Falciani F. Methods and approaches in the analysis of gene expression data. *J Immunol Methods.* 2001;250:93-112.
39. Sherlock G. Analysis of large-scale gene expression data. *Curr Opin Immunol*. 2000;12:201-205.
40. Alizadeh AA, Ross DT, Perou CM, van de Rijn M. Towards a novel classification of human malignancies based on gene expression patterns. *J Pathol.* 2001;195:41-52.
41. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A.* 1999;96:6745-6750.
42. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A.* 2000;97:12079-12084.
43. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet.* 1999;22:281-285.
44. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A.* 1999;96:2907-2912.
45. Raychaudhuri S, Sutphin PD, Chang JT, Altman RB. Basic microarray analysis: grouping and feature reduction. *Trends Biotechnol.* 2001;19:189-193.
46. Gaasterland T, Bekiranov S. Making the most of microarray data. *Nat Genet.* 2000;24:204-206.
47. Brown MP, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A.* 2000;97:262-267.
48. Hvidsten TR, Komorowski J, Sandvik AK, Laegreid A. Predicting gene function from gene expressions and ontologies. *Pac Symp Biocomput*. 2001;299-310.
49. Califano A, Stolovitzky G, Tu Y. Analysis of gene expression microarrays for phenotype classification. *Proc Int Conf Intell Syst Mol Biol.* 2000;8:75-85.
50. Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med.* 2002;8:68-74.
51. Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer. *N Engl J Med.* 2001;344:539-548.
52. King HC, Sinha AA. Gene expression profile analysis by DNA microarrays: promise and pitfalls. *JAMA.* 2001;286:2280-2288.
53. von Eggeling F, Junker K, Fiedle W, et al. Mass spectrometry meets chip technology: a new proteomic tool in cancer research? *Electrophoresis*. 2001;22:2898-2902.

Primer on Medical Genomics Part IV will appear in the November issue.