Experimental Design for Gene Expression Analysis

Answers Are Easy, Is Asking the Right Question Difficult?

Marcia V. Fournier, Paulo Costa Carvalho, David D. Magee, Maria da Gloria Costa Carvalho, and Krishnarao Appasani

Summary

More and more, array platforms are being used to assess gene expression in a wide range of biological and clinical models. Technologies using arrays have proven to be reliable and affordable for most of the scientific community worldwide. By typing microarrays or proteomics into a search engine such as PubMed, thousands of references can be viewed. Nevertheless, almost everyone in life science research has a story to tell about array experiments that were expensive, did not generate reproducible data, or generated meaningless data. Because considerable resources are required for any experiment using arrays, it is desirable to evaluate the best method and the best design to ask a certain question. Multiple levels of technical problems, such as sample preparation, array spotting, signal acquisition, dye intensity bias, normalization, or sample-contamination, can generate inconsistent results or misleading conclusions. Technical recommendations that offer alternatives and solutions for the most common problems have been discussed extensively in previous work. Less often discussed is the experimental design. A poor design can make array data analysis difficult, even if there are no technical problems. This chapter focuses on experimental design choices in terms of controls such as replicates and comparisons for microarray and proteomics. It also covers data validation and provides examples of studies using diverse experimental designs. The overall emphasis is on design efficiency. Though perhaps obvious, we also emphasize that design choices should be made so that biological questions are answered by clear data analysis.

Key Words: experimental design; gene expression; microarray, proteomics.

1. cDNA and DNA Microarrays

1.1. Background

How are complex organisms such as humans formed from a single cell? How are tissues differentiated? How do cells function in different environments? What changes occur in diseases? Such questions can be assessed using good biological models in combination with a comprehensive assessment of gene expression patterns. Whether you want to view the entire genome on a single array or focus on a target set of biologically relevant genes, microarrays allow the analysis of gene expression levels and can be applied in a broad spectrum of questions, including uncovering new regulatory pathways, validating drug targets, clarifying diseases, analyzing toxicological responses, or building robust databases (1–9). Although the full potential of arrays is yet to be realized, these tools have shown great promise in deciphering complex diseases such
as cancer (10–13). Array technologies have varying limitations, which need to be kept in mind when choosing among them. This chapter covers two types of microarray platforms—cDNA and oligonucleotides—that are currently used, and both are effective for assessing gene expression patterns (14,15).

Oligonucleotide microarrays use direct synthesis or deposition of oligonucleotides onto a solid surface and single color readout of gene expression from a test sample. Oligonucleotides offer greater specificity than cDNAs, because they can be tailored to minimize chances of cross-hybridization. Major advantages of these arrays include the uniformity of probe length and the ability to distinguish among splice variants. Another advantage particular to the commonly used Affymetrix GeneChip system (Affymetrix, Santa Clara, CA) (16) is the ability to perform multiple independent measurements of each transcript of interest, providing reliable assessments of each data point. In addition, this system allows the recovery of samples after hybridization to a chip and its sequential hybridization to multiple arrays, a considerable advantage when dealing with limited resources.

cDNA microarrays are typically limited by density and can analyze each transcript with only a single probe, compromising the robustness of the array. The primary benefit of cDNA arrays is that they can be made by individual investigators, are easily customized, and do not require a priori knowledge of a cDNA sequence, because clones can be used and then sequenced later if they are of interest.

1.2. One-Color Vs Two-Color Arrays

In one-color arrays, the experimental RNA sample is amplified enzymatically, biotin-labeled for detection, hybridized to the microarray, and detected through the binding of a fluorescent compound (streptavidin-phycoerythrin). Two-color arrays use the competitive hybridization of two messenger RNA (mRNA) samples labeled with dyes—cyanine (Cy)3 and Cy5—to measure the relative gene expression levels of the samples. Quantified signal intensities for Cy3 (R) and Cy5 (G) are intended to be proportionally consistent with the mRNA levels for the two samples across all spotted genes and slides. Inconsistencies in channel intensity result from various steps of microarray fabrication, RNA preparation, hybridization, scanning, or image processing. The ability to make a direct comparison between two RNA samples on the same microarray slide is a unique and powerful feature of the two-color arrays.

1.3. Experimental Control and Bias Issues

Experimental controls provide the standard means to implement quality assurance procedures in biological experiments. In addition, the experimental design should incorporate controls to minimize sources of bias in an experiment. In statistical terms, bias is defined as the difference in value between a sample and a population measurement. More generally, bias is any partiality that prevents the objective consideration of an issue or situation. Experimental controls provide a means to measure and normalize biases associated with sample conditions and technical aberrations. To eliminate unwanted bias, it is necessary to include controls for experimental conditions in the analysis (17). Microarray experiments typically involve small numbers of replicates where the assumptions of classical statistics (normally distributed values) do not apply. Typically, researchers calculate \( p \) values without testing the normality of the data. A more
1.3.1. Nonbiological Bias and Data Normalization:

In microarray experiments, systematic variation from a variety of nonbiological sources can affect measured gene expression levels. The process of normalization seeks to eliminate such variation and thus enhance the reliability of results obtained from subsequent higher-order statistical analysis of the data (18,19). Important low-level sources of technical variation are fluorescent intensities between two channels and the physical locations of spots on a microarray slide. Technical bias is introduced during array printing, extraction, labeling, and hybridization. Sophisticated normalization methods adjust such spatial and intensity bias. The diagnosis of bias and the assessment of normalization methods are currently accomplished using various plots (20).

Using plots to assess normalization, however, does not specifically address how to order methods to best remove inconsistent bias patterns, how to compare methods statistically, or how to verify the quality of single arrays. In addition, postdata collection error may occur when bias is introduced by erroneous data normalization (21). Therefore, it is desirable to avoid nonbiological bias by implementing a rigorous quality control during all steps of the experiments, including sample preparation, probe amplification, array spotting, hybridization, and signal detection. Thus, if using custom- or home-made arrays, it is important to perform a range of optimization experiments to ensure quality control before performing actual experiments. Ensuring quality control will allow the use of simple normalization methods such as normalization of log values to the median of each array.

1.3.2. Biological Variation

Variability is intrinsic to all organisms and is influenced by genetic or environmental factors. Thus, measurements taken from a particular cell culture are biased by that culture and may not represent a broader gene expression. Pooling samples will conceal biological variation in that expression levels for a gene may vary in each sample. But, once pooled, all the replicates should indicate the same level. In this case, you can assume that any remaining variation is nonbiological. Data from pooled RNA replicates (technical replicates) would be useful for assessing the quality of the original arrays and the hybridization conditions. From a biological perspective, pooling RNA eliminates an important experimental control as well as potentially useful gene expression information. For considering sample controls, it is better to use several pools and fewer replications than one pool of all the available samples and multiple replications (22).

The main goal of replication is to generate independent measurements for the purpose of reducing a type of bias. For most questions, biological bias and technical bias should be considered before constructing statistical tests. Hence, good designs should incorporate replication at these levels (22–25). For example, performing dye swaps in two-color array experiments provides technical replicates and minimizes technical bias caused by any difference in dye intensity.

Identifying the independent measurements in an experiment is a prerequisite for a proper statistical analysis. Details on how individual animals, cultures, or samples were handled through the course of an experiment can be important for identifying which
biological samples and technical replicates are independent. Although it is tempting to avoid biological replicates in an experiment because the results seem to be more reproducible (23), it is useful to use them because they ensure that the results are biologically significant. In contrast, by eliminating biological replicates (e.g., pooling RNA), you can be somewhat certain that the remaining variation is technical (nonbiological). Replicate spots on the same slide provide an effective means to measure technical variation, which could represent array printing quality, hybridization conditions, or spot-reading software. When initially calibrating laboratory conditions, you may focus on reducing technical error by using pooled RNA. After the conditions are optimized, the focus of measure may turn to controlling the biological variance.

1.3.3. Reference Samples

The ability to make a direct comparison between two samples on the same microarray slide is a unique and powerful feature of the two-color microarray. Thus, the main consideration for cDNA arrays is which samples should be cohybridized (22). For example, if an investigator has unlimited amounts of reference material, but a limited amount of test RNA, then a comparison of the test sample with a reference sample should be the choice. Reference or control samples can be biological controls (untreated cell or animal) or universal reference RNA.

A useful tool for monitoring and controlling intra- and interexperimental variation in experiments applying a two-color microarray is universal reference RNA (URR) by providing a hybridization signal at each microarray probe location (spot). In this case, all the experimental RNAs, including experimental controls, are hybridized to a URR rather than to each other. Measuring the signal at each spot as the ratio of experimental RNA to reference RNA targets, rather than relying on absolute signal intensity, decreases the variability by normalizing the signal output in any two-color hybridization experiment (26). URR is prepared from pools of RNA derived from individual cell lines representing different tissues. Moreover, experiments using URR can be compared over time and across operators, because the reference RNA is the same (27).

1.4. Types of Design

Experiments should be designed to maximize the efficiency and reliability of the data obtained (28). Careful attention to the experimental design will ensure that the use of available resources is efficient, obvious biases are avoided, and the primary question is answered (22,29). For example, primary questions may include identifying differentially expressed genes, defining groups of genes with similar patterns of gene expression, and identifying tumor subclasses.

Consideration for a good experimental design should take into account the aims of the study, the choices of the biological model or clinical setting, sources of variability, replicates, and the optimal sample size (21,29–31). Biological resources and cost considerations will usually dictate the amount of RNA available and the number of replicates to be used respectively.

Sources of RNA can be either tissue samples or cell lines. How much RNA is available will affect the number of times the experiment can be repeated and its validation. Sample isolation, RNA extraction, and labeling also affect the number of replicates required. Data validation is discussed in greater detail (see Subheading 1.5.).
Experimental Design for Gene Expression Analysis

A single-color microarray compares samples in parallel manner, where each sample is probed in an independent array (Fig. 1A). Comparisons can be either direct or indirect when using a two-color microarray (22). A direct comparison occurs when two samples are compared in the same array (Fig. 1B). In an indirect comparison, the expression levels of the experimental samples are measured separately on different slides by using a reference sample (Fig. 1C). The type of experiment usually dictates the type of design, although there are experiments where diverse types of designs are suitable. Principles are needed for choosing one design from among several possible designs. A focus on the primary question is necessary in such experiments. You should think about which design is suitable to answer the most interesting question. One approach is to ask which comparisons are of greater and which are of lesser interest and then seek a design that gives greater precision to the former and less precision to the latter.

In general, if an experiment needs a small sample size, using two-color arrays and a direct comparison can be advantageous because of the lower statistical variance (22,32). A parallel comparison should be the option for a large number of samples, because of the higher specificity and reproducibility between replicates. Moreover, parallel comparisons facilitate data analysis by using unsophisticated statistical tools, which provide a considerable advantage for a large data set.

How do you make design choices based on the type of experiment? Broadly, experiments can be classified into one of the following categories: treatment regimen, disease classification, time courses, and factorial studies.
1.4.1. Treatment Regimen Studies

The use of an untreated control is obvious in experiments involving treatment regimens. In two-color microarrays, all treated samples should be hybridized with the untreated control. There are two fundamentally distinct designs in human studies to assess treatment effects (32). In one design, the investigator assigns the exposure and then measures the outcome (experimental). In the other design, the exposure is measured (observational). Most retrospective studies fall under the latter category. Case-control design is a good option in studies of gene expression patterns where the treatment and collection of tissue samples are not designed in the laboratory. Cases are patients who died or had recurrence (high aggressive), and controls are those who lived beyond a time line after diagnosis and treatment (low aggressive). For example, in a recent study, 60 patients were selected from a total of 103 ER-positive, early stage cases presented to Massachusetts General Hospital between 1987 and 1997, from whom tumor specimens were snap-frozen and for whom minimal 5-yr follow-up was available. In this study, breast cancer patients treated with tamoxifen were compared for differential gene expression between tamoxifen responders and nonresponders (33). Additional examples on using microarrays for the development and assessment of therapeutics are covered in previous reviews (4,34–37).

1.4.2. Disease Classification Studies

Microarray studies allow class comparison, class discovery, or class prediction (10). In class comparison, the study aims to establish whether gene expression profiles differ between classes. In class discovery, the goal is to elucidate unrecognized subclasses—such as new tumor subclasses—based on gene expression profiles. In class prediction studies, information from gene expression profiles is used to predict a phenotype (10,27). Combinations between class discovery and class prediction have been used in several clinical studies to assess gene expression patterns identifying tumor subclasses (38–44).

Human tumor samples are heterogeneous mixtures of diverse cell types, including malignant cells, stromal elements, blood vessels, and inflammatory cells. Because of this heterogeneity, the interpretation of gene expression studies is not always simple. Groups attempting to focus on differences between malignant and nonmalignant components of a tumor may use laser capture microdissection of individual cells from a tumor section to isolate cancer cell RNA for microarray experiments. Sgroi et al. (45) combined laser capture microdissection and cDNA microarray analysis to compare global gene expression in normal and tumor cells from the same tissue specimen. In other study, cell type-specific surface markers and magnetic beads have been used to isolate various cell types in a tissue sample for gene expression studies (46,47). Designs using magnetic beads for cell type isolation allowed analysis of tumor cells and the surrounding microenvironment.

1.4.3. Time-Course Studies

A time-course experiment is a case of a multiple-slide experiment in which transcript abundance is monitored over time. Recently, several methods have been suggested to identify differentially expressed genes in multiple-slide microarray experiments based on statistical models such as the analysis of variance model and the mixed effect model (48,49). In time-course experiments, the comparisons to be made might not be obvious. The comparisons of greatest interest should determine the best design (Table 1). For example, in a time-course experiment where the main question is about gene expression changes occurring in relation to the starting point, the use of a starting point (time zero) as a reference is an appropriate choice. Assessing the activity of indi-
individual genes as a function of time is an example of time-course experiments. The oncogenic potential of cyclin D1 overexpression in a time dependent manner has been dissected by using gene expression profiling as a global assay (2). Comparing temporal gene expression patterns, following the expression of wild-type and mutant cyclin D1 proteins, revealed an expression profile of cyclin D1 target genes. In this case, the gene expression was analyzed from time 0 to 24 h postinfection with adenovirus. The genes were selected by expression levels in comparison with time 0 and 4 h and showed

### Table 1

| Time-Course Experiments |
|--------------------------|
| **One-color array using oligonucleotide platform** |
| - Parallel comparison |
| Sample | Array unit | Type of replicate | Strongness |
| T1 | A1 | | One or more samples can be used as a reference |
| T2 | A2 | | Easy data analysis |
| T3 | A3 | | Flexibility |
| T1’ | A4 | Biological replicate | Highly reproducible |
| T2’ | A5 | Biological replicate | Multiple representation of genes in the arrays |
| T3’ | A6 | Biological replicate | |
| **Two-color array** |
| - Design I: using direct comparison of time points (Cy3-Cy5) |
| Sample | Array unit | Type of replicate | Strongness |
| T1-T2 | A1 | | Comparison of time points in the same array |
| T1-T3 | A2 | | |
| T2-T3 | A3 | | |
| T2-T1 (dye swap) | A4 | Technical replicate | |
| T3-T1 (dye swap) | A5 | Technical replicate | |
| T3-T2 (dye swap) | A6 | Technical replicate | |
| - Design II: indirect comparison using a reference sample (Cy3-Cy5) |
| Sample | Array unit | Type of replicate | Strongness |
| T1-R | A1 | | Apply universal reference sample |
| T2-R | A2 | | Easy analysis |
| T3-R | A3 | | |
| Plus dye swaps | A4-A6 | Technical replicate | |
| - Design III: linear comparison |
| Sample | Array unit | Type of replicate | Strongness |
| T1-T2 | A1 | | Uses less microarray slides |
| T2-T3 | A2 | | |
| Plus dye swaps | A3-A4 | Technical replicate | |
| - Design IV: using T1 as reference |
| Sample | Array unit | Type of replicate | Strongness |
| T1-T2 | A1 | | Uses less microarray slides |
| T1-T3 | A2 | | Simple analysis |
| Plus dye swaps | A3-A4 | Technical replicate | |

In an experiment including three time points (T1, T2, and T3) at least five possible experimental designs can be considered. The design choice should be made based on the time-point comparisons most important for answering the question. T1, time zero; T2 and T3 experimental time points.
maximum expression at 24 h. This empirically determined expression profile was used to further study the mechanistic basis of the oncogenic consequences of cyclin D1 overexpression in human tissues.

1.4.4. Factorial Studies

Gene expression profiling can capture the activity of individual genes by using overexpression or depletion systems. For example, Hughes et al. (50) investigated in yeast the functions of previously uncharacterized genes by matching gene expression changes induced by their deletion against a collection of yeast reference profiles. Multifactorial designs consider differences that not only are caused by single factors but also result from the interaction of two or more factors. In such experiments, there are three major questions to be addressed: 1) How do individual factors affect cells (unique pathways affected), 2) What are the common pathways affected, and 3) What are the combinatorial effects (Table 2). One possible design would be the comparison of single treated or combination treated with a reference biological control (untreated). Parallel designs have the advantage of allowing comparisons between any of the experimental samples, easy analysis, and interpretation without the need for statistical tools. Direct designs can generate results where a lower variance of estimated effects is obtained compared with indirect designs (22). A balance between direct and indirect designs can be used as well. In this case, the main comparisons should be direct, whereas the secondary comparisons can be indirect. Examples of multifactorial studies include treatment combination (51–54) and double knockouts (55, 56).

1.5. Verification

The amount of verification can influence the choice of statistical method and sample size. The main goal for verifying the array data is to identify bona fide gene expression changes. Verification can be broad or confined to the original samples used to acquire the data. When array verification is confined to the samples used for the generation of the data, it is usually focused on one or a small group of genes. A group of genes can be assessed using reverse transcription-PCR, Northern blots, Westerns blots, or tissue immunostaining, providing a validation of patterns of gene expression. Verification also can be extended to an independent data set to validate the ability of selected sets of genes to make a certain distinction, such as a treatment response or a clinical outcome. This strategy has been used to unveil gene expression profiles that predict clinical outcomes of breast cancer (43). In this study, a group of 70 genes was identified as a molecular signature for a poor prognosis in early stage breast cancer patients. This molecular signature was validated in a second study by using an independent data set (57). Confirmation using an independent data set enhances confidence in the obtained results.

Data verification and validation also can use gene expression manipulation (e.g., expression systems, knockout, and small-interfering RNA). Studies that aim to discover relevant gene expression associated with a phenotype or to elucidate a gene function should use this approach.

An experimentally generated data set also can be compared with published data available online in public databases. One of the advantages of validation using gene expression databases is the fast assessment of data. However, the investigator should keep in mind that the variability in the quality and reproducibility between array platforms is a limiting factor. The choice of the verification method will depend on the biological resources and budget available. A combination of verification levels is usually a good choice.
Proteins are the functional units of cells and represent the end products of gene expression. Proteomics is a set of methods and tools for studying a cell’s proteome, the proteins expressed by a cell’s genome at a given time. This approach permits the study of posttranslational modifications and quantifies levels of protein expression over time in response to changes in environment or disease. In proteomics methodology, proteins are identified mostly by the use of two-dimensional electrophoresis (2DE) coupled with mass spectrometry (MS) and computer search algorithms. These combined tools are able to identify a purified, digested protein by comparing the mass spectra of the peptide with in silico digestion of proteins contained in a database (e.g., Swiss–Prot).

### Table 2

**Multifactorial Experiments**

#### One-color array by using oligonucleotide platform

- Parallel comparison

| Sample | Array unit | Type of replicate | Strongness |
|--------|------------|-------------------|------------|
| R      | A1         |                   | One or more samples can be used as a reference |
| A      | A2         |                   | Easy data analysis |
| B      | A3         |                   | Flexibility |
| AB     | A4         |                   | Highly reproducible |
| R'     | A5         | Biological replicate | Multiple representation of genes in the arrays |
| A'     | A6         | Biological replicate | |
| B'     | A7         | Biological replicate | |
| AB'    | A8         | Biological replicate | |

#### Two-color array

- Design I: indirect comparison using a reference sample

| Sample | Array unit | Type of replicate | Strongness |
|--------|------------|-------------------|------------|
| A-R    | A1         |                   | Uses less microarray slides |
| B-R    | A2         |                   | |
| AB-R   | A3         |                   | |
| dye swaps | A4       | Technical replicate | |
| dye swaps | A5       | Technical replicate | |
| dye swaps | A6       | Technical replicate | |

- Design II: loop

| Sample | Array unit | Type of replicate | Strongness |
|--------|------------|-------------------|------------|
| R-A    | A1         |                   | Mix direct and indirect comparisons |
| A-B    | A2         |                   | |
| B-AB   | A3         |                   | |
| AB-R   | A4         |                   | |
| dye swaps | A5       | Technical replicate | |
| dye swaps | A6       | Technical replicate | |
| dye swaps | A7       | Technical replicate | |
| dye swaps | A8       | Technical replicate | |

In multifactorial experiments there are 3 major questions to be addressed: how individual factors affect cells (the unique pathways affected), the common pathways affected, and combinatorial effects.

R, not treated; A, treatment 1; B, treatment 2; AB, combinatorial treatment.

### 2. Proteomics

#### 2.1. Background

Proteins are the functional units of cells and represent the end products of gene expression. Proteomics is a set of methods and tools for studying a cell’s proteome, the proteins expressed by a cell’s genome at a given time. This approach permits the study of posttranslational modifications and quantifies levels of protein expression over time in response to changes in environment or disease. In proteomics methodology, proteins are identified mostly by the use of two-dimensional electrophoresis (2DE) coupled with mass spectrometry (MS) and computer search algorithms. These combined tools are able to identify a purified, digested protein by comparing the mass spectra of the peptide with in silico digestion of proteins contained in a database (e.g., Swiss–Prot).
New advances through proteomics include cancer biomarkers, protein signatures, interactions between protein networks, and their relation to multicellular functions. A specific biomolecule in the body that can be used for detecting a disease and measuring its progress or response to a treatment constitutes a biomarker. Even though many advances have been made in mass spectrometry-based proteomics, it still lags behind microarray technology when dealing with experimental design and reproducibility. Integrating proteomic data in global databases and standardizing protocols is still a subject of great discussion, because results vary with different mass spectrometers.

2.2. Two-Dimensional Electrophoresis

2DE is a common and efficient technique for isolating proteins for subsequent MS identification. Numerous studies involving 2DE and MS have already identified various disease-related changes in the levels of protein expression. 2DE gels are prepared with the biological samples of healthy and sick subjects, so the identification of differentially expressed proteins can be done visually or by computer gel comparison algorithms. This powerful method is capable of separating several thousand proteins per experiment, according to two independent physicochemical properties of the protein: isoelectric point in the first dimension and molecular weight in the second dimension (58). Reproducibility is of the highest importance so that variations in the level of protein expression between samples can be studied. The use of immobilized pH gradients for the first dimension (59), as an alternative to carrier ampholytes, has made 2DE reproducible enough for proteome analysis. Although it is common to separate proteins through the linear pH range of 3.5–10 (isoelectric focusing), there are commercially available immobilized pH gradients with many different ranges so that the separation of proteins can be optimized to a desired range. Note that highly hydrophobic proteins are hard to keep in solution; thus, they can be lost during sample preparation and isoelectric focusing.

The molecular weight separation can range from 6 to 300 kDa, and gel staining can be done through various techniques, with silver (60) and Coomassie blue being the most popular. To find a protein of interest, different gels are created (i.e., healthy and patient subjects). Once a differentially expressed spot is found, it can be further analyzed. MS should then be used to identify the protein and its possible function.

2.3. Protein Identification Through Matrix-Assisted Laser Desorption Ionization-Time of Flight (MALDI-TOF)

MALDI-TOF MS is the method of choice for analyzing peptide mixtures because of its fast data acquisition (Fig. 2). After the spectra are obtained from the mixture, program filters such as smooth and background cutoffs can be applied to the spectra data. Other options, such as the selection of spectra mono-isotopic peaks, also should be used when exporting the data to Web interfaces for protein identification. Among such programs, Protein Prospector (61) and Mascot (www.matrixscience.com) have become popular, because they are freely available through the World Wide Web. After receiving a peptide mass list through a Web form, these programs perform an in silico digestion of the proteins available in their database and calculate the mass of each theoretical peptide. Possible identifications are presented as a report. If an ambiguous identification occurs, tandem mass spectrometry can be done to obtain the amino acid sequence or the tag of a small peptide. By inputting the sequence tag, together with the mass spectra data, the protein should be properly identified if it is contained in the database.
2.4. Differential Spectra Analysis for Cancer Biomarker Characterization

Generally, cancer is diagnosed and treated too late, after cancer cells have already invaded and formed metastasis (62). Serum tests based on single biomarkers, such as the prostate-specific antigen for the prostate, can sometimes give misleading results (63). Protein expression analysis in body fluids by using MS offers great hope for early cancer diagnosis screening tests by characterizing pathological protein patterns. These protein expression signatures can also be used in therapy response predictions, allowing appropriate medication selection (63). The quest for new biomarkers, so protein arrays can be used instead of a single target, is the hope for improving specificity and sensitivity in early cancer detection.

One experimental approach for early diagnosis by using differential protein patterns is based on the construction of antibody chips that assay the levels of several biomarkers simultaneously (64). In this case, the diagnosis is based on a “larger picture,” rather than on the expression of a single protein that could mislead a test. This methodology is straightforward; and the chip is simply a set of antigen tests that run in parallel.

Another experimental design is based on differential mass spectra analysis and consists of monitoring specific spectrum peaks, where peak intensity variation could suggest a pathology diagnosis (65). Such an approach could be limited, because it does not detect variations in other regions of the mass spectrum. A multidisciplinary approach...
to performing pathology classification based on protein profiling is attained by using artificial intelligence to scan across the entire spectrum. Machine-learning techniques—such as neural networks, and recently, supporting vector machines—can correctly classify an unknown spectrum among known groups.

Tumor biomarkers are usually secreted in extremely low quantities into human blood. To better identify the differentially expressed proteins, experimental approaches are targeted at collecting and studying directly what is being secreted by tumor tissues. This approach provides key insight into “what to look for” when trying to spot diluted tumor biomarkers in the serum. Detection in a serum is important because most medical diagnosis strategies use easily accessed body fluids.

2.5. Precautions During Experimental Design

Experimental design is a key issue when dealing with proteomics. The critical issues are eliminating bias, providing a careful analysis of the data without overfitting, and recognizing its limitations. Proteomic experiments deal with many experimental design challenges, such as limited and variable sample amounts, sample degradation, posttranslational modifications, and disease or drug alterations. When performing differential spectra analysis or protein identification, a few key points should be considered for experimental design:

1. It has been shown that both sensitivity and specificity decline significantly when samples are processed in the same laboratory after a delay of several months. Always keep samples at –80°C.
2. When using liquid chromatography coupled with an electron spray, randomize the sample order between control subject and patients, and, if possible, use several sample pools.
3. Always calibrate the mass spectrometer before an experiment. In MALDI-TOF, the use of commercially available controls that contain known masses can be mixed with the analyte and used as internal controls, or they can be spotted next to the analyte for MS/MALDI calibration.
4. In differential spectra analysis, biomarkers are expressed in very low amounts, so whether or not to deplete high abundant proteins (i.e., Albumin) is a key issue in experimental design. If such proteins are depleted, detection sensitivity for other proteins could be improved. However, many other proteins that stick to albumin also are removed, resulting in the loss of information and possible biomarkers.
5. The removal of contaminant peaks is an important step in protein identification. Trypsin autolysis and human keratin are among the most common contaminants having well-known and characterized peaks. In the majority of spot identification cases, the removal of contaminant peaks from the mass spectrum will result in better spot identification with higher scores. It is important to note that, although rare, the protein of interest may have masses identical to one or more of these peaks, and its removal may reduce the match score.
6. In differential spectra analysis, choose the best method of data normalization (e.g., total ion current, maximum peak, intensity, etc.).
7. Optimize the MALDI-TOF laser frequency and intensity for a given experiment. Lower frequencies allow the detector to better recover; high laser intensity can result in a “noisy” spectrum.
8. Validate the experiment by using antibodies. Keep in mind that differentially expressed protein peaks could be related to well-known acute phase proteins and thus might not represent putative biomarkers.
9. When performing protein identification by using Web programs, always consider possible posttranslational modifications and tryptic missed cleavages.
3. Conclusions

As informative as DNA microarray expression studies are, it has been shown that changes in mRNA expression often correlate poorly with changes in protein expression. However, both technologies give insight into pathology prediction and treatment (63). The future holds great challenges in proteomics and transcriptomics, especially when integrating large and disparate data types. Proteomics promises a major role in creating a predictive, preventative, and personalized approach to medicine. Great efforts are being made to standardize protocols so that a global proteome database can be constructed. This standardization is not currently possible, because spectra reproducibility, especially of complex mixtures among different mass spectrometers, is still a challenge. Although great advances in MS methods and bioinformatics have been made recently, much more effort is necessary to decipher the secrets encrypted within the large volumes of data generated by modern proteomic techniques, to unravel the true power of proteomics, and to find the hidden links with transcriptomics.

Microarray applications towards the diagnostics or the pathogen-derived DNA/proteome chips will have tremendous clinical value and they are not far away from the clinic. As of today, the lab-made Oligo or cDNA arrays are currently useful in research labs either for the classification of cancers or for infectious disease diagnosis, but their real applications did not reach the clinic. We hope that it is not too far from reality since US Food and Administration has approved Cytochrome P450 Affymetrix gene chip for the xenobiotics (drug resistance in liver disease) studies, which are commercially available through Roche. On the other hand, protein chips are available from Invitrogen (Carlsbad, CA) for the study of protein–protein interactions; however their clinical applications also have not reached the pathology laboratories yet. Several groups, including Michael Snyder’s (Yale), Joshua LaBaer’s (Harvard), DeResi’s (UCSF), Liotta and Petricoin’s (George Mason University,) and companies such as Affymetrix, Illumina, JPT Peptide Technologies, GE Healthcare, Pierce Biotechnology, CombiMatrix, Agilent and others, are developing both DNA, protein, and glycoarrays for the diagnosis of infectious diseases (like coronavirus, Francisella tularensis, Vibrio cholerae). In recent years, contributions from the laboratory of Patrick Brown’s group at Stanford are well appreciated in the field of microarrays, since he has looked at the global analysis of gene expression patterns in normal human tissues/cell lines, including several human disease tissues such as various forms of cancers, fibroblasts, peripheral blood, placenta and various compartments of eye, and developed a comprehensive “global atlas of gene expression patterns” (68), which will definitely be useful in the clinic in the coming decade.

Acknowledgments

We thank James Garbe for discussions and Robert Cowles for critical reading of this chapter. We apologize for inadvertently omitting the work of many investigators who could not be cited because of space limitations.

References

1. Debouck, C., Goodfellow, P. N. (1999) DNA microarrays in drug discovery and development. Nat. Genet. 21, 48–50.
2. Lamb, J., Ramaswamy, S., Ford, H. L., et al. (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. Cell 114, 323–334.
3. Bild, A. H., Yao, G., Chang, J. T., et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357.
4. Bild, A. H., Potti, A., Nevins, J. R. (2006) Linking oncogenic pathways with therapeutic opportunities. *Nat. Rev. Cancer* **6**, 735–741.
5. Bejjani, B. A., Shaffer, L. G. (2006) Application of array-based comparative genomic hybridization to clinical diagnostics. *J. Mol. Diagn.* **8**, 528–533.
6. Jayapal, M., Melendez, A. J. (2006) DNA microarray technology for target identification and validation. *Clin. Exp. Pharmacol. Physiol.* **33**, 496–503.
7. Wulfkuhle, J. D., Edmiston, K. H., Liotta, L. A., Petricoin, E. F., 3rd (2006) Technology insight: pharmacoproteomics for cancer–promises of patient-tailored medicine using protein microarrays. *Nat. Clin. Pract. Oncol.* **3**, 256–268.
8. Liu, E. T., Kuznetsov, V. A., Miller, L. D. (2006) In the pursuit of complexity: systems medicine in cancer biology. *Cancer Cell* **9**, 245–247.
9. Fournier, M. V., Martin, K. J. (2006) Transcriptome profiling in clinical breast cancer: From 3D culture models to prognostic signatures. *J. Cell. Physiol.* **209**, 625–630.
10. Ramaswamy, S., Golub, T. R. (2002) DNA microarrays in clinical oncology. *J. Clin. Oncol.* **20**, 1932–1941.
11. Quackenbush, J. (2006) Microarray analysis and tumor classification. *N. Engl. J. Med.* **354**, 2463–2472.
12. Hayes, D. N., Monti, S., Parmigiani, G., et al. (2006) Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J. Clin. Oncol.* **24**, 5079–5090.
13. Fan, C., Oh, D. S., Wessels, L., et al. (2006) Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.* **355**, 560–569.
14. Bowtell, D. D. (1999) Options available–from start to finish–for obtaining expression data by microarray. *Nat. Genet.* **21**, 25–32.
15. Holloway, A. J., van Laar, R. K., Tothill, R. W., Bowtell, D. D. (2002) Options available–from start to finish–for obtaining data from DNA microarrays II. *Nat. Genet.* **32 Suppl**, 496–501.
16. Dalma-Weiszhausz, D. D., Warrington, J., Tanimoto, E. Y., Miyada, C. G. (2006) The affymetrix GeneChip platform: an overview. *Methods Enzymol.* **410**, 3–28.
17. Novak, J. P., Sladek, R., Hudson, T. J. (2002) Characterization of variability in large-scale gene expression data: implications for study design. *Genomics* **79**, 104–113.
18. Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.* **32 Suppl**, 496–501.
19. Yang, Y. H., Dudoit, S., Luu, P., et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15.
20. Park, T., Yi, S. G., Kang, S. H., et al. (2003) Evaluation of normalization methods for microarray data. *BMC Bioinformatics* **4**, 33.
21. Morrison, D. A., Ellis, J. T. (2003) The design and analysis of microarray experiments: applications in parasitology. *DNA Cell. Biol.* **22**, 357–394.
22. Yang, Y. H., Speed, T. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* **3**, 579–588.
23. Churchill, G. A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat Genet* **32 Suppl**, 490–495.
24. Pan, W., Lin, J., Le, C. T. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol.* **3**, research0022.
25. Lee, M. L., Kuo, F. C., Whitmore, G. A., Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* **97**, 9834–9839.
26. Novoradovskaya, N., Whitfield, M. L., Basehore, L. S., et al. (2004) Universal Reference RNA as a standard for microarray experiments. *BMC Genomics* 5, 20.
27. Miller, L. D., Long, P. M., Wong, L., et al. (2002) Optimal gene expression analysis by microarrays. *Cancer Cell* 2, 353–361.
28. Gaasterland, T., Bekiranov, S. (2000) Making the most of microarray data. *Nat. Genet.* 24, 204–206.
29. Kerr, M. K., Churchill, G. A. (2001) Experimental design for gene expression microarrays. *Biostatistics* 2, 183–201.
30. Kerr, M. K., Churchill, G. A. (2001) Statistical design and the analysis of gene expression microarray data. *Genet. Res.* 77, 123–128.
31. Simon, R., Radmacher, M. D., Dobbin, K. (2002) Design of studies using DNA microarrays. *Genet. Epidemiol.* 23, 21–36.
32. Yang, P., Sun, Z., Aubry, M. C., et al. (2004) Study design considerations in clinical outcome research of lung cancer using microarray analysis. *Lung Cancer* 46, 215–226.
33. Ma, X. J., Wang, Z., Ryan, P. D., et al. (2004) A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 5, 607–616.
34. Clarke, P. A., te Poele, R., Workman, P. (2004) Gene expression microarray technologies in the development of new therapeutic agents. *Eur. J. Cancer* 40, 2560–2591.
35. Nelson, P. S. (2004) Predicting prostate cancer behavior using transcript profiles. *J. Urol.* 172, S28–32; discussion S33.
36. Mischel, P. S., Cloughesy, T. F., Nelson, S. F. (2004) DNA-microarray analysis of brain cancer: molecular classification for therapy. *Nat. Rev. Neurosci.* 5, 782–792.
37. Lee, C. H., Macgregor, P. F. (2004) Using microarrays to predict resistance to chemotherapy in cancer patients. *Pharmacogenomics* 5, 611–625.
38. Perou, C. M., Sorlie, T., Eisen, M. B., et al. (2000) Molecular portraits of human breast tumours. *Nature* 406, 747–752.
39. Ramaswamy, S., Tamayo, P., Rifkin, R., et al. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* 98, 15,149–15,154.
40. Sorlie, T., Perou, C. M., Tibshirani, R., et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* 98, 10,869–10,874.
41. Wang, Y., Klijn, J. G., Zhang, Y., et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365, 671–679.
42. Dai, H., van ’t Veer, L., Lamb, J., et al. (2005) A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer Res.* 65, 4059–4066.
43. van ’t Veer, L. J., Dai, H., van de Vijver, M. J., et al. (2002) Gene expression profiling predicts clinical outcome in breast cancer. *Nature* 415, 530–536.
44. Fournier, M. V., Martin, K. J., Kenny, P. A., et al. (2006) Gene expression signature in organized and growth-arrested mammary acini predicts good outcome in breast cancer. *Cancer Res.* 66, 7095–7102.
45. Sgroi, D. C., Teng, S., Robinson, G., et al. (1999) In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res.* 59, 5656–5661.
46. Allinen, M., Beroukhim, R., Cai, L., et al. (2004) Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 6, 17–32.
47. Klein, C. A., Seidl, S., Petat-Dutter, K., et al. (2002) Combined transcriptome and genome analysis of single micrometastatic cells. *Nat. Biotechnol.* 20, 387–392.
48. Park, T., Yi, S. G., Lee, S., et al. (2003) Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics* 19, 694–703.
49. Churchill, G. A. (2004) Using ANOVA to analyze microarray data. *Biotechniques* 37, 173–175, 177.
50. Hughes, T. R., Marton, M. J., Jones, A. R., et al. (2000) Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126.
51. Taxman, D. J., MacKeigan, J. P., Clements, C., Bergstralh, D. T., Ting, J. P. (2003) Transcriptional profiling of targets for combination therapy of lung carcinoma with paclitaxel and mitogen-activated protein/extracellular signal-regulated kinase kinase inhibitor. *Cancer Res.* **63**, 5095–5104.
52. Nasr, R., Rosenwald, A., El-Sabban, M. E., et al. (2003) Arsenic/interferon specifically reverses 2 distinct gene networks critical for the survival of HTLV–1-infected leukemic cells. *Blood* **101**, 4576–4582.
53. Zhelev, Z., Bakalova, R., Ohba, H., et al. (2004) Suppression of bcr-abl synthesis by siRNAs or tyrosine kinase activity by Glivec alters different oncogenes, apoptotic/antiapoptotic genes and cell proliferation factors (microarray study). *FEBS Lett.* **570**, 195–204.
54. Ruddy, M. J., Wong, G. C., Liu, X. K., et al. (2004) Functional cooperation between interleukin-17 and tumor necrosis factor-alpha is mediated by CCAAT/enhancer-binding protein family members. *J. Biol. Chem.* **279**, 2559–2567.
55. Oishi, K., Miyazaki, K., Kadota, K., et al. (2003) Genome-wide expression analysis of mouse liver reveals CLOCK-regulated circadian output genes. *J. Biol. Chem.* **278**, 41,519–41,527.
56. Kwak, M. K., Wakabayashi, N., Itoh, K., et al. (2003) Modulation of gene expression by cancer chemopreventive diithiolethiones through the Keap1-Nrf2 pathway. Identification of novel gene clusters for cell survival. *J. Biol. Chem.* **278**, 8135–8145.
57. van de Vijver, M. J., He, Y. D., van’t Veer, L. J., et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009.
58. Righetti, P. G. (1990) Recent developments in electrophoretic methods. *J. Chromatogr.* **516**, 3–22.
59. Bjellqvist, B., Ek, K., Righetti, P. G., et al. (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J. Biochem. Biophys. Methods* **6**, 317–339.
60. Rabilloud, T., Vuillard, L., Gilly, C., Lawrence, J. J. (1994) Silver-staining of proteins in polyacrylamide gels: a general overview. *Cell Mol. Biol. (Noisy-le-grand)* **40**, 57–75.
61. Clauer, K. R., Baker, P., Burlingame, A. L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71**, 2871–2882.
62. Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C., Liotta, L. A. (2002) Clinical proteomics: translating benchside promise into bedside reality. *Nat. Rev. Drug Discov.* **1**, 683–695.
63. Weston, A. D., Hood, L. (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J. Proteome Res.* **3**, 179–196.
64. Sun, Z., Fu, X., Zhang, L., et al. (2004) A protein chip system for parallel analysis of multi-tumor markers and its application in cancer detection. *Anticancer Res.* **24**, 1159–1165.
65. Nomura, F., Tomonaga, T., Sogawa, K., et al. (2004) Identification of novel and downregulated biomarkers for alcoholism by surface enhanced laser desorption/ionization-mass spectrometry. *Proteomics* **4**, 1187–1194.
66. Volmer, M. W., Radacz, Y., Hahn, S. A., et al. (2004) Tumor suppressor Smad4 mediates downregulation of the anti-adhesive invasion-promoting matricellular protein SPARC: Landscaping activity of Smad4 as revealed by a “secretome” analysis. *Proteomics* **4**, 1324–1334.
67. Krah, A., Schmidt, F., Becher, D., et al. (2003) Analysis of automatically generated peptide mass fingerprints of cellular proteins and antigens from *Helicobacter pylori* 26695 separated by two-dimensional electrophoresis. *Mol. Cell Proteomics* , 1271–1283.
68. Brown, P.O. (2006). Exploring along a crooked path. *Am. J. Human Genet.* **79**, 429–433.