Published in final edited form as:
Clin Lab Med. 2008 March ; 28(1): 145–viii.

Data Mining in Genomics

Jae K. Lee[a,b], **Paul D. Williams**[a], and **Sooyoung Cheon**[a]

a *Division of Biostatistics and Epidemiology, Department of Public Health Sciences Box 800717, University of Virginia, Charlottesville, Virginia, 22908, USA*

SYNOPSIS

In this paper we review important emerging statistical concepts, data mining techniques, and applications that have been recently developed and used for genomic data analysis. First, we summarize general background and some critical issues in genomic data mining. We then describe a novel concept of statistical significance, so-called false discovery rate, the rate of false positives among all positive findings, which has been suggested to control the error rate of numerous false positives in large screening biological data analysis. In the next section two recent statistical testing methods---significance analysis of microarray (SAM) and local pooled error (LPE) tests are introduced. We next introduce statistical modeling in genomic data analysis such as ANOVA and heterogeneous error modeling (HEM) approaches that have been suggested for analyzing microarray data obtained from multiple experimental and/or biological conditions. The following two sections describe data exploration and discovery tools largely termed as: supervised learning and unsupervised learning. The former approaches include several multivariate statistical methods to investigate co-expression patterns of multiple genes, and the latter approaches are the classification methods to discover genomic biomarker signatures for predicting important subclasses of human diseases. The last section briefly summarizes various genomic data mining approaches in biomedical pathway analysis and patient outcome and/or chemotherapeutic response prediction. Many of the software packages introduced in this paper are freely available at Bioconductor, the open-source Bioinformatics software web site (http://www.bioconductor.org/).

Keywords

ANOVA; False discovery rate; Genomic data; Heterogeneous error model (HEM); Hierarchical clustering; Linear discriminant Analysis; Local pooled error (LPE) test; Logistic Regression discriminant analysis; Microarray GeneChip™ gene expression; Missclassification penalized posterior (MiPP); Significance analysis of microarray (SAM); Supervised learning; Unsupervised learning

---

[b] Corresponding author. Jae K. Lee, Ph.D., Associate Professor of Biostatistics and Epidemiology, Department of Public Health Sciences, Old Medical School, Rm. 3914 (Biostatistics; physical location, 1335 Hospital Drive, Rm. 3181 (mail delivery), Hospital West Complex, University of Virginia School of Medicine, Charlottesville, VA 22908-0717, phone (434) 982-1033, dept (434) 924-8712, fax (434) 243-5787, 924-8437, email jaeklee@virginia.edu.

Co-Authors: Paul D. Williams, Department of Public Health Sciences, Old Medical School, Rm. 3914 (Biostatistics; physical location), 1335 Hospital Drive, Rm. 3181 (mail delivery), Hospital West Complex, University of Virginia School of Medicine, Charlottesville, VA 22908-0717

Sooyoung Cheon, Department of Public Health Sciences, Old Medical School, Rm. 3914 (Biostatistics; physical location), 1335 Hospital Drive, Rm. 3181 (mail delivery), Hospital West Complex, University of Virginia School of Medicine, Charlottesville, VA 22908-0717

Introduction

There has been a great explosion of genomic data in recent years. This is due to the advances in various high-throughput biotechnologies such as RNA gene expression microarrays. These large genomic data sets are information-rich and often contain much more information than the researchers who generated the data may have anticipated. Such an enormous data volume enables new types of analyses, but also makes it difficult to answer research questions using traditional methods. Analysis of these massive genomic data has several unprecedented challenges:

Challenge 1: Multiple comparisons issue

Analysis of high-throughput genomic data requires handling an astronomical number of candidate targets, most of which are false positives [1,2]. For example, a traditional statistical testing criterion of 5% significance level would result in 500 false positive genes, on average, from a 10K microarray study comparing two biological conditions for which there were no real biological differences in gene regulation. If there actually were a small number of, e.g. 100 genes that are differentially regulated, such *real* differentially expressed genes will be mixed with the above 500 false positives without any *a priori* information to discriminate the two groups of genes. Confidence on the 600 targets identified by such a statistical test is low, and further investigation of these candidates will have a poor yield. Simply tightening such a statistical criterion, e.g., a 1% or lower significance level, will result in a high false-negative error rate with failure to identify many important real biological targets. This kind of pitfall, the so-called *multiple comparisons issue* becomes even more serious when one tries to find novel biological mechanisms and biomarker prediction models that involve multiple interacting targets and genes, because the number of candidate pathways or interaction mechanisms grows exponentially. Thus, it is critical that data mining techniques effectively minimize both false positive and false negative error rates in these kinds of genome-wide investigations.

Challenge 2: High dimensional biological data

The second challenge is the high dimensional nature of biological data in many genomic studies [3]. In genomic data analysis, many gene targets are investigated simultaneously, yielding dramatically sparse data points in the corresponding high-dimensional data space. It is well known that mathematical and computational approaches often fail to capture such high dimensional phenomena accurately. For example, many search algorithms cannot freely move between local maxima in a high dimensional space. Furthermore, inference based on the combination of several lower dimensional observations may not provide a correct understanding of the real phenomenon in their joint, high-dimensional space. Consequently, unless appropriate statistical dimension reduction techniques are used to convert high dimensional data problems into lower dimensional ones, important variation and information in the biological data may be obscured.

Challenge 3: Small n and large p problem

The third challenge is the so-called “small n and large p” problem [2]. Desired performance of conventional statistical methods is achieved when the sample size of the data, namely “n”—the number of independent observations and subjects—is much larger than the number of candidate prediction parameters and targets, namely “p”. In many genomic data analyses this situation is often completely reversed. For example, in a microarray study tens of thousands of genes’ expression patterns may become the candidate prediction factors for a biological phenomenon of interest (e.g., response vs. resistance to a chemotherapeutic regimen), but the number of independent observations (e.g., different patients and/or samples) is often a few tens or hundreds at most. Due to the experimental costs and limited availability of biological

materials, the number of independent samples may be even smaller, sometimes only a few. Traditional statistical methods are not designed for these circumstances and often perform very poorly; furthermore it is important to strengthen statistical power by utilizing all sources of information in large-screening genomic data.

Challenge 4: Computational limitation

We also note that no matter how powerful a computer system becomes, it is often prohibitive to solve many genomic data mining problems by exhaustive combinatorial search and comparisons [4]. In fact, many current problems in genomic data analysis have been theoretically proven to be of NP (non-polynomial)-hard complexity, implying that no computational algorithm can search for all possible candidate solutions. Thus, heuristic—most frequently statistical—algorithms that effectively search and investigate a very small portion of all possible solutions are often sought for genomic data mining problems. The success of many bioinformatics studies critically depends on the construction and use of effective and efficient heuristic algorithms, most of which are based on the careful application of probabilistic modeling and statistical inference techniques.

Challenge 5: Noisy high-throughput biological data

The next challenge derives from the fact that high-throughput biotechnical data and large biological databases are inevitably noisy because biological information and signals of interest are often observed with many other random or confounding factors. Furthermore, a one-size-fit-all experimental design for high-throughput biotechniques can introduce bias and error for many candidate targets. Therefore, many investigations in bioinformatics can be successfully performed only when such variability of genomics data are well-understood. In particular, the distributional characteristics of each data set needs to be analyzed using statistical and quality control techniques on initial data sets so that relevant statistical approaches may be applied appropriately. This preprocessing step is critical for all subsequent bioinformatics analyses, and it is sometimes difficult to reconcile dramatically different results that may stem from slightly different preprocessing procedures. While there is no easy answer for such an issue, it is important to employ consistent preprocessing procedures within each and across different analyses, with good documentation of procedures used.

Challenge 6: Integration of multiple, heterogeneous biological data for translational bioinformatics research

The last challenge is the integration of genomic data with heterogeneous biological data and associated metadata, such as gene function, biological subjects' phenotypes, and patient clinical parameters. For example, multiple heterogeneous data sets including gene expression data, biological responses, clinical findings and outcomes data may need to be combined to discover genomic biomarkers and gene networks that are relevant to disease and predictive of clinical outcomes such as cancer progression and chemosensitivity to an anticancer compound. Some of these data sets exist in very different formats and may require combined preprocessing, mapping between data elements, or other preparatory steps prior to correlative analysis, depending on their biological characteristics and data distributions. Effective combination and utilization of the information from such heterogeneous genomic, clinical and other data resources remains a significant challenge.

In this paper we review novel concepts and techniques for tackling various genomic data mining problems. In particular, because DNA microarrays and GeneChips™ techniques have become an important tool in biological and biomedical investigations, we will focus on statistical approaches that have been applied to various microarray data analyses to overcome some of the challenges mentioned above.

A New Concept of Statistical Significance: False Discovery Rate

In order to avoid a large number of false positive findings, the family-wise error rate (FWER) has been classically controlled for the random chance of multiple hypotheses (or candidates) by evaluating the probability that at most one false positive is included at a cutoff level of a test statistic among all candidates. However, FWER has been found to be very conservative in microarray studies, resulting in a high false-negative error rate, often very close to 100% [1]. To avoid such a pitfall, a novel concept of statistical significance, the so-called ***false discovery rate*** (FDR) and its refinement, ***q-value,*** have been suggested [2,5] (qvalue package, www.bioconductor.org). FDR is defined as follows. Suppose there are **M** candidates for simultaneously testing to reject the null hypothesis of no biological significance. Assume $\mathbf{M_0}$ among **M** to be the number of true negative candidates and $\mathbf{M_1}$ ($=\mathbf{M}-\mathbf{M_0}$) to be the number of true positive candidates. At a cutoff value of a test statistic or data mining tool, let R denote the number of all positives (or significantly identified candidates), V the number of false positives, and S the number of false negatives (Table 1).

Then, the FDR is defined as V/R if $R > 0$, the ratio between false positives (V) and all positive findings (R=V+S). Note that FDR is thus derived based both on the null (no significance) and alternative (significant target) distributions. In contrast, the classical p-value (or type I error), here V/M_0, and the statistical power (1 - type II error), or S/M_1, are based only on one of the null and alternative distributions. Therefore, the FDR criterion can simultaneously balance between false positives and false negatives whereas the classical p-value and power can address only one of the two errors.

The FDR evaluation has been rapidly adopted for microarray data analysis, including the widely-used SAM (Significance Analysis of Microarrays) and other approaches [1,6]. Many different methods have been suggested for estimating FDR directly from test statistics, or indirectly from classical p-values of such statistics. The latter methods are convenient since standard p-values can be simply converted into their corresponding FDR values [5,7] and q-value, especially the latter based on a resampling technique. More careful FDR assessment can also be found in many other recent studies [7].

Pairwise Statistical Tests for Genomic Data

Each gene's differential expression pattern in a microarray experiment is usually assessed by (typically pairwise) contrasts of mean expression values among experimental conditions. Such comparisons have been routinely measured as fold changes whereby genes with greater than two or three fold changes are selected for further investigation. It has been found frequently that a gene that shows a high fold-change between comparison conditions might also exhibit high variability in general and hence its differential expression may not be significant. Similarly, a modest change in gene expression may be significant if its differential expression pattern is highly reproducible. A number of authors have pointed out this fundamental flaw in the fold-change based approach [1]. Thus, the emerging standard approach is based on statistical significance and hypothesis testing, with careful attention paid to reliability of variance estimates and multiple comparison issues.

The classical two-sample t-test and other traditional test statistics have been initially used for testing each gene's differential expression [6]. These classical testing procedures, however, rely on reasonable estimates of reproducibility or within-gene error, requiring a large number of replicated arrays. When a small number of replicates are available per condition, e.g., duplicate or triplicate, the use of within-gene estimates of variability does not provide a reliable hypothesis testing framework. For example, a gene may have very similar differential expression values in duplicate experiments by chance alone. Furthermore, the comparison of

means can be misled by outliers with dramatically smaller or larger expression intensities than other replicates. Because of this, error estimates constructed solely within genes may result in underpowered tests for differential expression comparisons and also result in large numbers of false positives. Several approaches to improving estimates of variability and statistical tests of differential expression have thus recently emerged as follows [8–10].

Significance Analysis of Microarrays (SAM)

SAM has been proposed to improve the unstable error estimation in the two-sample t-test by adding a variance stabilization factor which minimizes the variance variability across different intensity ranges [1]. Based on the observation that the signal-to-noise ratio varies with different gene expression intensities, SAM tries to stabilize gene-specific fluctuations. and is defined based on the ratio of change in gene expression to the standard deviation in the data for that gene. The relative difference d(i) in gene expression is defined as:

$$d(i)=(x_I(i) - x_U(i))/(s(i)+s0)$$

where $x_I(i)$ and $x_U(i)$ are the average expression values of gene i in states I and U, respectively. The gene-specific scatter s(i) is the standard pooled deviation of replicated expression values of the gene in the two states. To compare values of d(i) across all genes, the distribution of d (i) is assumed to be independent of the level of gene expression. However, as mentioned above, at low expression levels variability in d(i) can be high because of small values of s(i). To ensure that the variance of d(i) is independent of gene expression, a positive constant s_0 is added to the denominator. The value for s_0 is chosen to minimize the coefficient of variation, where the coefficient of variability of d(i) is computed as a function of s(i) in moving windows across all the genes.

Local Pooled Error (LPE)

Based on a more careful error-pooling technique, the so-called local-pooled-error (LPE) test has also been introduced. This testing technique is particularly useful when the sample size is very small, e.g., two or three per condition. LPE variance estimates for genes are formed by pooling variance estimates for genes with similar expression intensities from replicated arrays within experimental conditions [6]. The LPE approach leverages the observations that genes with similar expression intensity values often show similar array-experimental variability within experimental conditions; and that variance of individual gene expression measurements within experimental conditions typically decreases as a (non-linear) function of intensity. LPE has been introduced specifically for analysis of small-sample microarray data, whereby error variance estimates for genes are formed by pooling variance estimates for genes with similar expression intensities from replicated arrays within experimental conditions (LPE package, www.bioconductor.org). This is possible because common background noise can often be found within each local intensity region of the microarray data. At high levels of expression intensity, this background noise is dominated by the expression intensity, while at low levels the background noise is a larger component of the observed expression intensity, which can be easily observed in the so-called AM log-intensity scatter plot of two replicated chips among three different immune conditions [6] (Figure 1). The LPE approach controls the situation where a gene with low expression may have very low variance by chance and the resulting signal-to-noise ratio is unrealistically large. Statistical significance of the LPE-based test is evaluated as follows. First, each gene's medians m_1 and m_2 under the two compared conditions are calculated to avoid artifacts from outliers. The LPE statistic for the median (log-intensity) difference z is then calculated as:

$$z=(m_1 - m_2)/s_{LPEpooled}$$

where $s_{LPEpooled}$ is the pooled standard error from the LPE-estimated baseline variances from the two conditions. The LPE approach shows a significantly better performance than two-

sample t-test, SAM, and Westfall-Young's permutation tests, especially when the number of replicates is smaller than ten [6].

Statistical Modeling on Genomic Data

Genomic expression profiling studies are also frequently performed for comparing complex, multiple biological conditions and pathways. Several linear modeling approaches have been introduced for analyzing microarray data with multiple conditions. For example, an ANOVA model approach was considered to capture the effects of dye, array, gene, condition, array-gene interaction, and condition-gene interaction separately on cDNA microarray data [11], and a two-stage mixed model was proposed first to model cDNA microarray data with the effects of array, condition, and condition-array interaction and then to fit the residuals with the effects of gene, gene-condition interaction, and gene-array interaction [12]. Several approaches have also been developed under the Bayesian paradigm for analyzing microarray data, including Bayesian parametric modeling [13], Bayesian regularized t-test [8], Bayesian hierarchical modeling with a multivariate normal prior [14], and Bayesian heterogeneous error model (HEM) with two error components [15]. ANOVA and HEM approaches are introduced below.

ANOVA Modeling

The use of analysis of variance (ANOVA) models has been suggested to estimate relative gene expression and to account for other sources of variation in microarray data [16]. Even though the exact form of the ANOVA model depends on the particular data set, a typical ANOVA model for two-color based cDNA microarray data can be defined as

$$y_{ikg}=\mu+A_i+D_j+V_k+G_g+AD_{ij}+AG_{ig}+DG_{ig}+VG_{kg}+\varepsilon_{ijkg},$$

where y_{ijkg} is the measured intensity from array i, dye j, variety k, and gene g on an appropriate scale (typically the log scale). The generic term "variety" is often used to refer to the mRNA samples under study, such as treatment and control samples, cancer and normal cells, or time points of a biological process. The terms A, D, and AD account for the overall effects that are not gene-specific. The gene effects G_g capture the average levels of expression for genes and the array-by-gene interactions AG_{ig} capture differences due to varying sizes of spots on arrays. The dye-by-gene interactions DG_{jg} represent gene-specific dye effects. None of the above effects are of biological interest, but amount to a normalization of the data for ancillary sources of variation. The effects of primary interest are the interactions between genes and varieties, VG_g. These terms capture differences from overall averages that are attributable to the specific combination of variety k and gene g. Differences among these variety-by-gene interactions provide the estimates for the relative expression of gene g in varieties 1 and 2 by $VG_{1g} - VG_{2g}$. Note that AV, DV, and other higher-order interaction terms are typically assumed to be negligible and are considered together with the error terms. The error terms ε_{ijkg}'s are often assumed to be independent and normal with mean zero and a common variance. However, such a global ANOVA model is difficult to implement in practice due to its computational restriction. Instead, one often considers gene-by-gene ANOVA models like:

$$y_{ijkg}=\mu_g+A_i+D_j+V_k+AD_{ij}+VG_{kg}+\varepsilon_{ijkg}.$$

Alternatively, a two-stage ANOVA model may be used [12]. The first layer is for main effects non-specific to the gene effects:

$$y_{ijkg}=\mu+A_i+D_j+V_k+AD_{ij}+AGig+\varepsilon_{ijkg}.$$

Let r_{ijkg} be the residuals from this first ANOVA fit. Then, the second-layer ANOVA model for gene-specific effects is considered as:

$$r_{ijkg}=G_g+AG_{ig}+DG_{ig}+VG_{kg}+\nu_{ijkg}.$$

Excepting the main effects of G and V and their interaction effects, the other terms A, D, (AD), (AG), and (DG) can be considered as random effects. These within-gene ANOVA models can be implemented using most standard statistical packages, such as R (see chapter X), SAS, or SPSS.

Heterogeneous Error Model

Similarly to the statistical tests for comparing two sample conditions, the above within-gene ANOVA modeling methods are underpowered and have inaccurate error estimation in microarray data with limited replication. The heterogeneous error model (HEM) has been suggested as an alternative (HEM package at www.bioconductor.org). It is based on Bayesian hierarchical modeling and LPE error-pooling-based prior constructions, with two layers of error which decompose the total error variability into the technical and biological error components in microarray data [15]. The first layer is constructed to capture the array technical variation due to many experimental error components, such as sample preparation, labeling, hybridization, and image processing:

$$y_{ijkl}=x_{ijk}+\varepsilon_{ijkl}, \text{ where } \varepsilon_{ijkl} \sim \text{iid Normal}[0, \sigma^2(x_{ijk})], \text{ where } i=1,2,\ldots,G;\ j=1,2,\ldots,C;\ k=1,2,\ldots,m_{ij};\ l=1,2,\ldots,n_{ijk}.$$

The second layer is then hierarchically constructed to capture the biological error component:

$$x_{ijk}=\mu+g_i+c_j+r_{ij}+b_{ijk}, \text{ where } b_{ijk} \sim \text{iid Normal}[0, \sigma^2{}_b(ij)].$$

Here, the genetic parameters are for the grand mean (shift or scaling) constant, gene, cell, interaction effects, and the biological error; the last error term varies and is heterogeneous for each combination of different genes and conditions. Note that the biological variability is individually assessed for discovery of biologically-relevant expression patterns. The HEM approach shows a significantly better performance than standard ANOVA methods, especially when the number of replicates is small (Figure 2).

Unsupervised Learning: Clustering

Clustering analysis is widely applied to search for the groups (clusters) in microarray data because these techniques can effectively reduce the high-dimensional gene expression data into a two-dimensional dendrogram organized by each gene's expression association patterns (Figure 3). Currently, clustering analysis is one of the most frequently used techniques for genomic data mining in biomedical studies [17–19]. Some technical aspects of these approaches are summarized below. A clustering approach first needs to be defined by a measure or distance index of similarity or dissimilarity such as:

- Euclidean: $d(x, y) = \Sigma(x_k - y_k)^2$
- Manhattan: $d(x, y) = \Sigma|x_k - y_k|$
- Correlation: $d(x, y) = 1 - r(x, y)$, where $r(x, y)$ is a correlation coefficient.

Next, one needs to define an allocation algorithm based on one of the above distance metrics. Two classes of clustering algorithms have been used in genomic data analysis: hierarchical and partitioning allocation algorithms.

Hierarchical algorithms that allocate each subject to its nearest subject or group include:

- Agglomerative methods: average linkage based on group average distance, single linkage based on minimum nearest distance, and complete linkage based on maximum furthest distance
- Probabilistic methods: Bayes factor, posterior probabilities of subclusters

- Divisive methods: monothetic variable division, polythetic division

Partitioning algorithms that divide the data into a pre-specified number of subsets including:

- Self-organizing map: division into a geometrically preset grid structure of subclusters
- Kmeans: Iterative relocation into a predefined number of subclusters
- Pam (partitioning around medoids): similar to, but more robust than Kmeans clustering
- Clara: division of fixed-size sub-datasets for applications to large data sets
- Fuzzy algorithm: probabilistic fractions of membership rather than deterministic allocations

One of the most difficult aspects of using these clustering analyses is the interpretation of their heuristic, often unstable, clustering results. To overcome this shortcoming, several refined clustering approaches have been suggested. For example, the use of bootstrapping was suggested to evaluate the consistency and confidence of each gene's membership to particular cluster groups [11]. The *gene shaving* approach has been suggested to find the clusters directly relevant to major variance directions of an array data set [3]. Recently, *tight clustering*, a refined bootstrap-based hierarchical clustering is proposed to formally assess and identify the groups of genes that are most tightly clustered with each other [20].

Supervised Learning: Classification

Supervised classification learning on genomic data is often performed to obtain genomic prediction models for different groups of biological subjects and patients, e.g. macrophage cells under different immunologic conditions and different sub-classes of cancer patients such as acute lymphoblastic leukemia (ALL) vs. acute myeloid leukemia (AML). In fact, prediction based on genomic expression signatures has received considerable attention in many challenging classification problems in biomedical research [21,22]. For example, such analyses have been conducted in cancer research as alternative diagnostic techniques to the traditional ones such as classification by the origin of cancer tissues and/or microscopic appearance, which can be problematic for the prediction of many critical human disease subtypes [23]. Several different approaches to microarray classification modeling have been proposed, including gene voting [21], support vector machines (SVMs) [24], Bayesian regression models [22], partial least squares [25], and GA/KNN [26]. The following discussion considers strategies to evaluate and compare the performance of these different classification methods.

Measures for Classification Model Performance

Microarray data often have tens of thousands of genes on each chip whereas only a few tens of samples or replicated arrays are available in a microarray study. In the classification modeling on genomic data, it is thus essential to avoid over-fitting and to find an optimal subset of the thousands of genes for constructing classification rules and models that are robust in different choices of training samples and consistent in prediction performance on future samples. Typically, in this kind of supervised learning, separate training and test sets (of subjects with known classes) are used, the former to fit classification prediction models and the latter, which is independent of the former set, for rigorous model validation. Evaluation of prediction performance should then be carefully conducted among the extremely large number of competing models, especially in using appropriate performance selection criteria and in utilizing the whole data for model training and evaluation. Several different measures are currently used to evaluate performance of classification models: classification error rate, area under the ROC curve (AUC), and the product of posterior classification probabilities [27,28].

When a large number of candidate models, e.g., ~10^8 two-gene models on 10K array data, are compared in their performance, these measures are often saturated—their maximum performance levels are achieved by many competing models—so that identification of the best (most robust) prediction model among them is extremely difficult. Furthermore, these measures cannot capture an important aspect of classification model performance as follows. Suppose three samples are classified using two classification models (or rules)—one model provides the correct posterior classification probabilities 0.8, 0.9, and 0.4, and the other 0.8, 0.8, and 0.4 for the three samples. Assuming these were unbiased estimates of classification error probabilities (on future data), the former model would be preferred because this model will perform better in terms of the expected number of correctly classified samples in future data.

Note that the two models provide the same misclassification error rate 1/3. This aspect of classification performance cannot be captured by evaluating the commonly-used error rate or AUC criteria, which simply add one count for each correctly-classified sample, ignoring its degree of classification error probability.

To overcome this limitation, the so-called Misclassification Penalized Posterior (MiPP) criterion has been suggested recently [4]. This measure is the sum of the correct-classification (posterior) probabilities of correctly-classified samples subtracted by the sum of the misclassification (posterior) probabilities of misclassified samples. Suppose there are m classes π_I, i=1,…,m, from a population of N samples. Let Xij, j=1,…,n_i, be the j-th sample from the i-th class under a particular prediction model (e.g., one gene or two gene model) from a classification rule such as LDA or SVMs. MiPP is then defined as:

$$\Lambda = \sum_{\text{correct}} p_k(X_j) - \sum_{\text{wrong}} (1 - p_k(X_j)),$$

where $p_k(X_j)$ is the posterior classification probability of sample X_j into the k-th class. Here correct and wrong correspond to the samples that are correctly and incorrectly classified. In the two class problem, "correct'" simply means $p_k(X_j) > 0.5$, but in general, it occurs when $p_k(X_j) = \max_i = 1,\ldots,m\ p_i(X_j)$. It can be shown easily that MiPP is the sum of the posterior probabilities of correct classification penalized by the number of misclassified samples (N_M): $\Lambda = \Sigma p_k(X_j) - N_M$. Thus, MiPP is a continuous measure (compared to the discrete error rate) of classification performance that takes into account both the degree of classification certainty and the error rate, and is sensitive enough to distinguish subtle differences in prediction performance among many competing models.

Classification Modeling

Several classification modeling approaches are currently widely used in genomic data analysis:

1. Gene Voting: Gene voting [21] is an intuitively-derived technique that aggregates the weighted votes from all modeling gene signatures; the advantage of this technique is that it can be easily implemented without complicated computing and statistical arguments. It has been proposed for the prediction of subclasses of acute leukemia patients observed by microarray gene expression data [21]. This method gains accuracy by aggregating predictors built from a learning set and by casting their voting weights. For binary classification, each gene casts a vote for class 1 or 2 among p samples, and the votes are aggregated over genes. For gene g_j the vote is $v_j = a_j (g_j - b_j)$, where $a_j = (m_1 - m_2)/(s_1 + s_2)$ and $b_j = (m_1 + m_2)/2$ for sample means m_1 and m_2 and sample standard deviations s_1 and s_2. Using this method based on 50 gene predictors, 36 of 38 patients in an independent validation set have been correctly classified between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

2. LDA and QDA: Linear (or quadratic) discriminant analysis (LDA) is one of the classical statistical classification techniques based on the multivariate normal distribution assumption. It is frequently found that this technique is quite robust and powerful for many different applications despite the distributional assumption; the gene voting technique can be considered as a variant of LDA. LDA can be applied with leave-one-out classification, assuming each class follows a multivariate normal distribution. Each sample will then be allocated to group k to which its classification probability is maximized. The quadratic discriminant analysis can be similarly performed except that the covariance matrix of the multivariate normal distribution (for each of m classes) is now considered differently among m classes. Differences between LDA and QDA are typically small, especially if polynomial factors are considered in LDA. In general, QDA requires more observations to estimate each variance-covariance matrix for each class. LDA and QDA have consistently shown high performance not because the data is likely derived from Gaussian distributions, but more likely because the data support only a simple boundaries such as linear or quadratic [28].

3. Logistic Regression (LR): *The* logistic regression classification technique is based on the regression fit on probabilistic odds among comparing conditions. This technique requires no specific distribution assumption but is often found to be less sensitive than other approaches. LR methods simply maximize the conditional likelihood Pr(G=k|X), typically by a Newton-Raphson algorithm [29]. The allocation decision on a sample is based on the logit regression fit:

$$\text{Logit}(p_i)=\log(p_i/(1-p_i)) \sim \beta^T x,$$

where β is the LR estimated coefficient vector for the microarray data. LR discriminate analysis is often used due to its flexible assumption about the underlying distribution, but if it is actually from Gaussian distribution, LR shows a loss of 30% efficiency in the (misclassification) error rate compared to LDA.

4. Support Vector Machines (SVMs): Conceptually similar to gene voting, support vector machine is one of the recent machine-learning classification techniques based on the data projection to high dimensional kernel space. This technique does not require distributional assumption either, yet can perform better than other approaches in some complicated cases. However, it often needs large numbers of samples and predictor gene signatures for its optimal performance. SVMs separate a given set of binary labeled training data with a hyper-plane that is maximally distant from them, known as the maximal margin hyper-plane [24]. Base on a kernel, such as a polynomial of dot products, the current data space will be embedded in a higher dimensional space. The commonly-used kernels are:

    - Radial basis kernel: $K(x,y) = \exp(-|x-y|^2/2s^2)$
    - Polynomial kernel: $K(x,y) = <x,y>\text{^d}$ or $K(x,y) = (<x,y> + c)^d$, where <,> denotes the inner product.

Comparison of Classification Methods

The above classification techniques must be carefully applied in prediction model training on genomic data. In particular, if all the samples are used both for model search/training and for model evaluation in a large screening search for classification models, a serious selection bias is inevitably introduced [30]. To avoid such a pitfall, a stepwise (leave-one-out) cross-validated discriminant procedure that gradually adds genes to the training set has been suggested [4, 28]. It is typically found that the prediction performance is continuously improved (or not decreased) by adding more features into the model. This is again due to a sequential search

and selection strategy against an astronomically large number of candidate models; some of them can show over-optimistic prediction performance for a particular training set by chance. Note also that even though a leave-one-out or similar cross-validation strategy is used in this search, the number of candidate models is too big to eliminate many random ones that survive by chance from cross-validation. Thus, test data should be completely independent from the training data to obtain an unbiased estimate of each model's performance. To address these pitfalls, the stepwise cross-validated discriminant (SCVD) procedure sequentially adds one gene at a time to identify the most optimal prediction model based both on n-fold modeling and train-test validation strategies. SCVD can be used with any of the aforementioned classification methods.

*L*inear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression (LR), and support vector machines (SVMs) with linear or RBF kernels have been compared using the SCVD approach [28]. The leukemia microarray data in [21] had a training set of 27 ALL and 11 AML samples and an independent test set of 20 ALL and 14 AML samples. Since two distinct data sets exist, the model is constructed on the training data and evaluated on the test data set. Each rule identified a somewhat different subset of features that showed the best performance within each classification method (Table 2). In terms of error rate, it appears as if the SVM with a linear kernel is the most accurate rule. However, LDA only misclassified one sample and the SVM with the RBF kernel and QDA misclassified two samples on the independent test data. Logistic regression does not seem to perform as well as the other rules, misclassifying 4 out of 34 samples. Note again that comparing the rules on the basis of MiPP is somewhat tricky for SVMs since the estimated probabilities of correct classification from SVMs are based upon how far samples are from a decision boundary. As a result, these are not true probabilities as is the case with LDA, QDA, and LR. In an application to a different microarray study on colon cancer, the RBF-kernel SVM model with three genes was found to perform best among these classification techniques.

The MiPP-based SCVD procedure was the most robust classification model and could accurately classify samples with a very small number of features—only two or three genes for the two well-known microarray data sets, outperforming many previous models with 50–100 features—though different classification methods may perform differently in different data sets. This data is consistent with the notion that many correlated genes share more or less similar information and may discriminate similarly between different subtypes of a particular disease, and that multiple small-feature models may perform well as far as the construction of a classification model is concerned. As shown, the prediction performance on the training set is quickly saturated with 0% error rate and very close to the maximum MiPP value 38 (total sample size). On the contrary, both error rates and MiPP values greatly vary on the independent test set. Also, the error rates were found to be misleading and less informative than MiPP.

Genomic Pathway modeling

Many recent pathway modeling studies for transcriptional regulation and gene functional networks have been performed based on genomic expression data. These approaches can be largely divided into three categories—qualitative, quantitative, and integrative pathway modeling—based on the different types of genomic data used. We briefly introduce several pathway modeling approaches here with references. Note that pathway modeling is one of the most active research fields in current genomic sciences and substantial additional information can be found in the references.

Qualitative Pathway modeling

Pathway modeling has been carried out using functional and annotation information from several genomic databases. For example, computationally predicting genome-wide

transcription units (TU) based on pathway-genome databases (PGDBs) and other organizational (i.e. protein complexes) annotation improved TU organization information in Escherichia coli and Bacillus subtilis [31]. A classification of transcription factors is also proposed to organize pathways that connect extracellular signaling to the regulation of transcription and constitutively activate nuclear factors in eukaryotic cells. This latter classification was performed on the basis of known cellular characteristics that describe the roles of these factors within regulatory circuits in order to identify many down-stream functional mechanisms such as serine and tyrosine phosphorylation, Rel/NFkB family, CI (GLI), Wnt and Notch pathway, NFAT activation, and Ca2+ increase [32].

Quantitative Pathway modeling

Gene regulation networks have also been explored based on quantitative genomic expression data. For example, Bayesian network modeling was used for capturing regulatory interactions between genes based on genome-wide expression measurements on yeast [33]. Probabilistic models for context-specific regulatory relationships were also proposed to capture complex expression patterns of many genes in various biological conditions, accounting for known variable factors such as experimental settings, putative binding sites, or functional information in yeast stress data and compendium data [34]. These quantitative pathway models have been found to effectively characterize both relationships and magnitude of relevant genes' expression patterns, and have been extensively used in recent pathway modeling in various microarray studies [33–35].

Integrative Pathway Modeling

Integration of qualitative and quantitative gene network information has been attempted in recent pathway modeling studies. For example, a comprehensive genomic module map was constructed by combining gene expression and known functional & transcriptional information, where each module represents a distinctive set of directly regulated and associated genes that act in concert to carry out a specific function. In this study, different expression activities in tumors were described in terms of the behavior of these modules [35]. Regression on transcription motifs is proposed for discovering candidate genes' upstream sequences that undergo expression changes in various biological conditions. This method combines the known motif structural information and gene expression patterns based on an integrated regression analysis [36].

Genomic Biomarkers for Disease Progression and Chemosensitivity

Recently, genomic data have been used to predict cancer patients' outcome and tumor chemosensitivity. The prognosis of a cancer patient is frequently uncertain, and histologically-based prognostic indicators are sometimes inaccurate due to the inherent complexities involved. Deregulation of tumor cells leads to uncontrolled division, invasion, and metastasis; the specific patterns of deregulation in a particular tumor and which genetic pathways are altered likely affect the course of the disease. Based on such observations, genomic predictors have been developed for the progression of metastatic disease following removal of breast cancer tumors [37]. In another study the genomic predictors were generated using tumor samples from 78 patients, of whom 34 developed metastasis within five years of surgical resection [38]. The first step in the development of the predictor was to determine, using microarray analysis, which genes were significantly differentially expressed compared to a pool of all specimens. 4,968 out of 24,479 genes on the chip were found to have at least a two-fold change in expression and a p-value less than 0.01 in at least five tumors. To determine which of these genes could be used to predict metastasis, the correlation coefficients between gene expression values and metastasis development were calculated, and 231 genes were found with an absolute correlation coefficient greater than 0.3. Leave-one-out cross-validation on

sequentially larger subsets of genes determined that a group of 70 genes yielded the least erroneous classifier. Testing the classifier on a 19-patient dataset resulted in 2 incorrect predictions, and this was later expanded to predict survival times for breast cancer patients, accurately predicting disease-free long-term survivors against patients with poor prognosis [39].

A combination of in vitro assays and genomic data was used to predict the prognosis of non-small cell lung cancer patients based on Bayesian computational classification techniques [40]. In this study microarray profiling analysis was performed to determine gene expression profiles of the various tumors, using Affymetrix HG-U133 plus 2.0 chips. First filtering found about 2070 genes highly correlated to lung cancer recurrence. The k-means method was used to generate gene clusters, which were then analyzed using singular value decomposition to generate a metagene, the dominant average expression pattern of the cluster. These metagenes were used in a binary regression tree to partition tumor samples into subsets on which predictions of recurrence could be made; the prediction accuracy for 5 yr disease-free survival was over >90% among the the predicted long-term survivors compared to <40% among the predicted poor-prognosis patients.

Concluding Remarks

We believe that there are several reasons why these genomic data mining approaches have been successful and represent a promising direction for future work. First, it has been found that gene networks inferred from genomic expression signatures are highly relevant to patients' prognosis and chemotherapeutic responses [41–43]. Even though many individual genes' expression values in such networks are often variable and noisy, the whole gene networks have been found to be quite consistent in their overall expression patterns [44–46]. Second, genome-wide RNA expression profiling techniques such as microarrays and GeneChips™ have been dramatically improved in recent years, so that the expression patterns of the entire human genome can now be accurately and cost-effectively measured on patient samples. In fact, microarray RNA profiling is one of the most accurately quantifiable and comprehensive profiling biotechnologies among all current high-throughput biotechniques, including CGH, SKY, SAGE, 2D-gel, mass spectrometry, or protein arrays [47,48]. Third, as summarized in this paper, bioinformatics analysis methods and techniques for these microarray data have been significantly improved by us and others, especially in testing (SAM, LPE, SAM, false discovery rate, etc.), clustering (hierarchical, SOM, K-means, response projected clustering, etc.), classification (LDA, SVMs, logistic regression, random forest), and pathway analysis (GenMaPP, Ingenuity Pathway Analysis) for investigating the complex and extensive information in massive genomic data sets effectively and efficiently [1,4,24]. Finally, most importantly, based on the significant efforts by the NIH (Gene Expression Omnibus, GEO) and the EBI (ArrayExpress), many precious microarray data sets of cancer—both cell lines and patients—have been archived for public access. For example, GEO currently has archived over 5,550 microarray data sets on >150K different biomedical samples and human patients with >1,500 sets for cancer alone. Furthermore, despite their technical differences, microarray data sets from different time points, different laboratories, and even different platforms contain quite consistent information for many genes' expression patterns, so that we can successfully perform our investigations across those different genomic data sets. This large and rapidly increasing compendium of data demands data mining approaches and ensures that genomic data mining will continue to be a necessary and highly productive field for the foreseeable future.

Acknowledgements

This study is supported by NIH grant 1R01HL081690 of JKL.

References

1. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98(9):5116–21. [PubMed: 11309499]
2. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100(16):9440–5. [PubMed: 12883005]
3. Hastie T, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol 2000;1(2):RESEARCH0003. [PubMed: 11178228]
4. Soukup M, Cho H, Lee JK. Robust classification modeling on microarray data using misclassification penalized posterior. Bioinformatics 2005;21(Suppl 1):i423–i430. [PubMed: 15961487]
5. Benjamini Y, et al. Controlling the false discovery rate in behavior genetics research. Behav Brain Res 2001;125(1–2):279–84. [PubMed: 11682119]
6. Jain N, et al. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. Bioinformatics 2003;19(15):1945–51. [PubMed: 14555628]
7. Jain N, et al. Rank-invariant resampling based estimation of false discovery rate for analysis of small sample microarray data. BMC Bioinformatics 2005;6:187. [PubMed: 16042779]
8. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 2001;17(6):509–19. [PubMed: 11395427]
9. Efron B, Tibshirani R. Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol 2002;23(1):70–86. [PubMed: 12112249]
10. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol 2000;7(6):819–37. [PubMed: 11382364]
11. Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A 2001;98(16):8961–5. [PubMed: 11470909]
12. Wolfinger RD, et al. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 2001;8(6):625–37. [PubMed: 11747616]
13. Newton MA, et al. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 2001;8(1):37–52. [PubMed: 11339905]
14. Ibrahim, JGaCM-H.; Gray, RJ. Bayesian Models for Gene Expression with DNA Microarray Data. Journal of American Statistical Association 2002;97:88–99.
15. Cho H, Lee JK. Bayesian hierarchical error model for analysis of gene expression data. Bioinformatics 2004;20(13):2016–25. [PubMed: 15044230]
16. Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet Res 2001;77(2):123–8. [PubMed: 11355567]
17. Lee JK, et al. Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. Genome Biol 2003;4(12):R82. [PubMed: 14659019]
18. Scherf U, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24(3):236–44. [PubMed: 10700175]
19. Weinstein JN, et al. The bioinformatics of microarray gene expression profiling. Cytometry 2002;47 (1):46–9. [PubMed: 11774349]
20. Tseng GC, Wong WH. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. Biometrics 2005;61(1):10–6. [PubMed: 15737073]
21. Golub TR, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7. [PubMed: 10521349]
22. West M, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A 2001;98(20):11462–7. [PubMed: 11562467]
23. Su AI, et al. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 2001;61(20):7388–93. [PubMed: 11606367]
24. Furey TS, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16(10):906–14. [PubMed: 11120680]

25. Nguyen DV, Rocke DM. Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics 2002;18(12):1625–32. [PubMed: 12490447]
26. Li L, et al. Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. Comb Chem High Throughput Screen 2001;4(8):727–39. [PubMed: 11894805]
27. Hand, DJ. Construction and Assessment of Classification Rules. Chichester: John Wiley & Sons; 1997.
28. Soukup M, Lee JK. Developing optimal prediction models for cancer classification using gene expression data. J Bioinform Comput Biol 2004;1(4):681–94. [PubMed: 15290759]
29. Pampel, FC. Sage University Papers Series on Quantitative Applications of the Social Sciences. 2000. Logistic Regression: A Primer.
30. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 2002;99(10):6562–6. [PubMed: 11983868]
31. Romero PR, Karp PD. Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. Bioinformatics 2004;20(5):709–17. [PubMed: 14751985]
32. Brivanlou AH, Darnell JE Jr. Signal transduction and the control of gene expression. Science 2002;295 (5556):813–8. [PubMed: 11823631]
33. Friedman N, et al. Using Bayesian networks to analyze expression data. J Comput Biol 2000;7(3–4): 601–20. [PubMed: 11108481]
34. Segal E, et al. Rich probabilistic models for gene expression. Bioinformatics 2001;17(Suppl 1):S243–52. [PubMed: 11473015]
35. Segal E, et al. A module map showing conditional activity of expression modules in cancer. Nat Genet 2004;36(10):1090–8. [PubMed: 15448693]
36. Conlon EM, et al. Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci U S A 2003;100(6):3339–44. [PubMed: 12626739]
37. van't Veer LJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530–6. [PubMed: 11823860]
38. van't Veer LJ, et al. Expression profiling predicts outcome in breast cancer. Breast Cancer Res 2003;5 (1):57–8. [PubMed: 12559048]
39. Dressman HK, et al. Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy. Clin Cancer Res 2006;12(3 Pt 1):819–26. [PubMed: 16467094]
40. Potti A, et al. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. N Engl J Med 2006;355(6):570–80. [PubMed: 16899777]
41. Miller LD, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 2005;102(38):13550–5. [PubMed: 16141321]
42. Havaleshko DM, et al. Prediction of drug combination chemosensitivity in human bladder cancer. Mol Cancer Ther 2007;6(2):578–86. [PubMed: 17308055]
43. Paik S, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351(27):2817–26. [PubMed: 15591335]
44. Horvath S, et al. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci U S A 2006;103(46):17402–7. [PubMed: 17090670]
45. Bild AH, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439(7074):353–7. [PubMed: 16273092]
46. Potti A, et al. Genomic signatures to guide the use of chemotherapeutics. Nat Med 2006;12(11):1294–300. [PubMed: 17057710]
47. Ma XJ, et al. Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Arch Pathol Lab Med 2006;130(4):465–73. [PubMed: 16594740]
48. Puskas LG, et al. Gene profiling identifies genes specific for well-differentiated epithelial thyroid tumors. Cell Mol Biol (Noisy-le-grand) 2005;51(2):177–86. [PubMed: 16171553]

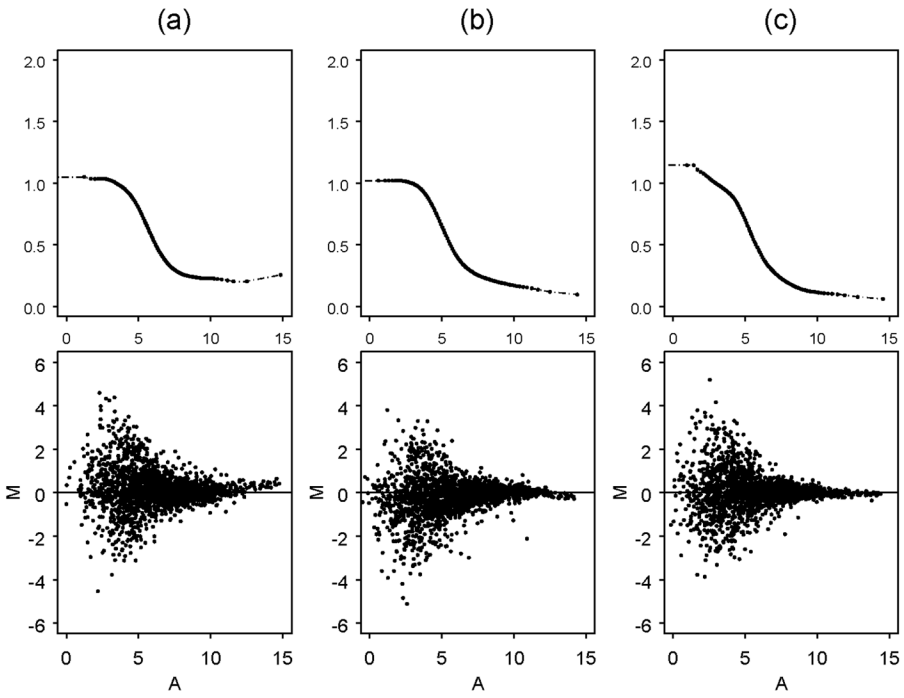

Figure 1.
Log intensity ratio (M) as a function of average gene expression between replicated chips (A). Top panels represent the estimated error distributions (based on a non-parametric regression) for (a) naive, (b) 48 hour activated, and (c) T-cell clone D4 conditions in the mouse immune response microarray study.

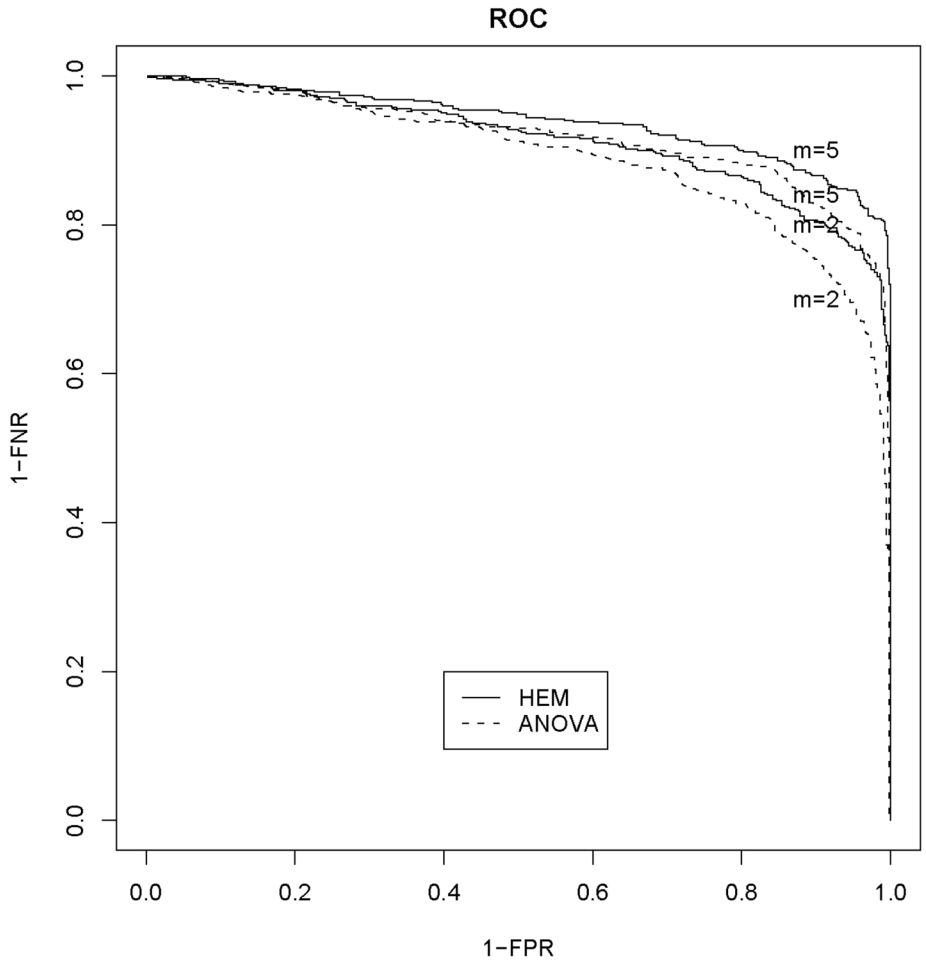

Figure 2.
ROC curves from HEM (solid lines) and ANOVA (dotted lines) models with two and five replicated arrays; The horizontal axis is 1 - FPR (false positive error rate) and the vertical axis is 1 – FNR (false negative error rate).

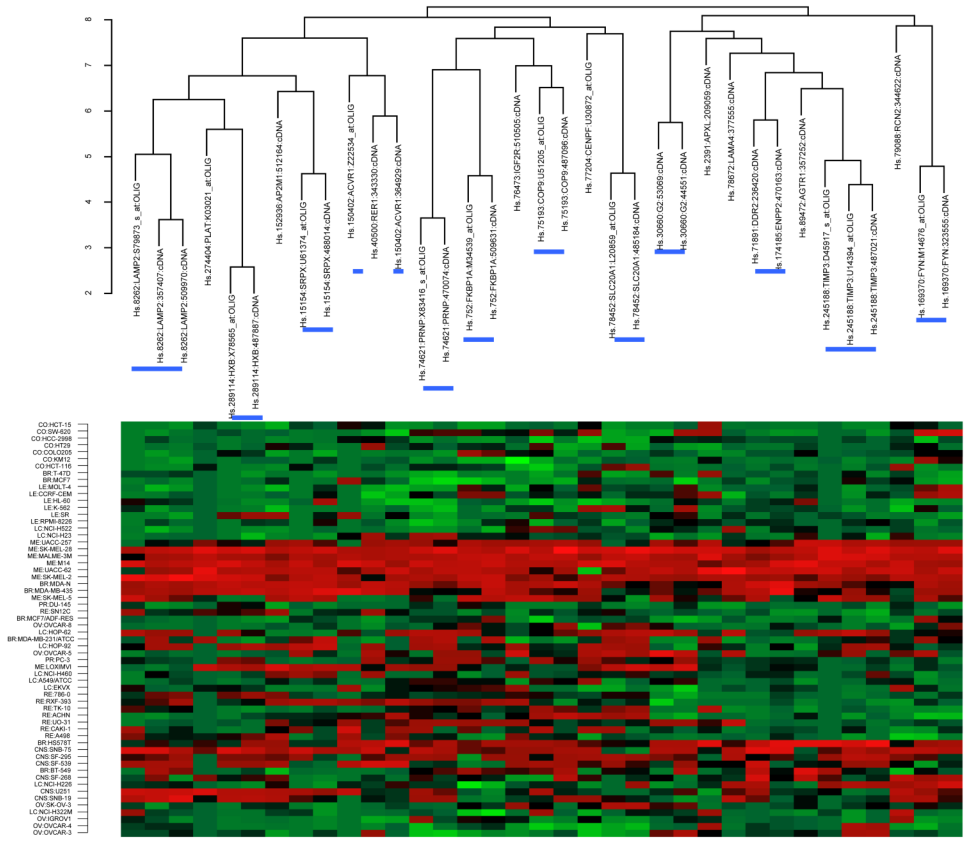

Figure 3.

Dendrogram (top panel) and heatmap (bottom panel) of hierarchical clustering analysis for the concordant cDNA and oligo array expression patterns on the NCI-60 cancer cell lines. A region of heatmap occupied by melanoma genes are shown from the combined set of 3,297 oligo and cDNA transcripts. Each gene expression pattern is designated as coming from the cDNA or oligo array set: The concordant oligo and cDNA microarray expressions are marked with blue bars.

Table 1

Classification of the candidate hypotheses: true negative (U), false positive (V), false negative (T), true positive (S).

	Null Accept	Null Reject	Total
Null true	**U**	**V**	M_0
Alternative true	**T**	**S**	M_1
Total	W	R	M

Table 2

Classification results of the classification rules and the corresponding gene model.

Method	Gene model	Error rate on training data	MiPP on training data	Error rate on test data	MiPP on test data
LDA	1882+1144	0%	37.91	2.9%	**31.46**
QDA	4847+5062	0%	37.96	5.8%	**29.81**
Logistic	1807+4211+575	0%	37.99	11.8%	**25.64**
SVM K=linear	2020+4377+1882	0%	35.16	0%	**29.26**
SVM K=RBF	**4847+3867+6281**	**0%**	**32.52**	**5.9%**	**21.71**