On integrating multi-experiment microarray data

Georgia Tsiliki, Dimitrios Vlachakis and Sophia Kossida

Bioinformatics and Medical Informatics Group, Biomedical Research Foundation, Academy of Athens, 4 Soranou Ephessiou 115 27, Greece

With the extensive use of microarray technology as a potential prognostic and diagnostic tool, the comparison and reproducibility of results obtained from the use of different platforms is of interest. The integration of those datasets can yield more informative results corresponding to numerous datasets and microarray platforms. We developed a novel integration technique for microarray gene-expression data derived by different studies for the purpose of a two-way Bayesian partition modelling which estimates co-expression profiles under subsets of genes and between biological samples or experimental conditions. The suggested methodology transforms disparate gene-expression data on a common probability scale to obtain inter-study-validated gene signatures. We evaluated the performance of our model using artificial data. Finally, we applied our model to six publicly available cancer gene-expression datasets and compared our results with well-known integrative microarray data methods. Our study shows that the suggested framework can relieve the limited sample size problem while reporting high accuracies by integrating multi-experiment data.

1. Introduction

The growth of high-throughput technologies, such as microarrays and next generation sequencing (NGS), has been accompanied by active research in data analysis, producing new analysis tools and statistical methodologies. Microarray technologies, in particular, have matured rapidly over the past few years and present major challenges in managing, sharing, analysing and re-using the large amount of data generated [1] despite
the considerable international efforts in standardizing data format, content and description [2,3]. The majority of high-throughput experiments become available upon publication via public repositories such as Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) [4] and ArrayExpress (http://www.ebi.ac.uk/arrayexpress) [5]. Additionally, specialized, mainly in cancer genome, databases have been collecting high-throughput data from the same biological samples. For instance, The Cancer Genome Atlas (TCGA) project [6] provides a platform to download and analyse datasets from 20 different types of cancer.

Applications which link bioinformatic tools and databases have emerged, showing the way to easily visualize and analyse biomedical data. For instance, BioMart software [7] and its cancer specialized version IntOGen are repositories which store readily combined datasets and provide platforms to easily visualize data. The Galaxy Project (https://main.g2.bx.psu.edu) offers a web-based platform allowing researchers to perform and share analyses. Similarly, the GenePattern platform provides access to more than 180 tools for genomic analysis, to enable reproducible in silico research (http://www.broadinstitute.org/cancer/software/genepattern/).

An increasingly interesting research topic is how to combine datasets or results from several studies, given the similarity of experimental protocols, in order to readily increase the sample size of the data, and for that increase the reliability and efficiency of biological investigations. When considering multiple studies, the technological, laboratory-based and biological variability of microarray data should be accounted for. Two major difficulties are that the expression measurements can be absolute or relative values, depending on the technology, and also the challenges associated with cross-referencing measurements made between different technologies [6,8]. Despite these difficulties, the results of integrated analysis clearly demonstrate the potential for increased statistical power and novel discoveries [9].

Although integration methodologies for gene-expression data have the potential to strengthen or extend the results obtained from the individual studies in a comparatively inexpensive manner, there is little consensus as to how to perform such data transformation effectively. For example, a simple strategy would be to first apply the platform-specific normalization methods, then standardize the biological samples and median centre all gene values; however other approaches can be considered when combining datasets from different array platforms. Most prior work in large-scale microarray integration has been performed in one of two contexts, namely statistical meta-analyses [10,11], which treat each gene in each study independently, and transformation methodologies [12–15], which transform gene-expression measurements from different studies into a common scale before the unification of the data; alternatively, there are studies which combine the two methodologies [16]. According to the latter, integration is performed between differentially expressed genes from the available studies, assuming, for instance, that the parameters which capture the relationship between genes and phenotypes are related across studies [9,16]. Under this perspective, the cross-study integration is tailored to the problem of interest, and potentially relies on fewer assumptions than direct data integration.

Recently, efforts to combine NGS and microarray data have been conducted in order to overcome microarrays’ weakness of restricting the expression profiling data to specific annotations and content. NGS promises to eliminate this weakness while increasing sensitivity, specificity and accuracy [17]. Until now, many comparisons of the two technologies have been performed, demonstrating the good concordance between them [18–20], whereas some computational approaches are available for their data integration, discovering novel patterns in the data [21–23].

We propose a new framework based on a two-way Bayesian partitioning model, or else biclustering, that simultaneously clusters rows and columns of data matrix to address the challenging task of integrating gene-expression microarray data. The suggested framework overcomes the weakness of conventional meta-analysis by assuming inherently different normal distributions for different groups of genes expressed under various conditions. Furthermore, we do not consider phenotypic or any other functional information, to make use of the full data. A Gaussian model based on the composite likelihood of the data is used to infer patterns in the data under a fully Bayesian framework. Unlike other studies briefly mentioned above, which
used clustering algorithms such as k-means and mixture models that require the specification of the number of clusters, our methodology applies a Markov chain Monte Carlo (MCMC) search to sample from the posterior distribution of the partition model parameters [24]. Based on the useful empirical assumption that the genes are expressed via a binary activation pattern across biological samples, i.e. they are either regulated or non-regulated in each biological sample [15], our model assumes that in each dataset there exist a group of significantly low gene-expression values. We finally report the posterior probability of each gene value being derived by a normal distribution with zero mean value, marginalized over 100 random bootstrap samples of the data. This is a common probabilistic scale which unifies disparate gene-expression data, and acts as a summary measure of the significance of each gene.

The suggested methodology can be directly applied to RNA-sequencing data, the alternative to microarrays; however this involves an important truncation of the data in terms of gene entries since we are trimming datasets to only include the common genes among them. An interesting application would be to include complementary views of the genome and in that way employ the ability of RNA sequencing to quantify extreme transcription levels, although that is beyond the scope of this work.

2. Methods

We developed a Bayesian two-way partition model to estimate the optimal partition of multi-experiment data given the parameters of the underlying distributions for each cluster. Similar to the model suggested by Wang & Chen [15], we made the useful empirical assumption that the genes are expressed via a binary activation pattern across biological samples, i.e. they are either regulated or non-regulated in each biological sample. We assume that non-regulated genes may contain expression values that are normally distributed around zero with a small variance, and that regulated genes are normally distributed with a cluster-specific mean value and variance parameters. Finally, we report the posterior probability of each value being derived by a normal distribution with zero mean value given the data partition, a measure which allows us to unify datasets on a probability scale.

(a) Bayesian partition modelling

Let us consider a data matrix $X = \{x_{ij}^s\}_{p \times n \times S}$, where $i$ and $j$ index $p$ genes and $n$ biological samples, respectively, and $s = 1, \ldots, S$ index the number of different microarray studies considered with $X_i^s = (x_{i1}^s, \ldots, x_{in}^s)$. In the context of gene-expression microarray analysis, $x_{ij}^s$ represents the logarithm of the expression level of gene $i$ measured in biological sample $j$ and study $s$. Only the common genes across the $S$ studies are considered. Given $X$, we seek for an unordered collection of non-empty groups of genes which exhibit similar patterns across specific groups of biological samples or vice versa, while the groups themselves are different. Specifically, observations belonging to the same subset are assumed to be exchangeable, and observations from different subsets are assumed to be independent; we refer to these groups as biclusters. Thus, biclustering searches directly for submatrices of the data matrix $X$, i.e. biclusters, whose entries meet a predefined criterion.

We assume that all genes belonging to a bicluster follow a normal distribution, where $\phi(x|\mu, \sigma^2)$ denotes its probability density function with mean $\mu$ and variance $\sigma^2$. Furthermore, we assume that there always exists a bicluster consisting of low gene-expression values, i.e. derived from a normal distribution with a zero mean value. Those gene-expression values could belong to non-regulated genes given the phenotype of interest (e.g. normal or cancerous biological samples). We consider that they form a single bicluster, which we denote by $C_0$, and follow a normal distribution with zero mean value, $C_0 = \{x_{ij}; x_{ij}^s \sim N(0, \sigma_0^2)\}$. The data matrix $X$ consists of $K + 1$ clusters, including $C_0$. Each of the $K$ remaining clusters, say the $k$th, consists of gene-expression values that follow normal distributions with bicluster-specific mean and variance parameters, $x_{ij}^s \sim N(\mu_k, \sigma_k^2)$. 


In order to define the number and boundaries of the inferred biclusters, we introduce a binary indicator matrix $R = \{r_{ij}^s \in \{0, 1\}, \ i = 1, \ldots, p, \ j = 1, \ldots, n, \ s = 1, \ldots, S\}$ to specify the inclusion of the $(i, j)$th gene-expression value of the $s$ study to each of the $K + 1$ estimated biclusters [25,26]. The gene-expression value $x_{ij}^s$ is assigned to a bicluster when $r_{ij}^s = 1$. Alterations between zero and one values indicate the biclusters’ borders. We assume that a bicluster should at least consist of five values. As can be seen in figure 1, the binary pattern creates different biclusters, in this case five biclusters are formed (step (b)), and then one is selected at random with equal probability to be the $C_0$ bicluster (step (c)). A cluster is defined by a group of $r_{ij}^s = 1$ values surrounded by $r_{ij}^s = 0$, or vice versa. In that way $(K + 1)$ biclusters are formed. Note that each bicluster yields two marginal clusters for rows (genes) and columns (biological samples), and thus the $r_{ij}^s$ indicators are designed to identify a set of biological samples that have a common pattern of responses across a set of genes [27].

Given $R, K$, the composite likelihood can be expressed as

$$P(X|R, K, \Theta) = \prod_{s=1}^{S} \prod_{i=1}^{p} \prod_{j=1}^{n} \phi(x_{ij}^s | 0, \sigma_{0}^2) \prod_{k=1}^{K} \phi(x_{ij}^s | \mu_k, \sigma_{k}^2), \quad (2.1)$$

where $\Theta = (\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2, \sigma_0^2)$. The mean values $\{\mu_k\}$ and the bicluster membership indicators $\{r_{ij}\}$ are designed to identify a set of biological samples that have a common pattern of responses across a set of genes. We adopt standard conjugate priors for the partition model parameters, i.e.

$$P(\sigma_k^2) \sim \text{InverseGamma}(\alpha_k, \beta_k)$$
$$P(\sigma_0^2) \sim \text{InverseGamma}(\alpha_0, \beta_0)$$
$$P(\mu_k | \sigma_k^2) \sim \text{N}(\tau_k, \sigma_k^2)$$
$$P(r_{ij}) \sim \text{Bernoulli}(\pi_{rij})$$

The prior parameters $\alpha_k, \beta_k, \alpha_0, \beta_0, \tau_k, \sigma_k^2$ are specified based on the biological samples’ means and variances. We set Bernoulli’s distribution probability of success equal to $\pi_{rij} = 1/2$. Additionally, we set a uniform prior probability distribution to select which of the $(K + 1)$
biclusters will be assigned as \(C_0\) with \(P(C_0) = 1/(K + 1)\), also imposing the threshold of each bicluster having at least five gene-expression values. Then, the joint posterior distribution is

\[
P(R, K, \Theta | X) \propto P(X|R, K, \Theta) P(\sigma^2_0) P(c_0) P(\mu_k | \sigma_k^2) P(\sigma_k^2).
\]

The objective function in this case is the marginal posterior distribution \(P(R, K | X)\), which is derived by placing priors on all the parameters in the model and marginalizing over all parameters except the indicator variable \(R\) and the number of clusters \((K + 1)\), which are the parameters of interest. Note that, unlike mixture modelling methodologies, our model contains the parameters \(R\) and \(K\) which are directly relevant to the basic partitioning problem.

(b) Computational strategy

We use the MCMC sampler [24] to sample \(R\) and \(K\) from the posterior distribution. Our algorithm cycles through all rows and columns sequentially, flips the indicator variables of each row or column, and then decides whether to accept the change based on the acceptance rate. The detailed procedure of our algorithm is as follows.

1. Initialize the partition structure:
   (a) randomly assign binary indicators \(r^s_{ij}\) values for each \((i, j)\) gene-expression value to be either one or zero, which sets the initial number of \((K + 1)\) biclusters,
   (b) randomly select one of the \((K + 1)\) biclusters to be \(C_0\) bicluster.
   
   Alterations of one–zero values indicate the borders and length of each bicluster, while we impose the threshold of each bicluster having at least five elements. A bicluster is defined by a group of identical \(r_{ij}\) values encompassed by \(1 - r_{ij}\) values.

2. Cycle through all rows and columns sequentially. At each cycle propose which \(r^s_{ij}\) values to flip with probability \(p_m \in (0, 1)\).

3. Propose merging two biclusters with the same \(r^s_{ij}\) indicator values with probability \(p_{sm} \in (0, 1)\), and split a randomly selected cluster with probability \(1 - p_{sm}\). The proposed change is then accepted with probability equal to the Bayes factor in favour of the proposed model over the current one.

4. Repeat the cycle until convergence.

(c) Resampling

The above partition model is repeated independently for 100 different bootstrap samples of individuals per study \(s\). Particularly, for each study \(s\), and for all the common genes \(p\) in the \(S\) studies, we consider 100 bootstrap samples consisting of 1000 individuals randomly selected with replacement from \(n\). Finally, for each of the formed data matrices and for each \(x^s_{ij}\) value, we compute the posterior probability \(p_{ij}\) of this value belonging to \(C_0\) bicluster given \(R\) and \(K\) parameters, i.e. we test model \(M_0: x^s_{ij} \in C_0\) against \(M_1: x^s_{ij} \notin C_0\)

\[
p_{ij} = \frac{p(M_0 | R)p(M_0)}{p(M_0 | R)p(M_0) + p(M_1 | R)p(M_1)} = \frac{\text{number of times } x^s_{ij} \text{ value belongs to } C_0 \text{ over all bootstrap samples}}{\text{number of all bootstrap samples}},
\]

which is simplified since we consider \(p(M_0) = p(M_1)\). Thus, each gene-expression value is now transformed to a posterior probability value which depicts how likely it is for this value to belong to \(C_0\) bicluster, given the data. The bootstrap resampling is used in order to improve the accuracy with which posterior probabilities are estimated. It is suggested here as a form of simple model averaging over the different partitions of the available genes. Finally, we
merge the gene-expression matrices to form a unified matrix which accounts for any gene dependencies. The methodology was implemented in C++ using gsl library (http://www.gnu.org/software/gsl/). The source code can be downloaded at http://www.bioacademy.gr/bioinformatics/Xplatform/.

3. Results

The suggested framework enables the integration of gene-expression results performed under diverse microarray study conditions, therefore relieving concerns from the intractable ‘large p, small n’ problem. The validation results presented below intend to assess our framework, and in particular the percentage of genes shared in common between individual and integrated data, and also its classification potential given the pathology of interest. An initial assessment is performed with simulated data generated under two ‘biological’ conditions. Additionally, we assess our method’s performance with publicly available data from six different studies and two pathologies, i.e. breast cancer and lung cancer. Results obtained from our model, based on simulated and publicly available data, are also compared to those derived by recently published multi-platform integration methodologies.

(a) Performance assessment for simulated data

To evaluate the performance of our methodology, we simulated data by adopting the strategy followed by Wang & Chen [15]. We generated gene-expression values for two independent studies $S_1$ and $S_2$. Each dataset consists of 5000 genes and 200 biological samples under two conditions, treatment (40 biological samples) and control (160 biological samples). In each dataset, 500 genes are simulated to be differentially expressed between the two conditions. The differentially expressed genes between the two conditions for each dataset were selected by using Welch $t$-test $p$-values with Benjamini–Hochberg correction [28].

We used our method to integrate the two datasets and form a merged data, denoted by $S_1 + S_2$, and subsequently select the differentially expressed genes using corrected Welch $t$-test $p$-values. The data were also integrated by using the Gaussian Mixture Modeling-Information Gain (GMM-IG) method by Wang & Chen [15]. GMM-IG is a model-based two-stage methodology; during the first phase, GMM is used to quantify gene expression measurements of the data by using a mixture of two normal distributions per gene, while during the second phase an IG method is applied to select and rank the important genes from integrated datasets. GMM-IG methodology resembles our method in that it assumes two distinct biological states for each gene, i.e. regulated and non-regulated for case and control biological samples, respectively. This is considered to be the null model for the partition framework we present. In particular, although we consider one group of non-regulated gene values, we allow the regulated gene values to be derived by more than one normal distributions, thus unlike GMM-IG we are not fixing the number of clusters but estimating its optimum. Furthermore, GMM-IG treats each gene separately and binarizes it based on its measurements derived from the mixture of two normal distributions in normal and cancerous biological samples. Our partition framework models all gene values together, and in that way considers possible interactions between gene values belonging to the same cluster. Rather than binarizing the data using phenotypic information (e.g. cases versus control biological samples), we report the posterior probability of each gene value belonging to the non-regulated group of genes. Finally, our method only suggests how different datasets can be transformed and merged, without suggesting a gene ranking procedure which can be done by using any statistical significance test on the posterior probabilities data matrix, such as the Welch $t$-test. Nevertheless, a comparison of the two methodologies is presented here, since the integrated data are generated under two distinct biological states, and we would expect both methodologies to perform well.

Figure 2a shows a correspondence at the top (CAT) plot [15,29], which assesses the agreement of the top differentially expressed genes by showing the proportion of true differentially expressed genes found in each dataset. Particularly, we rank the 500 true differentially expressed genes for
Figure 2. Correspondence at the top and integration-driven discovery rate plots on simulated data. (a) CAT plot showing the proportions of genes in common between the reported and the 500 true differentially expressed genes on simulated data. Results are shown for $S_1$, $S_2$, the integrated dataset using our methodology, $S_1 + S_2$, and GMM-IG method. A randomly selected list of 500 genes is also included (black line). (b) The integration-driven discovery rate (IDR) plot for the proportions of genes that were found to be differentially expressed only in the integrated data $S_1 + S_2$ using our methodology. (c) The IDR plot for the proportions of genes that were found to be differentially expressed only in the integrated data using the GMM-IG framework. In all cases, differentially expressed genes are ranked by Welch t-test corrected p-values. (Online version in colour.)

$S_1$, $S_2$ and $S_1 + S_2$ (integrated using our framework), ranked based on their corrected p-values ($p \leq 0.05$). A random selection of 500 genes is also shown (black dashed line), as well as those derived by the GMM-IG method. As can be seen in figure 2, the proportions of the estimated differentially expressed genes reported for $S_1$ and $S_2$ datasets are similar; nevertheless the agreement is higher for the integrated data which indicates the reliability of the integrated methodology. The agreement between our methodology and GMM-IG is higher compared with the randomly ranked gene list; also we can observe a good agreement between the two transformation methodologies when the gene list size ranges between 250 and 340 genes, whereas the proportion in common found from our method increases with the gene list size.

We also show the integration-driven discovery rate (IDR) as defined in [30], namely we report the percentage of genes identified as differentially expressed in the integrated data that were not identified in any of the individual studies. In figure 2b, we can observe the IDR values for $S_1 + S_2$ dataset relative to the significance level of the corrected Welch t-test p-values. It is evident from
the plot that the highly differentially expressed genes found in $S_1 + S_2$ are mostly novel entries compared to those of $S_1$ and $S_2$. In figure 2c, we can observe that the GMM-IG-integrated data follow a similar behaviour.

(b) Breast cancer biomarker case study

In this section, empirical gene-expression datasets from three breast cancer studies [31–33] are integrated under the suggested framework. All datasets and clinical information files were downloaded from the GEO repository as well as the relevant published studies. Table 1 summarizes the content of the three breast cancer datasets denoted by $D_1^B$, $D_2^B$ and $D_3^B$, respectively. We validated our results based on the estrogen receptor (ER) status of biological samples, which has proved to be an important risk factor for breast cancer genesis [37], having two categorical values ER+ and ER−. Details about the ER status of each study’s biological samples are given in the parentheses of the last column of table 1.

Data were normalized and log transformed according to platform type to ensure that the mean expression value for each study is zero and the expression values for a given gene are approximately normally distributed under each condition. Within each dataset, genes with missing values in greater than 30% of the biological samples were removed, and the remaining missing values were imputed using $k$-nearest neighbours imputation algorithm with $k = 10$ [38], as implemented in R (http://www.bioconductor.org/packages/release/bioc/html/impute.html). We restrict our analysis to the set of common genes in the available studies recognized by cross-referencing HGNC symbols (Hugo Gene Nomenclature Committee, http://www.genenames.org/) using annotate R package (http://www.bioconductor.org/packages/release/bioc/html/annotate.html), whereas probe IDs mapped to the same gene symbol were averaged. This resulted in a set of 4549 genes common across the three breast cancer datasets.

According to the suggested methodology, values of each data matrix are transformed to posterior probability values $p_{ij}$, measuring how likely each $x_{ij}$ value is to belong to the non-regulated group of gene values $C_0$, over the 100 bootstrap samples which however retain the ER analogy of the original data (for example 44 : 13 for $D_1^B$). Transformed data matrices are then merged forming a unified matrix denoted by $D_1^B + D_2^B + D_3^B$, which consists of 4549 genes and 466 biological samples (biological samples with unknown ER status are discarded). To evaluate the performance of the unified data matrix as opposed to data derived by other transformation methodologies, we selected a set of differentially expressed genes, with respect to ER status, based on corrected Welch $t$-test $p$-values ($p \leq 0.05$). To evaluate the significance of selected differentially expressed genes, we compared their classification potential to that of gene lists derived by four

<table>
<thead>
<tr>
<th>breast cancer datasets</th>
<th>microarray platforms</th>
<th>accession number</th>
<th>probe sets</th>
<th>no. samples (ER+, ER−, NA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1^B$, Sorlie et al. [31]</td>
<td>spotted DNA/cDNA</td>
<td>GEO: GSE3193</td>
<td>9216</td>
<td>58 (44, 13, 1)</td>
</tr>
<tr>
<td>$D_2^B$, Sotiriou et al. [32]</td>
<td>affymetrix U133A</td>
<td>GEO: GSE2990</td>
<td>12 625</td>
<td>189 (149,34,6)</td>
</tr>
<tr>
<td>$D_3^B$, Wang et al. [33]</td>
<td>affymetrix U133A</td>
<td>GEO: GSE2034</td>
<td>22 000</td>
<td>225 (186,69,0)</td>
</tr>
<tr>
<td>lung cancer datasets</td>
<td>no. samples</td>
<td>stages: 1, 2, 3, 4, n.a.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_1^L$, Harvard [34]</td>
<td>affymetrix U95Av2</td>
<td>Broad Institute - Cancer</td>
<td>12 453</td>
<td>139 (60, 24, 10,15,20)</td>
</tr>
<tr>
<td>$D_2^L$, Michigan [35]</td>
<td>affymetrix HuGeneFL</td>
<td>caArray: beer-00153</td>
<td>6663</td>
<td>86 (67,0,19,0,0)</td>
</tr>
<tr>
<td>$D_3^L$, Stanford [36]</td>
<td>spotted DNA/cDNA</td>
<td>ArrayExpress:</td>
<td>23 100</td>
<td>68 (9,2,6,10,27)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>microarray platforms</th>
<th>accession number</th>
<th>probe sets</th>
<th>no. samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>spotted DNA/cDNA</td>
<td>GEO: GSE3193</td>
<td>9216</td>
<td>58 (44, 13, 1)</td>
</tr>
<tr>
<td>affymetrix U133A</td>
<td>GEO: GSE2990</td>
<td>12 625</td>
<td>189 (149,34,6)</td>
</tr>
<tr>
<td>affymetrix U133A</td>
<td>GEO: GSE2034</td>
<td>22 000</td>
<td>225 (186,69,0)</td>
</tr>
<tr>
<td>affymetrix U95Av2</td>
<td>Broad Institute - Cancer</td>
<td>12 453</td>
<td>139 (60, 24, 10,15,20)</td>
</tr>
<tr>
<td>affymetrix HuGeneFL</td>
<td>caArray: beer-00153</td>
<td>6663</td>
<td>86 (67,0,19,0,0)</td>
</tr>
<tr>
<td>spotted DNA/cDNA</td>
<td>ArrayExpress:</td>
<td>23 100</td>
<td>68 (9,2,6,10,27)</td>
</tr>
</tbody>
</table>
Table 2. Classification comparison for empirical cancer datasets: The average classification accuracy for 100 ten-fold cross validation tests over $D^B_3$ for breast cancer data and $D^L_3$ for lung cancer data, using the 50 highly significant genes selected from each of the five methodologies.

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>naive Bayesian classifier</th>
<th>decision tree</th>
<th>10-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>breast cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>our method</td>
<td>0.84 ± 0.15</td>
<td>0.94 ± 0.13</td>
<td>0.96 ± 0.06</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>metaMA</td>
<td>0.87 ± 0.12</td>
<td>0.90 ± 0.10</td>
<td>0.90 ± 0.20</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>XDE</td>
<td>0.80 ± 0.15</td>
<td>0.93 ± 0.10</td>
<td>0.90 ± 0.09</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>XPN</td>
<td>0.77 ± 0.10</td>
<td>0.93 ± 0.13</td>
<td>0.93 ± 0.07</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>GMM-IG</td>
<td>0.80 ± 0.14</td>
<td>0.95 ± 0.09</td>
<td>0.95 ± 0.07</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>lung cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>our method</td>
<td>0.80 ± 0.17</td>
<td>0.82 ± 0.10</td>
<td>0.94 ± 0.10</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>metaMA</td>
<td>0.79 ± 0.11</td>
<td>0.81 ± 0.10</td>
<td>0.84 ± 0.18</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>XDE</td>
<td>0.81 ± 0.10</td>
<td>0.81 ± 0.12</td>
<td>0.83 ± 0.07</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>XPN</td>
<td>0.80 ± 0.10</td>
<td>0.82 ± 0.11</td>
<td>0.92 ± 0.08</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>GMM-IG</td>
<td>0.84 ± 0.13</td>
<td>0.95 ± 0.11</td>
<td>0.91 ± 0.10</td>
<td>1.00 ± 0.00</td>
</tr>
</tbody>
</table>

transformation methods, namely we consider a meta-analysis method metaMA [11], a cross-
study method for the analysis of differentially expressed genes XDE [9], and two transformation
methods XPN [14] and GMM-IG [15]. Their main difference is that XDE and XPN methods use
cross-platform corrections, unlike metaMA, GMM-IG and our methodology. Additionally, the
phenotypic information is included in the model parameters for metaMA, XDE and GMM-IG
methods, whereas for XPN and GMM-IG methods pre-specify the number of clusters in the data.

In gene-expression data analysis, the most important goal may be to find a set of genes
showing notably similar upregulation and downregulation under a set of conditions. We
selected a 100 gene signature consisting of the most significant differentially expressed genes
with respect to ER status, in order to evaluate their significance as candidate biomarkers.
In all cases, we selected the 100 most significant genes with respect to ER. Four classifiers,
KNN, support-vector machine (SVM), naive Bayesian classifier and decision tree, were
employed to evaluate the estimated gene sets as implemented within the e1071 R package
(http://cran.r-project.org/web/packages/e1071/index.html). For each classifier, we performed
10-fold cross-validation tests over $D^B_3$ dataset using the 50 most highly significant genes based
on corrected Welch t-test p-values ($p \leq 0.05$). Particularly, the datasets were randomly divided
into two sets retaining their ER ratio: 90% of biological samples were assigned to the training
set and 10% of the biological samples were assigned to the validated set. The training set IDs
were used to train the classifier and the classification performance of the trained classifier was
then evaluated on the validated set. We repeated the procedure 100 times calculating the mean
classification accuracy. In table 2, we report the average classification accuracies. We can observe
that our method’s average classification accuracies are higher than or comparable to those by
metaMA, XDE, XPN and GMM-IG methodologies, whereas SVM is appearing to be the most
conservative of the four classifiers used. Our methodology outperforms XPN, which also uses
partition modelling. This could be attributed to the fact that the number of clusters in XPN
are estimated by k-means algorithm, whereas our method estimates them by maximizing the
suggested likelihood model. Overall, we can observe that our method’s classification accuracy
results resemble those of GMM-IG for all four classifiers; this should be expected due to the binary
nature of the specific classification problem (two ER biological states) which is the underlying
model for both methods.
Figure 3. DR plot for the proportions of genes that were found to be differentially expressed only in the breast cancer-integrated data. IDR was computed for the integrated, using our methodology, dataset \((D^b_1 + D^b_2 + D^b_3)\), relative to the threshold imposed to the corrected Welch \(t\)-test \(p\)-values. IDR plots for XPN and GMM-IG transformation methods are also shown. (Online version in colour.)

Figure 3 shows the IDR plot for the proportion of genes found to be differentially expressed between ER+ and ER− biological samples in the integrated data and were not reported as such in any of the individual studies. Here, our Bayesian partition model is denoted by BPM. We compare our method’s performance to those of XPN methodology, which is also based on partition modelling, and GMM-IG method, which was reported to perform well in terms of classification accuracy as shown in table 2. Both XPN and GMM-IG methods perform well; however GMM-IG outperforms the other two methods for strict \(p\)-value threshold values, being the only method containing a gene ranking process. However, in accordance with the simulated data results, our method performs better as the \(p\)-value threshold increases, which suggests that our methodology estimated additional differentially expressed genes for the integrated breast cancer data.

(c) Lung cancer biomarker case study

We also consider three lung cancer gene-expression studies originally derived in [34–36], respectively. Details about the three lung cancer datasets denoted by \(D^L_1\), \(D^L_2\) and \(D^L_3\) can be found in table 1. In the parentheses of the last column of the table, we include the stages of the disease ranging from 1 to 4, where 4 denotes metastatic cancer. The normalized and merged datasets were made publicly available by Parmigiani et al. [39] and were downloaded from the lungExpression R package available on the Bioconductor website (ftp://ftp.questnet.net.au/pub/bioconductor/packages/2.8/data/experiment/html/lungExpression.html). Datasets were merged by UniGene identifiers (http://www.ncbi.nlm.nih.gov/unigene) resulting in a common set of 3171 genes; probe IDs mapped to the same UniGene symbol were averaged and IDs with many mappings were excluded. The integrated dataset \(D^L_1 + D^L_2 + D^L_3\) consists of 246 biological samples in total (biological samples with unknown metastatic status are discarded).

Similar to the breast cancer data, we selected a 50 gene signature consisting of the highly significant differentially expressed genes with respect to the disease stage classification, in order to evaluate their significance as candidate biomarkers. For each classifier, we performed 10-fold cross-validation tests over \(D^L_2\) dataset. To estimate the statistically significant genes, we used analysis of variance and selected genes based on \(p\)-values \((p \leq 0.05)\). We compared our finding to those derived by metaMA, XDE, XPN and GMM-IG methods as shown in table 2. Based on the
average classification accuracies reported, our methodology seems to outperform all methods but GMM-IG, which should be attributed to the fact that GMM-IG includes a ranking stage, and thus the 50 gene signature in the case of GMM-IG is carefully scrutinized. Nevertheless, our method outperforms GMM-IG with the decision tree classifier and seems to perform similar to XPN which is also based on a partition model. In figure 4, we show the IDR plot for the proportion of genes found to be differentially expressed between the biological samples of different disease stages in the integrated data using our methodology, denoted by ‘our BPM’, and were not reported as such in any of the individual studies. Again we can observe that the number of common differentially expressed genes increases with the p-value threshold, although in this case values range from 0.3 to 0.72. All three methodologies exhibit similar performance; nevertheless GMM-IG methodology performs well for gene lists with $p \leq 0.02$, whereas our method performs better for longer gene lists.

4. Discussion

Gene expression analysis methods aim mostly at identifying subsets of genes that could explain the biological processes underlying microarray experiments. However, as the availability of microarray data increases, there is a growing need for cross evaluation of multiple independent microarray datasets along with careful summarization of results. Multi-platform transformation methods are applied for that purpose reducing study-specific biases and aiming to yield results which offer improved reliability and validity. Here, we propose a two-way partition model for multi-experiment integration that reports the posterior probability of each gene value based on its partition membership and the underlying assumption that genes are either regulated or non-regulated in each biological sample. We assumed that expression values which are normally distributed around zero with a small variance are likely to belong to non-regulated genes, whereas values which follow a normal distribution with non-zero mean value and larger variance are more likely to belong to regulated genes. In this context, use of a fully Bayesian model has several desirable features. The model borrows strength across both genes and experiments and can thereby provide better estimates of the gene-specific effects, without using study-specific phenotypic information.
The analysis on simulated and empirical cancer datasets showed that our method increased power and reproducibility. Furthermore, we have established that our methodology can produce comparable results with other meta-analysis and transformation methods, when identification of differentially expressed genes and classification of experiments are considered, even though a selection procedure is not included in the suggested framework.

One of the major advantages of our method is that we can unify gene-expression datasets based on the estimated partition of the data without taking into account the biological samples' phenotypic structure or specifying the number of clusters in the data nevertheless that could be imposed to the model. A possible extension of the model is to filter the data by a particular biological function of interest and in that way resemble other methodologies which aim to integrate only functionally important subset of genes. An important future research direction is to apply our method to NGS data, such as RNA sequencing data or copy number variations, and consider amendments in order to analyse multiple diverse data which often offer complementary genome information.

Acknowledgement. The authors are members of the BM1006 COST Action, SeqAhead: Next Generation Sequencing Data Analysis Network.

Data accessibility. The microarray data used for this study were derived from the NCBI Gene Expression Omnibus and ArrayExpress databases with accession numbers: GSE3193, GSE2990, GSE2034 and E-GEOD-3398.

Funding statement. We gratefully acknowledge funding from the EU DICODE (Mastering Data-Intensive Collaboration and Decision Making) Collaborative Project (FP7, ICT-2009.4.3, contract no. 257184), and the 09SYN-13-901 EPAN II Cooperation grant (co-funded by the EU and the Greek State).

References


29. Irizarry RA *et al.* 2005 Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350. ([DOI:10.1038/nmeth756](http://dx.doi.org/10.1038/nmeth756))


