Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 37 (2004) 293–303

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Methodological Review

A primer on gene expression and microarrays for machine learning researchers

Winston Patrick Kuo[a,b,c,d], Eun-Young Kim[a,b], Jeff Trimarchi[c], Tor-Kristian Jenssen[e], Staal A. Vinterbo[a,b], Lucila Ohno-Machado[a,b,*]

[a] *Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
[b] *Division of Health Sciences and Technology, Harvard University and MIT, Cambridge, MA, USA*
[c] *Department of Genetics, Harvard Medical School, Boston, MA, USA*
[d] *Department of Oral Medicine, Infection, and Immunity, Harvard School of Dental Medicine, Boston, MA, USA*
[e] *PubGene, Inc., Oslo, Norway*

Received 20 June 2004
Available online 8 September 2004

Abstract

Data originating from biomedical experiments has provided machine learning researchers with an important source of motivation for developing and evaluating new algorithms. A new wave of algorithmic development has been initiated with the publication of gene expression data derived from microarrays. Microarray data analysis is particularly challenging given the large number of measurements (typically in the order of thousands) that are reported for relatively few samples (typically in the order of dozens). Many data sets are now available on the web. It is important that machine learning researchers understand how data are obtained and which assumptions are necessary in the analysis. Microarray data have the potential to cause significant impact in machine learning research, not just as a rich and realistic source of cases for testing new algorithms, as has been the UCI machine learning repository in the past decades, but also as a main motivation for their development. In this article, we briefly review the biology underlying microarrays, the process of obtaining gene expression measurements, and the rationale behind the common types of analyses involved in a microarray experiment. We outline the main challenges and reiterate critical considerations regarding the construction of supervised learning models that use this type of data. The goal of this article is to familiarize machine learning researchers with data originated from gene expression microarrays.

Keywords: Bioinformatics; Microarrays; Machine learning

1. Introduction

Understanding the inter-cellular and intra-cellular processes underlying many diseases is essential for improving the capacity to diagnose and treat patients. Unraveling the complexity underlying these biological processes is expected to provide novel predictive tools to enable, for example, the sub-classification of diseases and eventual identification of proper therapeutic drug targets. The potential for further characterization of regulatory networks and pathways controlling the cellular homeostasis that are altered in diseases has been seen as one of the main promises of global analyses of gene expression profiles. Microarrays that measure the expression of thousands of genes simultaneously have been perceived by many as a first step towards this ambitious goal [1–4].

Biomedical researchers are trying to discover relationships between genes and disease or developmental stages, as well as relationships among genes. For

* Corresponding author. Fax: +1 617 739 3672.
E-mail addresses: wkuo@genetics.med.harvard.edu (W.P. Kuo), machado@dsg.harvard.edu (L. Ohno-Machado).

1532-0464/$ - see front matter © 2004 Elevier Inc. All rights reserved.
doi:10.1016/j.jbi.2004.07.002

example, an application of microarrays can be used for discovery of novel biomarkers for cancer, which can provide more accurate diagnosis and monitoring tools for early detection of a particular subtype of disease or assessment of effectiveness of a particular treatment protocol. Since microarray experiments have been growing in terms of usage across laboratories, the amount of data is rapidly growing, thus creating an environment for new computational strategies.

1.1. A new type of data

The availability of public repositories of data that are suitable for machine learning research is extremely important to the field of biomedical informatics. The UCI machine learning repository (http://www.ics.uci.edu/~mlearn/MLRepository.html), which contains clinical and biological datasets, has allowed a large number of researchers to demonstrate the performance of new statistical and machine learning algorithms in the past two decades. New centralized (http://www.ncbi.nlm.nih.gov/geo/) and decentralized sources of biomedical data (mostly from supplemental data in journals or researchers' web sites) have started to accumulate since the advent of high-throughput measurement technologies such as gene expression microarrays. This new and free source of data is receiving increasing attention from researchers in the machine learning community, and the number of publications is increasing at a rapid pace. The distributed collection of data has an important role as a rich source for testing new algorithms for pattern recognition, and it also serves as an important motivation for their development, as it presents important challenges to the direct application of existing algorithms. As in other domains, to properly employ machine learning models to these data, it is imperative that the researcher understands their potential and limitations. The goal of this article is to review certain aspects of gene expression microarray measurements, describe common analytical approaches, and familiarize machine learning researchers with data generated by these technologies.

1.2. Measuring gene expression

Until about a decade ago, the ability to identify and analyze gene expression patterns has been technically limited to a handful of genes per study. These traditional methods include Northern blot and real-time PCR [5–8], which, although fruitful and still in use for biological validation experiments, have limitations in terms of the number of gene expression patterns that can be analyzed in practice. This limitation has been overcome by the development of several high-throughput technologies that allow more comprehensive coverage of genes, although the measurement for each gene is usually less accurate than that resulting from low-throughput technologies. Some of the methods include differential display [9], serial analysis of gene expression (SAGE) [10], and massive parallel signature sequencing (MPSS) [11]. Other high-throughput methods, though still in their infancy, can measure protein expression levels: protein microarrays [12], and mass spectrometric analysis [13]. The goal of DNA microarray technologies is to measure the level of expression for large sets of genes, in a global fashion. Although less precise than traditional low-throughput methods, the information gained from measuring the expression of thousands of genes simultaneously is considered significant, particularly in exploratory phases of research. These technologies are based on the measurement of messenger RNA (mRNA), which is described in the context of its biological role in Section 2. Some common issues in measuring gene expression are pervasive across these techniques. For example, it is not the case that the amount of mRNA produced is always directly proportional to a known function that is important in disease processes such as translation into protein or regulation of another gene [13]. However, the analyses of gene expression patterns usually equate the amount of mRNA detected for a certain gene with its functional status.

There are two common microarray platforms for investigating gene expression: complementary DNA (cDNA) and oligonucleotide microarrays [14,15]. They differ in experimental protocols, lengths of probes, number of tissues measured per array thus implying challenges in the integration and comparison of data sets from different platforms [16]. Once issues with standardization are resolved and new algorithms for their analyses are developed, the range of applications of microarrays will be potentially vast. They have been used to study expression profiles of genes in areas of development [17], the study of progression of a disease [18], survival [19], and response to various drug compounds [20].

2. The biology behind gene expression microarray measurements

Genomics is a broad category describing the development and application of genetic information that has the potential to lead to qualitative changes in the way in which biomedical research is conducted in terms of diagnostics, risk assessment, therapeutics, and health care outcome. The post- era has brought with it the promise of change in the way basic experiments are conducted, enabling biomedical researchers to examine biological systems more comprehensively. Since the inception of the human genome project (HPG) about a decade ago, the 3.2 billion base pairs that make up the genome have been sequenced to near perfection. The human genome is believed to have between 30,000 and 40,000 genes

(i.e., "coding" regions of DNA that are important in the assembly of proteins or regulation of other regions) [21], each composed of hundreds to thousands of nucleotides of four types: adenine (A), cytosine (C), thymine (T), and guanine (G). The sequence information of our genome serves as the basis for development of DNA microarrays. However, knowledge of the sequence of nucleotides in a gene does not directly lead to knowledge regarding the level of expression of that gene (i.e., whether the gene is up-, down-, or neutrally regulated)

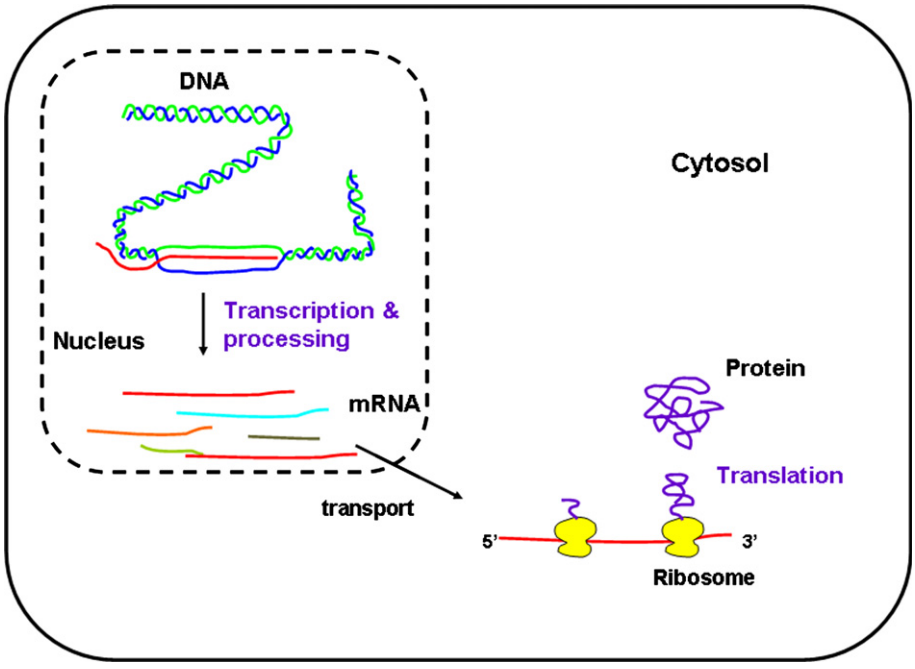

Fig. 1. Transcription of DNA to mRNA and translation of mRNA to protein. Activities of the cell are controlled by instructions contained in the DNA sequences, through mRNA that carries the genetic information (transcription) from the cell to the cytoplasm, where proteins are produced (translation).

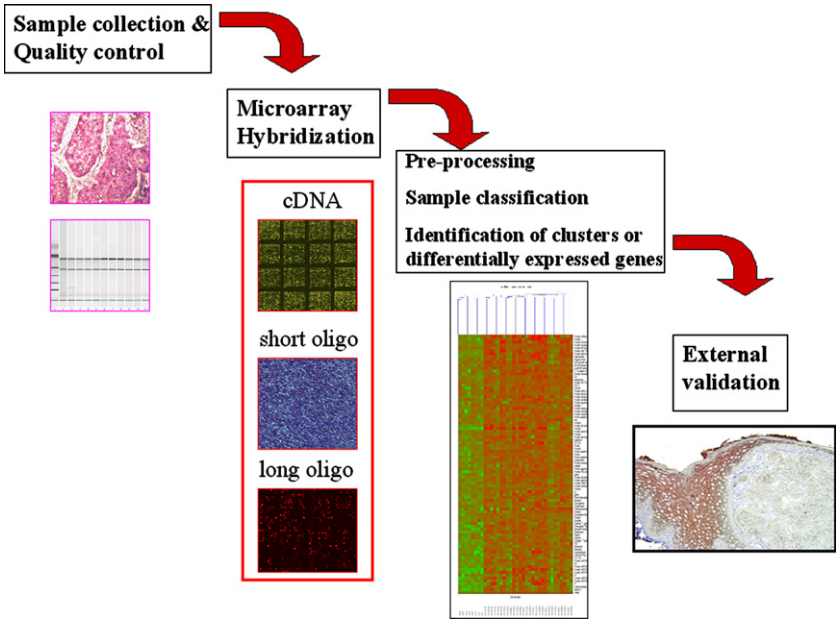

Fig. 2. Overview of a microarray experiment: procedure used in gene expression profiling for target identification and biological validation. mRNA extraction from a tissue biopsy and quality control of the samples is followed by labeling of a probe(s), hybridization onto a microarray, data acquisition, and analysis. Biological verification of the results can be performed using a variety of approaches, such as immunohistochemistry.

in a particular environmental or developmental condition. Although all cells in the body possess the same DNA, gene expression varies according to developmental stage, tissue, age, and environmental conditions. In the study of health and disease, the goal of microarray analysis is to characterize certain gene expression "profiles" in which the levels of expression of particular sets of genes are highly associated.

Gene expression is an intermediate step before the assembly of proteins from their building blocks, the amino acids. When a gene is expressed, messenger RNA (mRNA) is produced ("transcribed") from the gene's DNA sequence, and it serves as a template to guide the synthesis of a protein, allowing particular amino acids to be systematically incorporated into a protein (Fig. 1). The mRNA transcript is a complement of a corresponding part of the DNA coding region. The purpose of a gene expression microarray is to measure how much mRNA corresponding to a particular gene is present in the cell(s) or tissue of interest.

In general, a microarray (or chip) is composed of thousands of DNA sequences (probes), corresponding to segments of genes that are placed in specific arrangement, typically on a glass slide or a silicon microchip. The DNA sequences can be short, as in the case of oligonucleotide arrays, or long, as in the case of cDNA arrays. The principle behind microarrays is that complementary sequences will bind to each other under the proper conditions, whereas non-complimentary sequences will not bind. For example, if the DNA sequence on an array is 10 nucleotides long, TACCGAA CTG, the sequence ATGGCTTGAC will "hybridize" to the probe ('A' nucleotides complement 'T' and 'C' nucleotides complement 'G'). Probes are designed to be specific to a gene. On a microarray, many thousands of spots are placed onto a grid, each spot containing the DNA sequence from a particular gene. When a sample of interest contains many copies of mRNA, many bindings can take place, indicating a gene from which the mRNA transcribed is highly expressed.

The steps involved in common gene expression microarray studies are depicted in Fig. 2. The experiments begin with the collection of samples of a certain tissue (or cell) of interest, and the extraction of mRNA from these samples. Since mRNA can be easily degraded, special attention is required for the collection, preparation, and storage of these samples. A quality control step is essential before conducting an experiment, such as running the mRNA samples on an agarose gel. Since mRNA is inherently unstable, cDNA, which is more stable and easier to work with, is produced in the laboratory from the mRNA, and represents an equivalent sequence of nucleotides.[1] This cDNA is labeled with a fluorescent dye and will hybridize (i.e., bind specifically with pre-determined sequences of nucleotides representative of a certain gene) to sequences that are immobilized on the microarray. Quantifying the fluorescence signal intensity allows one to assess the amount of hybridization. Those sequences that do not hybridize will be washed away leaving no signal. Images from the fluorescent probes are read by a scanner and translated into numerical values. The mRNA abundance in a cell or tissue (or corresponding cDNA that is made from it in the laboratory) is therefore a proxy for the measure of gene expression: when certain genes become "expressed," many copies of mRNAs corresponding to those genes are produced. These copies will hybridize with microarray probes that are complementary (Fig. 3). The major assumption is that the abundance of mRNA corresponding to a certain gene is positively correlated with the expression of a certain gene.

[1] cDNA derived from the specimens can be used for both cDNA and oligonucleotide arrays.

3. How to access the data?

Many data sets are publicly available via the internet, additionally there are hundreds of life sciences databases reported in the literature [22], which stresses the difficulty to where and how to search information in a fast and efficient manner [23]. Fig. 4 illustrates common formats for publication of gene expression data from microarrays. The published data usually constitute a transformed version of the initial data set, and has usually been subject to pre-processing in the form of filtering and normalization. Gene expression data derived from microarrays can be obtained in web supplements to journal publications or in public repositories. There are a number of efforts well underway to create public gene expression databases. Two leading contenders that have become the de-facto public databases for arrays are Array Express at the European Bioinformatics Institute (http://www.ebi.ac.uk/arrayexpress/), and the NCBI's Gene Expression Omnibus GEO (http://www.ncbi.nlm.nih.gov/geo/). Other important sources of microarray data are the Stanford Microarray Database (http://genome-www5.stanford.edu/), the Duke Microarray Center database (http://mgm.duke.edu/genome/dna_micro/work/), and the Whitehead Institute Cancer Genomics database (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi). In this context, it is important that this information be archived in standardized fashion, which is usually not the case for many journal supplements and individual laboratory web sites. This effort towards standardization has been initiated by the Microarray Gene Expression Data (MGED www.mged.org) Society, which has taken the initiative to develop and enforce guidelines, formats and tools for submission of microarray data [24]. This allows researchers to share com-

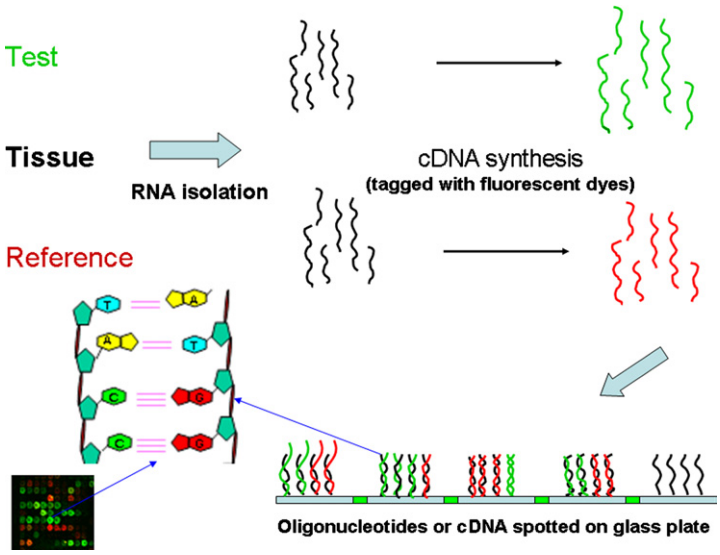

Fig. 3. Microarrays contain thousands of probes that can vary in length (from 25 to over 1000 bp) that are affixed onto a solid surface. Microarray experiments can be broken down to two groups based on their labeling bias, one-dye or two-dye experiments. Essentially in two sample experiments, two samples are labeled either with a Cy3-dye or a Cy5-dye, producing a ratio unit measurement, whereas in a single-dye experiment, an absolute unit of measurement is generated.

Two different platforms (cDNA, Affymetrix)

Block	Column	Row	Name	ID	X	Y	Dia.	F635 Medi ▪ ▪ ▪ ▪ ▪
1	1	1	ESTs	BE995540	2190	6730	130	5518
1	2	1	ESTs, Mo(	BE995610	2460	6730	100	7413
1	3	1	ESTs	BE995608	2730	6730	80	5691
1	4	1	EST	BE995606	2970	6720	130	4974
1	5	1	heat shock	BE995564	3220	6740	120	8851

1	Analysis Name	Probe Set Name	Stat Pairs	Stat Pairs	Signal	Detection	Detection p-value ▪ ▪ ▪ ▪ ▪
1	052203_MRP1_A11	AFFX-MurIL2_at	20	20	38.4	A	0.966323
1	052203_MRP1_A11	AFFX-MurIL10_at	20	20	100.1	A	0.368438
1	052203_MRP1_A11	AFFX-MurIL4_at	20	20	218.1	A	0.250796
1	052203_MRP1_A11	AFFX-MurFAS_at	20	20	138.5	A	0.275146
	052203_MRP1_A11	AFFX-BioB-5_at	20	20	1336.5	P	0.015183
	052203_MRP1_A11	AFFX-BioB-M_at	20	20	1715.6	P	0.000169
	052203_MRP1_A11	AFFX-BioB-3_at	20	20	657.6	P	0.001248
	052203_MRP1_A11	AFFX-BioC-5_at	20	20	3311.9	P	0.000127
	052203_MRP1_A11	AFFX-BioC-3_at	20	20	2299.2	P	0.000044
	052203_MRP1_A11	AFFX-BioDn-5_at	20	20	4379	P	0.00011

Fig. 4. Illustration of different raw data formats that were generated using different extraction software after scanning, GenePix 4.0 for cDNA data and MAS 5.0 for Affymetrix GeneChips. In cDNA data (two-dye, two-sample experiment), there are different attributes that are used in the analyses; intensity, background data for both dyes, and flags for bad spots. In Affymetrix data (one-dye, one sample), one can use the detection calls (absent, marginal, and, present) and signal intensity for analyses.

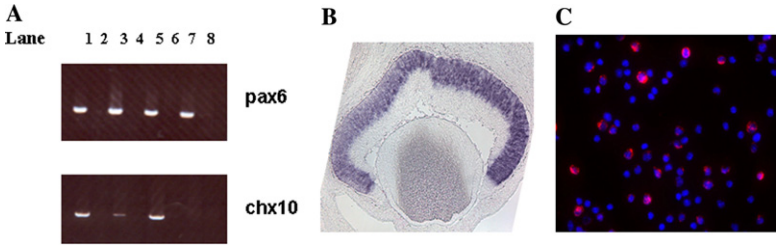

Fig. 5. (A) Representative example of gene specific PCR for microarray validation. PCR using primers specific for gene pax6 or chx10 was performed on the cDNA preparations from E15 single retinal progenitor cells (lanes 1, 3, 5, and 7) or controls (lanes 2, 4, 6, and 8). (B) Representative examples of in situ hybridizations for microarray validation. E15 retinas were either sectioned (B) or dissociated (C) and hybridized with a probe for gene Fgf15.

mon information and make valid comparisons among experiments. MGED is an international organization of scientists involved with gene expression profiles. Their primary contributions are proposed standards for publication and data communication. MGED proposed minimum information about a microarray experiment (MIAME) as a potential publication standard [25]. MAGE-ML is a proposed mark-up language for microarray experiments (http://www.mged.org/Workgroups/MAGE/mage-ml.html).

A typical experiment has thousands of genes, few samples, and minimal information other than gene expression measurements. Some experiments link clinical data to the information about gene expression in the arrays, but this information is usually minimal (e.g., survival days, age, and gender) and rarely published, so it is not appropriate to attempt to utilize the resulting models in the clinical setting. Not all available microarray data sets are appropriate for machine learning research. Many experiments contain too few arrays. Although each array contains information on thousands of genes, it may not make sense to try to extract generalizable patterns from this type of experiments. It is important to emphasize that, for most machine learning applications, the unit of analysis is not the gene, but rather the array (tissue sample). Just as it does not make sense to establish phenotypic patterns of disease (e.g., define the profile of laboratory tests that is related to diabetes) by analyzing the results for one or two patients with disease and a couple of healthy subjects, it does not make sense to establish gene profiles by examining a few arrays. The sample size has to be sufficiently large for the construction of generalizable models. Since accruing subjects or obtaining specimens and performing measurements in microarrays is still costly, there is an important limitation in terms of which data sets can be effectively used for machine learning research.

Table 1 lists some data sets that have been previously used in machine learning experiments. The extent to which patterns can be learned and generalized from these data sets is variable. The list contains data from experiments with a relatively large number of cases that can be used by machine learning researchers who would like to get familiar with this type of data. Considerable pre-processing of these data sets is necessary before they can be used for machine learning research, although pre-processed versions of these data may often be obtained upon request.

4. Common modeling techniques and computational challenges

The advantage of DNA microarrays is that they allow the study of multiple transcriptional events in a single chip, which corresponds to one experiment. Therefore, the values for thousands of variables (genes) can be simultaneously measured for a particular biological sample. The cost of processing each sample, however, is relatively high, so there are few cases (biological samples) per study. Significant steps towards increasing the reliability of measurements have been taken in the past few years. The main challenge in microarray studies resides in proper study design and efficient and realistic interpretation of the information. There are several issues that need to be considered for a study, such as: (i) the type of DNA microarray platform selected for the experiment (each having different protocols, sensitivities, and specificities); (ii) mRNA preparation (such as type of specimen, availability, heterogeneity); and (iii) data analysis (pre-processing, unsupervised, and supervised learning). Since each step of a microarray experiment is subject to different sources of variability, there is large variance in microarray measurements; it is common practice to have duplicate or triplicate arrays for the same sample [26]. A major hurdle is the extraction of meaningful information from the large amount of expression data generated from microarray studies (large m, small n problem, where m is the number of variables or gene measurements, and n is the number of observations from which those measurements are obtained). Other issues that should not be overlooked when analyzing microarray data are outliers and missing values.

4.1. Pre-processing

In a typical microarray analysis, the initial step is pre-processing of the data, which includes filtering and

Table 1
List of some microarray data sets used in machine learning research

Specimen	No. of samples	No. of genes	Platform	Author
Adenocarcinomas	279	9376 common	Affymetrix, cDNA	Ramaswamy et al. [63]
Breast cancer	117	~25,000	cDNA	van't Veer et al. [64]
Drosophila melanogaster	66	4028	cDNA	Arbeitman et al. [65]
Prostate cancer	52	~12,600	Affymetrix	Singh et al. [66]
CNS embryonal tumors	99	6817	Affymetrix	Pomeroy et al. [67]
Primary tumors	144	16,063	Affymetrix	Ramaswamy et al. [35]
Small round blue cell tumors	63	6567	cDNA	Khan et al. [3]
Lung carcinomas	186	12,600	Affymetrix	Bhattacharjee et al. [21]

normalization. Filtering is an approach to reduce the number of genes for further data analysis. Measurements of genes that represent many copies per cell (high abundance) tend to be more reliable and more consistent than measurements that represent genes with low abundance regardless of the microarray technology utilized. A natural explanation may be that there is inherent noise in the hybridization and/or signal detection processes, which for weak signals of low-abundant transcripts results in lower signal to noise ratio compared to the stronger signals of high-abundant transcripts. This effect is seen in one-dye (one sample) platforms, but in two-dye platforms, from which ratios are calculated, these effects are amplified and the ratio estimates are confounded by uncertainty of two numbers rather than by only one. As a result, it has become common practice to 'filter' out weak signals, for which it is assumed that the signal-to-noise ratio is too low for the data to be useful. This filtering process includes the identification and removal of array elements prior to further analyses. It is important to investigate the amount of true signal that is being filtered out in this filtering process. Future research on microarray analysis should formally address this issue. Some important transcription factors exhibit low but differentiable levels of expression under different circumstances. This type of information may be lost in the filtering process.

After filtering out unreliable observations, many sources of systematic and experimental variation that confound the observed gene expression levels still remain in the microarray experiments. Normalization is a method to adjust the means or variances of the variation in the measured expression intensities. This improves the monitoring of biological differences and allows the comparison of expression levels across multiple experiments. A review of normalization procedures is beyond the scope of this article. Readers should refer to [21,27–29] for details.

4.2. Statistical and machine learning models

In addition to coding well-known algorithms for data pre-processing and evaluation, new algorithms and approaches for clustering, classification, and evaluation are needed. Various mathematical and statistical tools have been developed to cluster genes by integrating them with biological pathways or to perform class predictions that identify expression patterns that correlate with phenotypic characteristics. The development of new machine learning models to extract knowledge (such as relationships between genes and disease and among genes themselves) [30] from large data sets is very important in this stage of genomic research. There is a need to expand the set of models available to researchers, allow them to select the adequate models for their data, investigate new ways to determine the importance of individual variables and individual observations, and make it possible to combine models.

Differential expression of genes under various conditions or time points is usually the focus of the analysis. This type of analysis can be done using one gene at a time (univariate) or several genes at a time (multivariate). Univariate analysis of gene expression data sets is sometimes useful and has been utilized in several studies. Examples of univariate analyses include the inspection of fold-differences, calculation of p values using methods analogous to t tests, and ANOVA [31–33]. As our understanding of biology increases, it becomes evident that there are many instances in which a combination of genes, rather than one gene in isolation, may contribute to the biological process under investigation. Therefore, it is critical that we analyze gene expression data using a multivariate approach, even though univariate analysis may still play a role in filtering genes in the pre-processing phase [34].

Unsupervised learning is done in order to investigate which genes behave in a similar manner or which genes exhibit high covariance. This was a very popular technique when the number of cases studied was small (for example, when tissues at a certain developmental stage or condition were pooled together to be analyzed by a single array). There are several examples in the literature on the use of hierarchical clustering, self-organizing maps, multidimensional scaling, and several clustering algorithms in gene expression data [35–38]. This exploratory analysis technique is a good first step towards identifying clusters of related genes, but its use for identifying disease progression markers or clustering cases into categories of interest is limited. Unsupervised techniques are in many cases inappropriately utilized as a replacement to supervised techniques (e.g., in studies in which researchers want to classify cases into known categories such as benign versus malignant). In this context, the analysis cannot guarantee that the clusters of interest will result from the data, although this may indeed happen. It is worth noticing that distances and other non-weighted measures of association between objects can be misleading in the task of forming clusters of interest. This is especially true when the sample size is small.

As larger experiments involving microarrays are being published, there is an increasing number of publications reporting the analysis of microarray data using supervised classification algorithms, such as support vector machines [39], artificial neural networks [3], regression [40], and various types of rule-induction algorithms [41]. Tools to support classification algorithms that work with a high ratio of variables/cases are still rare. The problem with using conventional statistical multivariate techniques, such as discriminant analysis and regression, is that they do not work when the number of variables exceeds the number of cases, which so

far has been the case in all microarray experiments. Therefore, techniques for reducing the number of variables that are directly used in the models have been utilized. Some of these, such as principal components analysis, are unsupervised in nature, and therefore do not guarantee that the reduction is done in a way that optimizes the predictive model. Furthermore, they merely "compress" the variables and the resulting components do not have any particular meaning. Utilizing partial least squares partially resolves the first issue (as it is a supervised variable "compression" technique), but still does not allow direct identification of important variables (genes), which is the goal of many analyses. In order to do this, variable selection algorithms are needed.

Identifying which genes contribute the most for the estimate in a particular predictive model can be done via variable selection methods. A number of variable selection techniques exist and some have been used in the bioinformatics literature. Some algorithms, such as forward stepwise selection of variables, utilize a gradient descent approach, with or without a stochastic component, making them more or less susceptible for stopping after reaching non-optimal solutions (i.e., local minima). Some authors indicate that this greedy approach can perform as well as more complex ones in a number of data sets [42]. Other algorithms, such as those based on evolutionary techniques (such as genetic algorithms) can rarely be used without a significant pre-selection of variables, given time constraints. Discovering new paradigms for variable selection in predictive models is critical for defining few genetic markers for disease progression in models derived from microarray data. Hence the increasing number of research articles that deal with variable selection for microarray data analysis [31].

Variable selection techniques can be divided into those that perform the selection using a purely univariate approach, and those that are multivariate. Purely univariate techniques have been used in the bioinformatics literature [43,44], but multivariate techniques are just beginning to be investigated. For example, the Goodman–Kruskal association index was used in the context of partitions to select biomarkers for malignancy in [45]. New methods for variable selection are needed. Published work in this area includes variations of genetic algorithms to promote *en bloc* selection of variables, which were shown to result in variable selections that outperform classic sequential forward, backward, or stepwise selection procedures [46]. In microarray analysis, data sets consist of thousands of variables, and often no more than dozens of cases. Related to the variable selection problem is that of overfitting, which has been investigated in the context of remedial strategies such as bootstrapping [47–49], cross-validation (including jackknife) [50,51], shrinkage, and other methods [52].

Some machine learning researchers overlook the importance of conducting proper evaluation of models, especially the issue of overfitting. The "curse of dimensionality" is well illustrated in microarray data, and researchers who are testing new algorithms may get a false impression that their models will generalize well to new data just because they can perfectly fit the training data.

As in similar problems, to attenuate the problem of overfitting, without increasing the cost of the experiment, two approaches can be potentially useful:

1. Resampling of training cases, such that test cases are not used for building a model. No new information is added with this strategy, but a better assessment of the generalizability of the model is achieved this way. Techniques such as cross-validation and bootstrap can be used and are discussed in [53]. Given that the number of cases in a particular category may be small, it is advisable to create cross-validation partitions by randomly sampling cases from each category in a way that the proportion of cases remains the same in each partition. A large number of bootstrap samples and related models can be generated by randomly sampling the arrays with replacement. Non-sampled cases can be used for testing of each model.
2. Decrease in the number of variables. Variable selection methods are used for this purpose, as discussed before.

A topic of ultimate importance for the mathematical validation of results is the choice of evaluation indices. Unsupervised learning models produce results that are difficult to interpret and evaluate objectively. For clustering models, the most acceptable strategy is to develop clusters according to several different objective functions (e.g., maximization of the inter-cluster Euclidean distances over intra-cluster distances) and evaluate the resulting clusters according to [1] measures of cluster concordance, and [2] known properties of the objects (e.g., functional classification) [54,55]. For classification problems, regions under the ROC curve [56,57] and standard confidence intervals can be used to assess discrimination ability of the models even in multi-categorical cases [58]. When new supervised learning algorithms are being tested, it is important to compare their results with those deriving from established methodologies such as regression. When different models are compared on the same data, adjustments to the confidence interval need to be made to account for the correlation of results. This results in narrower confidence intervals and improved potential to detect differences in performance. Calculating confidence intervals without adjusting for correlations is more conservative and hence less likely to demonstrate differences.

5. Biological validation of results

After a comprehensive data analysis, the list or cluster of genes that have been identified as linked to a particular condition or developmental process requires further investigation to determine its biological significance. It is imperative for biomedical investigators to assess the false-positive rate and conduct independent biological validations to confirm the results generated computationally. The most commonly used techniques to verify gene expression data include Northern blots and PCR-based approaches (both quantitative and semi-quantitative) (Fig. 5A). The advantages of these methods are twofold. First, these methods can be performed quantitatively. Secondly, Northern blot and PCR can be used to screen through a large number of candidates relatively rapidly. Other methods of validation include in situ hybridization (Fig. 5B) and immunohistochemistry. These approaches offer the extra benefit of showing exactly where in a particular tissue the candidate genes are expressed [59,60]. However, traditionally these methods are neither quantitative nor high-throughput. Recently, however, several studies have demonstrated that in situ hybridizations can be used to screen a large number of candidate genes [17] and, in some cases, this can be performed in a quantitative manner (Fig. 5C) [17,61].

A candidate gene-by-gene validation approach, which involves experiments with a few potentially important genes, is a viable and successful method. This approach is not appropriate to provide a more comprehensive understanding of the larger process of biological networks and pathways. Often, the same signaling molecules or transcription factors are commonly expressed in multiple tissues or stages. Studies suggest that the same gene can play distinct developmental roles in these circumstances, being significant in one tissue but unessential or redundant in another. Moreover, the function of any gene is often context dependent. For example, the ability of a signaling molecule to activate a specific differentiation pathway largely depends on the target cell's competence to receive and interpret the signal and its ability to utilize an existing signal transduction system that, in the presence of appropriate co-factors and nuclear transport assemblage, will activate or down-regulate downstream genes. In short, the functional importance of any gene depends on the presence or absence of products of many other genes. If, as some studies suggest, binary and more complex combinations of signaling molecules are needed to control certain biological processes, then this problem is magnified many-fold. Here again, a larger scale genomic approach coupled with sophisticated analyses could be potentially useful for elucidating the molecular network behind particular biological phenomena. This can be illustrated by using tools that integrate both microarray results and resources from private and public databases to generated networks/pathways that can assist our understanding in how these genes and/or proteins interact with each other as seen in [62].

6. Conclusion

In conclusion, new sources of data such as those derived from gene expression microarrays offer new challenges for the development and evaluation of statistical and machine learning algorithms. The large number of variables per observation can give researchers a false impression of having ''lots of data'' that are useful for machine learning research. In fact, most of the current data sets contain few observations, hence the main challenge is to select a small set of variables that are representative of the data, and build models that are potentially generalizable to a new set of cases. Careful internal and biological validation allows quantitative assessment of the potential for generalization of the machine learning models derived from the data. It is important to note that generalization will depend on a series of other issues as well, such as: microarray platform, reference sample, manipulation of the samples, normalization procedures, and intrinsic noise in the measurements.

Acknowledgments

The authors thank Prof. Jose A. Lozano and two anonymous reviewers for their insightful comments.

References

[1] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403(6769):503–11.
[2] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7.
[3] Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7(6):673–9.
[4] Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406(6797):747–52.
[5] Alwine JC, Kemp DJ, Parker BA, et al. Detection of specific RNAs or specific fragments of DNA by fractionation in gels and transfer to diazobenzyloxymethyl paper. Methods Enzymol 1979;68:220–42.
[6] Harkin DP, Bean JM, Miklos D, et al. Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. Cell 1999;97(5):575–86.
[7] Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res 1996;6(10):986–94.
[8] Mills JC, Roth KA, Cagan RL, Gordon JI. DNA microarrays and beyond: completing the journey from tissue to cell. Nat Cell Biol 2001;3(8):E175–8.

[9] Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257(5072):967–71.
[10] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270(5235):484–7.
[11] Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 2000;18(6):630–4.
[12] Haab BB, Dunham MJ, Brown PO. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. Genome Biol 2001;2(2). RESEARCH0004.
[13] Pandey A, Mann M. Proteomics to study genes and genomes. Nature 2000;405(6788):837–46.
[14] Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274(5287):610–4.
[15] Adams MD, Kelley JM, Gocayne JD, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 1991;252(5013):1651–6.
[16] Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS. Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 2002;18(3): 405–12.
[17] Blackshaw S, Harpavat S, Trimarchi J, et al. Genomic analysis of mouse retinal development. PLoS Biol 2004;2(9):E247.
[18] Kuo WP, Jenssen TK, Park PJ, Lingen MW, Hasina R, Ohno-Machado L. Gene expression levels in different stages of progression in oral squamous cell carcinoma. Proc AMIA Symp 2002:415–9.
[19] Jenssen TK, Kuo WP, Stokke T, Hovig E. Associations between gene expressions in breast cancer and patient survival. Hum Genet 2002;111(4–5):411–20.
[20] Scherf U, Ross DT, Waltham M, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24(3):236–44.
[21] Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001;98(24):13790–5.
[22] Baxevanis AD. The molecular biology database collection: 2002 update. Nucleic Acids Res 2002;30(1):1–12.
[23] Stein L. Creating a bioinformatics nation. Nature 2002;417(6885):119–20.
[24] Ikeo K, Ishi-i J, Tamura T, Gojobori T, Tateno Y. CIBEX: center for information biology gene expression database. Crit Rev Biol 2003;326(10–11):1079–82.
[25] Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 2001;29(4):365–71.
[26] 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 USA 2000;97(18):9834–9.
[27] Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004;20(3):307–15.
[28] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31(4):e15.
[29] Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4(2):249–64.
[30] Kuo WP, Mendez E, Chen C, et al. Functional relationships between gene pairs in oral squamous cell carcinoma. Proc AMIA Symp 2003:371–5.
[31] Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 2002;18(11):1454–61.
[32] Baggerly KA, Coombes KR, Hess KR, Stivers DN, Abruzzo LV, Zhang W. Identifying differentially expressed genes in cDNA microarray experiments. J Comput Biol 2001;8(6):639–59.
[33] Didier G, Brezellec P, Remy E, Henaut A. GeneANOVA—gene expression analysis of variance. Bioinformatics 2002;18(3):490–1.
[34] Phillips TJ, Belknap JK. Complex-trait genetics: emergence of multivariate strategies. Nat Rev Neurosci 2002;3(6):478–85.
[35] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 2001;98(26):15149–54.
[36] Nikkila J, Toronen P, Kaski S, Venna J, Castren E, Wong G. Analysis and visualization of gene expression data using self-organizing maps. Neural Netw 2002;15(8–9):953–66.
[37] Fuller GN, Hess KR, Rhee CH, et al. Molecular classification of human diffuse gliomas by multidimensional scaling analysis of gene expression profiles parallels morphology-based classification, correlates with survival, and reveals clinically relevant novel glioma subsets. Brain Pathol 2002;12(1):108–16.
[38] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95(25):14863–8.
[39] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16(10):906–14.
[40] Shannon WD, Watson MA, Perry A, Rich K. Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. Genet Epidemiol 2002;23(1):87–96.
[41] Ressom H, Reynolds R, Varghese RS. Increasing the efficiency of fuzzy logic-based gene expression data analysis. Physiol Genomics 2003;13(2):107–17.
[42] Weber G, Vinterbo S, Ohno-Machado L. Multivariate selection of genetic markers in diagnostic classification. Artif Intell Med 2004;31(2):155–67.
[43] Welle S, Brooks AI, Thornton CA. Computational method for reducing variance with Affymetrix microarrays. BMC Bioinformatics 2002;3(1):23.
[44] Butte AJ, Ye J, Haring HU, Stumvoll M, White MF, Kohane IS. Determining significant fold differences in gene expression analysis. Pac Symp Biocomput 2001:6–17.
[45] Jaroszewicz S, Simovici DA, Kuo WP, Ohno-Machado L. The Goodman–Kruskal coefficient and its applications in genetic diagnosis of cancer. IEEE Trans Biomed Eng 2004;51(7):1095–102.
[46] Vinterbo S, Ohno-Machado L. A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction. Proc AMIA Symp 1999:984–8.
[47] Zhang Z, Harrison P, Gerstein M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res 2002;12(10):1466–82.
[48] Zhao LP, Prentice R, Breeden L. Statistical modeling of large microarray data sets to identify stimulus-response profiles. Proc Natl Acad Sci USA 2001;98(10):5631–6.
[49] Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S, Tainsky MA. Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. Bioinformatics 2003;19(11):1348–59.
[50] Lyons-Weiler J, Patel S, Bhattacharya S. A classification-based machine learning approach for the analysis of genome-wide expression data. Genome Res 2003;13(3):503–12.
[51] Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 2004;5(2):155–76.
[52] Krzanowski W. Selection of variables to preserve multivariate data structure, using principle components. Appl Stat 1987:22–33.

[53] Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification?. Bioinformatics 2004;20(3):374–80.
[54] Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 1971;66:846–50.
[55] Jaccard P. Nouvelles rescherches sur la distribution florale. Bull Soc Vaud Sci Nat 1908;44:223–70.
[56] Song HH. Analysis of correlated ROC areas in diagnostic testing. Biometrics 1997;53(1):370–82.
[57] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148(3):839–43.
[58] Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Med Decis Making 2000;20(3):323–31.
[59] Chiang MK, Melton DA. Single-cell transcript analysis of pancreas development. Dev Cell 2003;4(3):383–93.
[60] Tietjen I, Rihel JM, Cao Y, Koentges G, Zakhary L, Dulac C. Single-cell transcriptional analysis of neuronal progenitors. Neuron 2003;38(2):161–75.
[61] Kamme F, Salunga R, Yu J, et al. Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. J Neurosci 2003;23(9):3607–15.
[62] Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001;28(1):21–8.
[63] Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet 2003;33(1):49–54.
[64] van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530–56.
[65] Arbeitman MN, Furlong EE, Imam F, et al. Gene expression during the life cycle of *Drosophila melanogaster*. Science 2002;297(5590):2270–5.
[66] Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1(2):203–9.
[67] Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002;415(6870):436–42.