BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

Regularized Estimation and Biomarker Selection in Microarray Meta-Analysis

Shuangge Ma [a] and Jian Huang [b]

[a]Department of Epidemiology and Public Health, Yale University, New Haven, CT, USA
[b]Departments of Statistics and Actuarial Science, and Biostatistics, University of Iowa, Iowa City, IA, USA

ABSTRACT

Motivation: In pharmacogenetic studies, it is common that multiple microarray studies are conducted to investigate the relationship between a phenotype and gene expressions. An important goal of such studies is to discover influential genes that can be used as disease biomarkers and construct predictive models. To increase statistical power, meta analysis should be used to combine results from these studies. However, it is difficult to apply the standard meta analysis approaches because of high-dimensionality of microarray data and because different microarray platforms and experimental settings used in different studies may not be directly comparable.

Results: We propose a Meta Threshold Gradient Descent Regularization (MTGDR) approach for regularized meta analysis. The proposed approach is capable of selecting the same sets of influential genes across different studies, while allowing for different estimates for different platforms or experiments. To demonstrate the proposed approach, we use microarray data with binary outcome as an example in the context of logistic regression models. We analyze datasets from pancreatic and liver cancer studies using the proposed approach.

Availability: R code is available upon request.

Contact: shuangge.ma@yale.edu

1 INTRODUCTION

Microarrays are capable of profiling human tissues on a genome wide scale and have been extensively used in pharmacogenetic studies, where expressions of thousands of genes are measured along with certain clinical outcomes. A major goal of such studies is to identify genes that can be used as predictive biomarkers for disease diagnosis and prognosis and as targets for therapy. It is now very common that multiple microarray studies are carried out to identify influential genes that are related to the same phenotype in the same species (Choi et al. 2004; Ghosh et al. 2003; Wang et al. 2004; and Warnat et al. 2005).

Most microarray studies include a large number of genes and a much smaller number of subjects. This “small n large d” scenario makes analysis of microarray data challenging. One way to increase statistical power is to simultaneously consider multiple datasets with similar setups by use of meta analysis. However, meta analysis can be complicated due to the high-dimensionality of microarray data and technical differences between platforms, which can lead to differences in the intrinsic nature of the produced expression data. Arrays that hybridize one sample at a time (e.g. synthesized oligonucleotide arrays) measure gene expression based directly on the signal intensity of each probe set. Spotted cDNA arrays hybridized with fluorescent labeled targets, in contrast, typically measure the ratio of the signal from a test sample to the signal of a co-hybridized reference sample. Thus one unit increase in the expression levels measured in a study using cDNA arrays is usually not directly comparable to one unit increase in a study using oligonucleotide arrays. For example, it has been shown that data from Affymetrix GeneChip oligonucleotide microarrays correlate poorly with the data from custom-printed cDNA microarrays (Kuo et al. 2002). Thus data from different platforms may not be directly combined.

Several papers have considered the problem of detecting differentially expressed genes based on multiple datasets. Examples include an approach using proper transformations to directly integrate raw gene expression data (Warnat et al. 2005); a Lasso based method (Ghosh et al. 2003); a random effects model based method (Stevens and Doerge 2005); and a Bayesian approach (Jung et al. 2006) among many others. Although such studies are informative, they do not directly lead to predictive models.

There are also publications that consider the problem of constructing predictive models from multiple microarray studies. For example, a majority voting with impact factors algorithm is proposed in Fung and Ng (2004), where the main goal is to construct models for predicting categorical outcomes. Predictive model building is also considered by Jiang et al. (2004), where gene shaving methods based on random forrest and Fisher’s linear discrimination are applied. Statistical software, including the R packages metaArray (Ghosh and Choi 2006), MergeMaid (*http://astor.som.jhmi.edu/MergeMaid/*) and RankProd (Hong et al. 2006), has been developed for microarray meta analysis. Although aforementioned studies investigate predictive model building, biomarker selection is either not considered or carried out with relatively ineffective approaches. In these studies, a critical underlying assumption is that although different platforms are not directly comparable, the underlying biological models (for example the sets of differentially expressed genes or the sets of genes with predictive power) are the same across studies.

On the other hand, for microarray data generated under a single experimental setting, simultaneous biomarker selection and predictive model building has been extensively investigated. Examples include the Lasso in binary classification (Ghosh and Chinnaiyan 2004) and survival analysis (Gui and Li 2005a), the Threshold Gradient Directed Regularization–TGDR, in classification (Ma and Huang 2005) and survival analysis (Gui and Li 2005b; Ma and Huang 2007), and the support vector machine with SCAD penalty in Zhang et al. (2006). We refer to Li

(2007) for a thorough review of existing methods. The regularized approaches are capable of selecting a small number of influential genes along with predictive model building. They are usually much more informative than simply detecting differential genes. Although great successes have been demonstrated by these approaches, they cannot be used directly in meta analysis, since different platforms and experimental settings are not directly comparable. We further explain this issue in Section 4.

In this paper, we propose a Meta Threshold Gradient Directed Regularization (MTGDR) method for simultaneously biomarker selection and predictive model building for microarray meta analysis. The MTGDR takes advantage of recent development in regularized biomarker selection methods (with single microarray dataset) and is capable of analyzing several datasets generated under different platforms or experimental settings. It thus fills the gap between available meta analysis methods and single-dataset regularization methods. Compared with available meta analysis methods, the MTGDR can select a small number of biomarkers with joint predictive power and lead to parsimonious predictive models. Compared with single-dataset regularized methods, the MTGDR allows different estimates for different experiments and hence can accommodate different experimental settings.

Notations and data settings are first introduced in Section 2. The MTGDR algorithm is described in Section 3. We demonstrate the proposed method on microarray data with binary outcomes. However, the proposed approach is also applicable to quantitative outcomes. We analyze four pancreatic cancer data in Section 4 and four liver cancer data in Section 5. Discussions are provided in Section 6.

2 DATA SETTINGS

For simplicity of notation, we assume that the same set of d genes are measured in all M different experiments with $M > 1$. Let $Y^1, \ldots, Y^M$ be the clinical outcomes and let $Z^1, \ldots, Z^M$ represent the gene expressions measured in these studies. We postpone discussions of possibly different sets of genes from different studies to the Discussion section. For $m = 1, \ldots, M$, we assume Y^m is associated with Z^m through the model $Y^m \sim \phi(Z^{m\prime}\beta^m)$, where a' denotes the transpose of a, and where ϕ is a known regression function and is assumed to be the same across all M studies.

We assume that the same statistical model holds across different experiments. This assumption has been generally made in microarray meta analysis. However, we allow different regression coefficients β^m. The rationale is that one unit gene expression increase in experiment 1 (say for example a cDNA study) is not equivalent to one unit increase in experiment 2 (say for example an Affymetrix study). This assumption shares the same spirits as the fixed effect models in standard meta analysis (Stevens and George 2005).

Although the proposed MTGDR approach is generally applicable regardless of clinical outcome types and statistical models, we describe it for binary outcome data. Let $Y = 1$ denote the presence and $Y = 0$ denote the absence of disease. We assume the commonly used logistic regression model, where for study m, the logit of the conditional probability is $logit(P(Y^m = 1|Z^m)) = \alpha^m + Z^{m\prime}\beta^m$. Here α^m is the unknown intercept for experiment m.

Suppose that there are n_m iid observations in experiment m. For experiment m, the log-likelihood is:

$$R^m(\alpha^m, \beta^m) = \sum_{j=1}^{n_m} Y_j^m \log\left(\frac{\exp(\alpha^m + \beta^{m\prime} Z_j^m)}{1 + \exp(\alpha^m + \beta^{m\prime} Z_j^m)}\right) + (1 - Y_j)\log\left(\frac{1}{1 + \exp(\alpha^m + \beta^{m\prime} Z_j^m)}\right). \quad (1)$$

Since the intercept α^m will not be subject to regularization, for simplicity, we denote $R^m(\alpha^m, \beta^m)$ as $R^m(\beta^m)$.

3 MTGDR METHOD

3.1 Regularized microarray biomarker selection

In microarray studies, it is usually assumed that although tens of thousands of genes are surveyed, only a small number of them are actually associated with the clinical outcome of interest. Statistically this is the basis for biomarker selection and the sparsity assumption, i.e, most components of the regression coefficient β are zero. For microarray data with binary outcome and logistic model, regularized estimation methods can be used include the Lasso, SCAD and TGDR among others. The proposed MTGDR is based on the TGDR, which is introduced by Friedman and Popescu (2004) in the context of linear regression and has been used for biomarker selection in microarray classification (Ma and Huang 2005) and survival analysis (Gui and Li 2005b). For completeness, we briefly describe the TGDR algorithm below.

Consider experiment m only. Denote $\Delta\nu$ as the small positive increment as in ordinary gradient descent searching. In the implementation of this algorithm, we choose $\Delta\nu = 10^{-3}$, Denote $\nu_k = k \times \Delta\nu$ as the index for the point along the parameter path after k steps. Let $\beta^m(\nu_k)$ denote the parameter estimate of β^m corresponding to ν_k. For a fixed threshold $0 \leq \tau \leq 1$, the TGDR algorithm consists of the following iterations:

1. Initialize $\beta^m(0) = 0$ and $\nu_0 = 0$.
2. With current estimate β^m, compute the negative gradient $g^m(\nu) = -\partial R^m(\beta^m)/\partial\beta^m$. Denote the j^{th} component of $g^m(\nu)$ as $g_j^m(\nu)$. If $max_j\{|g_j^m(\nu)|\} = 0$, stop the iteration.
3. Compute the threshold vector $f^m(\nu)$ of length d, where the j^{th} component of $f^m(\nu)$:
$$f_j^m(\nu) = I(|g_j^m(\nu)| \geq \tau \times max_l|g_l^m(\nu)|).$$
4. Update $\beta^m(\nu + \Delta\nu) = \beta^m(\nu) - \Delta\nu \times g^m(\nu) \times f^m(\nu)$ and update ν by $\nu + \Delta\nu$, where the product of f^m and g^m is component-wise.
5. Steps 2-4 are iterated k times. The number of iteration k is determined by cross validation.

We refer to Friedman and Popescu (2004) and Ma and Huang (2005) for more detailed descriptions of the TGDR algorithm. For microarray data, with cross validated k, many components of β^m are estimated to be exactly zero. Biomarker selection is achieved by only including genes with nonzero coefficients. Compared with regularized methods such as Lasso, the TGDR approach is a non-linear boosting-like approach. It can be used to analyze a single dataset, or pooled dataset by simply merging different datasets. However, it is not a meta analysis method.

3.2 MTGDR algorithm

We propose the following Meta-TGDR (MTGDR) approach for regularized microarray meta analysis. We make the following two essential assumptions: (1) although the same logistic regression model holds, the regression coefficients β^m may be different across studies; (2) the sets of genes with nonzero coefficients (i.e., the identified genes) are the same across studies. Assumption (1) is mainly due to the concern of different platforms; Assumption (2) assumes that although different experiments are not directly comparable, the biological conclusions should be comparable, i.e, we should conclude the same sets of genes to be significantly associated with the outcome.

Let $\beta = (\beta^1, \ldots, \beta^M)$ and $R(\beta) = R^1(\beta^1) + \ldots + R^M(\beta^M)$. Here β is a $d \times M$ matrix. Using notations similar to those in Section 3.1, the MTGDR algorithm can be described as follows.

1. Initialize $\beta = 0$ (component-wise) and $\nu_0 = 0$.
2. With current estimate β, compute the negative gradient matrix $g(\nu) = -\partial R(\beta)/\partial\beta$, where the (j, m) element of g is $g_{j,m}(\nu) = -\partial R^m(\beta^m)/\partial\beta^m_j$.
3. Compute the length d vector of meta gradient G, where the j^{th} component of G is $G_j(\nu) = \sum_{m=1}^{M} g_{j,m}(\nu)$.
4. Compute the meta threshold vector $F(\nu)$ of length d, where the j^{th} component of $F(\nu)$:

$$F_j(\nu) = I(|G_j(\nu)| \geq \tau \times max_l|G_l(\nu)|).$$

5. Update the (j, m) element of β: $\beta_{j,m}(\nu + \Delta\nu) = \beta_{j,m}(\nu) - \Delta\nu g_{j,m}(\nu)F(\nu)$ and update ν by $\nu + \Delta\nu$.
6. Steps 2-5 are iterated k times, where k is determined by cross validation.

The MTGDR algorithm shares some similarities with the TGDR: it starts with the zero estimate and coefficients for important genes (defined as those with large meta gradients) are updated at each iteration. With $\tau > 0$ and finite k, only a small number of genes may have nonzero coefficients.

In steps 2 and 5, the gradients are computed for each experiment (dataset) and estimates are updated accordingly. By doing so, we allow different estimates for different experiments, which satisfies assumption (1). In step 3, the meta gradient, which is defined as the sum across different experiments, is computed. A meta threshold vector is computed in step 4. By doing so, we force the threshold vector to be the same (for each gene) across experiments. So when a gene is include, it is included in all models across experiments, which corresponds to assumption (2).

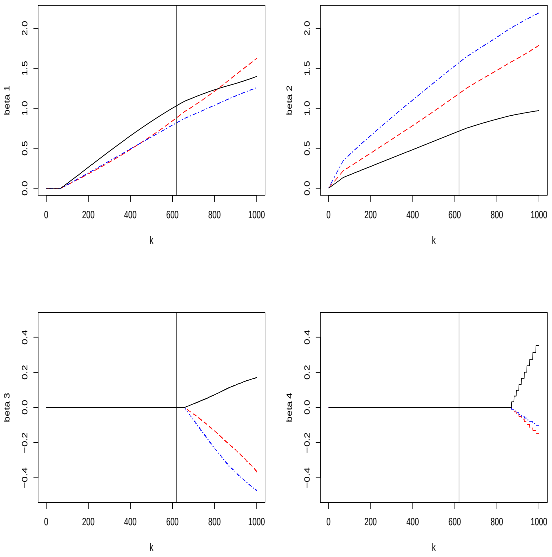

Fig. 1. Parameter path as a function of k. Dashed red line: simulated experiment 1; Dash-dotted blue line: simulated experiment 2; Solid black line: simulated experiment 3. Vertical lines: the cross validated k.

The meta gradient in step 3 is the most straightforward definition that considers the common effect in all studies. Consider for example gene 1 only has significant effect in experiment 1; whereas gene 2 has moderate effects in all experiments. Then the sum of gradients (combined effects) for gene 2 may be larger than that for gene 1. Gene 2 is thus more likely to be selected since consistent effects are demonstrated across studies, whereas gene 1 may demonstrate significant effect in experiment 1 simply because of experimental variation or purely by chance. If a gene shows significant effects in all experiments but the gradients have both positive and negative signs, then the sum may be small and hence this gene may not be selected. The rationale is that if a gene is selected, it is supposed to show similar biological effects across studies (for example up-regulation of this gene is positively associated with the clinical outcome). So if both positive and negative associations are observed, then inconsistent biological conclusions are reached in different experiments. Hence the corresponding gene should not be selected. It is worth pointing out that with the proposed MTGDR, it is still possible that a gene selected has coefficients with different signs in different experiments. For example, if a gene demonstrates dramatically large positive effect in one study but no or small negative effect in other studies, this gene is still possible to be selected.

3.3 Tuning parameter selection

The MTGDR involves tuning parameters k and τ, which jointly determine the property of the estimate. We use V-fold cross validation to determine k and τ. With V-fold cross validation, the tuning parameters with the best predictive power are selected. Partial protection against over-fitting is also provided. In our study, we set $V = 3$ mainly due to the small sample sizes concern.

3.4 Evaluation

With binary outcome, prediction evaluation can be simply based on prediction error. We consider the following Leave-One-Out (LOO) approach. We first remove one subject from the data. With reduced data, we carry out cross validation and MTGDR estimation. We then use this estimate to make prediction for the one removed record. With the logistic model, the predicted probability can be computed. We use 0.5 as cutoff and predict the class label. We then repeat this procedure over all subjects. Prediction error can then be computed. This approach has been used in Dettling and Buhlmann (2003) and Ma, Song and Huang (2007).

3.5 A graphic demonstration

We use the following small numerical example to demonstrate the MTGDR parameter path. For $m = 1, 2$ and 3, we generate data from $logit(P(Y^m = 1|Z^m) = \beta_1^m Z_1^m + \beta_2^m Z_2^m + \beta_3^m Z_3^m + \beta_4^m Z_4^m$. In this simulated meta analysis, we have three independent experiments and four genes per experiment. Z_i^js are generated independently and $N(0, 0.5)$ distributed. We set $\beta^1 = (2.0, 2.0, 0, 0)$, $\beta^2 = (1.5, 1.5, 0, 0)$ and $\beta^3 = (1.0, 1.0, 0, 0)$. In all three experiments, only the first two genes are associated with the binary outcomes, although their corresponding coefficients are different. We simulate 50 observations under each experiment.

The 3-fold cross validation select $\tau = 1.0$ and $k = 620$. We show in Figure 1 the parameter paths as a function of k for $\tau = 1.0$. We can see that for any k the estimated coefficients for one gene are either all zero or all nonzero across experiments. This corresponds to the assumption that if a gene is significantly associated with the outcome, the estimated coefficients should be nonzero in all experiments. We can also see that individual parameter paths are similar to Lasso paths. This property has been demonstrated for the TGDR in Friedman and Popescu (2004). With cross validated tuning parameter, only the first two genes enter the final predictive models and the true underlying models are recovered.

4 PANCREATIC CANCER STUDY

4.1 Data settings

Pancreatic ductal adenocarcinoma (PDAC) is a major cause of malignancy-related death. Apart from surgery, there is still no effective therapy and even resected patients usually dies within one year postoperatively. Several studies have applied microarray technology to pancreatic cancer, targeting at identification of predictive pancreatic cancer biomarkers. We use four datasets in our study: Iacobuzio-Donahue et al. (2003), Logsdon et al. (2003), Crnogorac-Jurcevic et al. (2003) and Friess et al. (2003). These four datasets have been selected and used in the meta analysis of Grutzmann et al. (2005). We show data descriptions in Table 1 and refer to original publications for more experimental details. Two of the four studies use cDNA arrays and two used oligonucleotide arrays. Cluster ID and gene names are assigned to all of the cDNA clones and Affymetrix probes based on UniGene Build 161. The two sample groups considered in our analysis are PDAC and normal pancreatic tissue. Data on chronic pancreatitis are available from Logsdon et al. (2003) and Friess et al. (2003). Following Grutzmann et al. (2005), those CP samples are not used in our study.

Grutzmann et al. (2005) identified a consensus set of 2984 UniGene IDs. Our dataset is provided by Dr. Grutzmann and contains the same set of 2984 genes. We further screen genes with more than 30% missingness in any of the four datasets. 1204 genes pass this screening. For Affymetrix expression measurements, we add a floor or 10 and make log2 transformations. We fill in missing values with medians across samples (for each dataset separately), and then standardize each gene to zero mean and unit variance. For cDNA studies, we fill in missing values with sample medians for each dataset separately, and then standardize each gene to zero mean and unit variance.

4.2 Individual TGDR analysis

We first analyze the four datasets separately. For each dataset, the logistic regression model is assumed and we use the TGDR described in section 3.1 for regularized estimation and biomarker selection. Tuning parameters are determined via 3-fold cross validation. For each dataset, we use the LOO approach described in section 3.4 to compute prediction error.

Table 2. Pancreatic cancer datasets: genes with nonzero coefficients in MTGDR.

UniGene	P1	P2	P3	P4
Hs.107	-0.078	-0.074	-0.096	-0.062
Hs.12068	-0.265	-0.387	-0.189	-0.250
Hs.16269	0.038	0.055	0.060	0.017
Hs.169900	-0.879	-0.992	-0.693	-0.775
Hs.180920	-0.144	-0.244	-0.223	-0.189
Hs.241257	0.096	0.128	0.124	0.062
Hs.287820	1.051	1.157	1.055	0.736
Hs.317432	-0.023	-0.012	-0.053	-0.022
Hs.5591	-0.082	-0.170	-0.149	-0.149
Hs.62	0.111	0.100	0.104	0.126
Hs.66581	-0.024	-0.028	-0.034	-0.013
Hs.75335	-0.270	-0.259	-0.250	-0.250
Hs.76307	0.435	0.303	0.616	0.416
Hs.78225	0.011	0.010	0.018	0.010
Hs.83383	-0.074	-0.094	-0.066	-0.085

For the four datasets, 7 (P1), 10 (P2), 6 (P3) and 1 (P4) genes are selected in the final models, respectively. There is only one common gene selected in both P2 and P3. Otherwise there is no overlap between the four sets of selected gene. So the gene selection results are not different across these studies. In addition, the prediction performance is unsatisfactory. For example, if we use estimate from P2 to make predictions for P1, P3 and P4, the error rates are 0.27, 0.43 and 0.36, respectively. We conclude that separate TGDR analysis results are unsatisfactory in terms of reproducibility across studies and prediction across studies.

4.3 Pooled TGDR analysis

In the second set of analysis, we ignore the fact that the four datasets are from different studies and different platforms and simply pool them together. The pooled sample size is now 56.

We apply the TGDR for regularized estimation and biomarker selection with the 56 subjects. 22 genes are identified and included in the final model. The LOO approach mis-classifies 2 subjects, leading to a prediction error 0.036, which is much improved from separate TGDR analysis in section 4.2. As has been noted in Grutzmann et al. (2005), the four selected datasets are relatively easy to classify. This partly explains the satisfactory prediction performance of simply pooling all data together.

4.4 MTGDR analysis

We use the proposed MTGDR method to analyze the pancreatic cancer data. Tuning parameters are chosen via 3-fold cross validation. 15 genes are selected in the final models. We show the gene IDs and corresponding estimate in Table 2. We can observe from Table 2 that (1) if a gene has nonzero coefficient for one dataset, then it has nonzero coefficients for all datasets; (2) the estimated coefficients for one gene can be different across all studies; this is the extra flexibility allows by the MTGDR compared with pooled analysis; and (3) although the estimated coefficients are

Table 1. Pancreatic cancer gene expression datasets used in the meta-analysis. PDAC: number of pancreatic adenocarcinoma tissue samples analyzed; N: number of normal pancreatic tissues; CP: number of chronic pancreatitis; Array: type of array used in the study; UG: number of unique UniGene cluster presented on the arrays.

Dataset	P1	P2	P3	P4
Author	Logsdon	Friess	Iacobuzio-Donahue	Crnogorac-Jurcevic
PDAC	10	8	9	8
N	5	3	8	5
CP	5	8	–	–
Array	Affy. HuGeneFL	Affy. HuGeneFL	cDNA Stanford	cDNA Sanger
UG	5521	5521	29621	5794

different for one gene across studies, their signs are the same. The same signs lead to similar biological conclusions–i.e., whether up-regulation of genes are positively or negatively associated with the clinical outcome of interest. This consistency in terms of biological conclusions partly supports the validity of the proposed MTGDR.

The prediction error is computed using the LOO approach described in section 3.4. 2 subjects cannot be properly predicted, leading to a prediction error 0.036. The surprisingly satisfactory prediction performance of the pooled analysis (compared with the MTGDR) can be partly explained by the fact that estimated coefficients in Table 2 are similar across genes. So forcing them into one value (as in the pooled analysis) will still lead to satisfactory prediction.

The MTGDR identifies 15 genes, which is fewer than the pooled analysis, and yet its prediction performance is the same as the pooled analysis with 22 genes. With $d >> n$, more parsimonious models are expected to be more reliable. In addition, identifying a smaller number of predictive biomarkers can lead to a more focused hypothesis for further investigation.

5 LIVER CANCER STUDY

5.1 Data settings

Gene expression profiling studies have been carried out on hepatocellular carcinoma (HCC), which is among the leading causes of cancer death in the world. A microarray meta analysis is carried out in Choi et al. (2004), where the main goal is to detect differentially expressed genes. Two sets of data are analyzed in Choi et al. (2004). The first set contain five independent studies (referred as data P1–P5 in Choi et al. 2004). As in the pancreatic cancer data, datasets P1–P5 have limited overlapped genes. So in our study, we focus on data D1–D4 in Choi et al. (2004) only. Dataset information is shown in Table 3 (partly reproduced from Table 1 of Choi et al. 2004).

Datasets D1–D4 were generated in three different hospitals in South Korea. Although the studies were designed in a controlled setting, Choi et al. (2004) "failed to directly merge the data even after normalization of each dataset."

In studies D1–D3, expressions of 10336 genes were measured. In study D4, expressions of 9984 genes were measured. We focus on the 9984 genes that are measured in all four studies. We first pre-process the data as follows: (1) if a gene has more than 30% of missing for any one dataset, then this gene is removed from downstream analysis. 3122 out of 9984 genes pass this screening. (2) if a subject has more than 30% missing expressions for the 3122 genes, then this subject is removed from downstream analysis. 8 subjects are removed, leading to an effective sample size of 125. We show the number of subjects actually used in the analysis in Table 3. (3) For each dataset, we then fill in missing expression values with medians across samples. (4) We compute the two-sample t-statistic for each gene and each dataset. (5) We then assign a rank for each gene and each dataset, based on the t-statistic. (6) The overall rank for one gene is defined as the sum of ranks for all four datasets. The 1000 genes with the highest ranks are selected for downstream analysis.

We note that the proposed MTGDR has no limitation on the number of genes that can be used in the analysis. However previous empirical studies show that more reliable models can be obtained by excluding noisy genes prior to the analysis. Gene pre-processing is hence generally adopted. We refer to Ma (2006) for a more detailed discussion.

5.2 Individual TGDR analysis

As for the pancreatic cancer datasets, we first carry out logistic model + TGDR regularization analysis for each individual dataset. 27 (D1), 10 (D2), 20 (D3) and 6 (D4) genes are included in the final models, respectively, where the optimal tuning parameters are chosen via 3-fold cross validation. The gene sets identified are quite different. For example, the gene sets from datasets D1 and D2 have no overlap, while those from D1 and D3 have only one overlapped gene.

We further note that the genes selected from one dataset cannot be used for prediction in a different dataset. For example, if we use the genes selected from data D1 for prediction in datasets D2–D4, the prediction errors are 0.43 (D2), 0.24 (D3) and 0.35 (D4), which are rather unsatisfactory.

The inconsistency of gene discoveries and large prediction errors motivate us to consider combined analysis of four datasets.

5.3 Pooled TGDR analysis

In the second analysis, we simply pool four datasets together. The four liver datasets are generated in similar experimental settings and are expected to behave similarly. Using the TGDR for regularization, 34 genes are included in the final model. The LOO prediction error is 0.27 (34 subjects are misclassified). The prediction performance is improved comparing to the individual TGDR analysis.

Table 3. Liver cancer gene expression datasets used in the meta-analysis. # tumor: number of tumor samples. # normal: number of normal samples. Numbers in the "()" are the actual number of subjects used in the analysis. Ver. 2 chips have different spot location from Ver. 1 chips. They were printed using the same arrayer.

Dataset	D1	D2	D3	D4
Experimenter	Hospital A	Hospital B	Hospital C	Hospital C
# tumor	16 (14)	23	29	12 (10)
# normal	16 (14)	23	5	9(7)
Chip type	cDNA(Ver.1)	cDNA(Ver.1)	cDNA(Ver.1)	cDNA(Ver.2)
(Cy5:Cy3)	sample:normal liver	sample:placenta	sample:placenta	sample:sample

Table 4. Liver cancer datasets: genes with nonzero coefficients in MTGDR.

Gene Information	D1	D2	D3	D4
1.2.F.7/noseq/	-0.076	-0.100	-0.078	-0.035
1.3.A.8/clone MGC:5207 IMAGE:2901089	0.147	0.199	0.030	0.054
10.1.B.9/cDNA FLJ20844 fis, clone ADKA01904	-0.020	-0.016	-0.002	-0.002
11.3.F.6/noseq/	-0.275	-0.519	-0.225	-0.170
15.1.G.7/Cyt19 protein (Cyt19), mRNA	0.023	0.019	-0.001	0.009
15.2.D.10/EST387826 cDNA	-0.041	-0.031	-0.003	-0.015
15.3.E.9/hypothetical protein MGC11287	0.016	0.034	0.015	0.014
15.4.E.1/Rab9 effector p40 (RAB9P40), mRNA	0.166	0.243	-0.012	0.083
17.2.B.11/ATPase, H+ transporting, lysosomal 9kD	0.145	0.258	0.108	0.020
18.3.F.6/nomatch/	0.072	0.073	0.070	0.045
19.1.G.5/Ras association (RalGDS/	0.168	0.176	-0.036	0.042
2.2.E.11/triosephosphate isomerase 1 (TPI1), mRNA	0.012	0.012	0.004	0.011
2.2.G.10/UDP-glucose pyrophosphorylase 2 (UGP2)	-0.296	-0.274	-0.043	-0.178
21.3.A.4/noseq/	0.016	0.011	0.002	0.001
23.3.H.1/thioredoxin-like, 32kD (TXNL)	0.285	0.226	0.066	0.033
25.2.A.5/noseq/	0.016	0.014	0.001	0.009
26.2.D.2/adipose differentiation-related protein (ADFP)	-0.169	-0.114	-0.219	-0.118
26.4.B.5/Human zyxin related protein ZRP-1 mRNA	0.161	0.127	0.042	0.070
3.2.E.10/Human G protein-coupled receptor V28 mRNA	-0.707	-0.589	-0.359	-0.375
4.1.D.1/multiple endocrine neoplasia I (MEN1), mRNA	-0.086	-0.075	-0.130	-0.090
4.2.H.5/solute carrier family 22, member 1	-0.014	-0.120	-0.144	-0.092
4.3.C.1/noseq/	-0.058	-0.020	-0.008	0.007
4.4.B.9/noseq/	-0.438	-0.670	-0.460	-0.502
5.1.A.9/noseq/	-0.001	-0.007	-0.002	-0.001
5.1.D.1/malate dehydrogenase 2, NAD (mitochondrial)	0.135	0.043	0.063	0.060
6.2.E.3/tubulin, beta polypeptide (TUBB), mRNA /	0.024	0.012	0.004	0.011
6.3.B.3/noseq/	0.104	0.104	-0.023	0.015
6.4.D.11/non-metastatic cells 2, protein expressed NME2	0.053	0.072	0.020	0.025
6.4.F.5/H2A histone family, member Z (H2AFZ), mRNA	0.047	0.062	-0.001	0.042
7.3.A.5/nomatch/	-0.329	-0.432	-0.297	-0.222
7.3.G.9/guanine nucleotide binding protein, q polypeptide	0.073	0.019	0.049	0.029
8.2.B.11/cystatin B (stefin B) (CSTB), mRNA	0.040	0.112	0.051	0.046
8.2.D.8/RNA helicase-related protein (RNAHP), mRNA	-0.739	-1.369	-1.002	-1.140
8.3.A.7/proline-rich Gla polypeptide 2	-0.001	-0.019	-0.024	-0.026

5.4 MTGDR analysis

We analyze the liver cancer data using the proposed MTGDR, with optimal tuning parameters selected using the 3-fold cross validation. 34 genes have nonzero coefficients in the final models. We provide the gene information and corresponding estimates in Table 4. We can see that the characteristics of Table 4 and close to those in Table 2. However, we note that for some genes, the signs of the four estimates can be different. For example, for gene 15.4.E1/Rab9 effector p40, three out of four estimated coefficients are positive, and one is negative. As discussed above, different signs of estimates may indicate conflicting biological conclusions. However, we observe that the negative coefficient is very small. Similar small estimates are observed for other conflicting cases.

Prediction performance is evaluated using the LOO approach. There are 20 subjects that are misclassified, leading to a classification error of 0.16, which is a significant improvement over the pooled analysis.

6 DISCUSSIONS

Multiple pharmacogenetic studies have been carried out to detect predictive genomic biomarkers and construct predictive models. It is thus critical to develop microarray meta analysis methods that can effectively combine different datasets. In this article, we propose the MTGDR method for regularized meta analysis. This method can accommodate different platforms and experimental settings, and can lead to the same or similar biological conclusions across studies. Our analysis of pancreatic and liver cancer data indicates that parsimonious models with satisfactory prediction performance can be obtained using the proposed approach.

The main goal of this article is statistical methodology development, where we focus on constructing parsimonious predictive models with satisfactory prediction performance. The biological implications of the meta analysis results are not further pursued.

In our data analysis, the same sets of genes across studies are considered. When different sets of genes are included in different studies, the MTGDR is still applicable by setting gradients for missing genes zero. We note that meta analysis will be less powerful, if the sets of genes measured vary greatly.

We considered studies with binary outcome and the logistic regression model only. The MTGDR method is generally applicable, as long as the objective function $R(\beta)$ is well defined and differentiable. So the MTGDR can be applied to continuous outcomes including censored survival data. However, empirical studies need to be carried out to evaluate its performance with continuous outcomes.

ACKNOWLEDGMENT

We thank Dr. Robert Grutzman and Dr. Jung Kyoon Choi for providing the datasets.

REFERENCES

Choi, J., Choi, J., Kim, D., Choi, D., Kim, B., Lee, K., Yeom, Y. Yoo, H., Yoo, O. and Kim, S. (2004) Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Letters* **565**, 93-100.

Crnogorac-Jurcevic, T., Missiaglia, E., Blaveri, E., Gangeswaran, R., Jones, M., Terris, B., Costello, E., Neoptolemos, J.P. and Lemoine, N.R. (2003) Molecular alterations in pancreatic carcinoma: expression profiling shows that dysregulated expression of S100 genes is highly prevalent. *Journal of Pathology* **201**, 63-74.

Dettling, M. and Buhlmann, P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics* **9**, 1061-1069.

Friedman, J. and Popescu, B.E. (2004) Gradient directed regularization. *Technical Report, Stanford Department of Statistics.*

Friess, H., Ding, J., Kleeff, J., Fenkell, L., Rosinski, J.A., Guweidhi, A., Reidhaar-Olson, J.F., Korc, M., Hammer, J. and Buchler, M.W. (2003) Microarray-based identification of differentially expressed growth-and metastasis-associated genes in pancreatic cancer. *Cellular and Molecular Life Sciences* **60**, 1180-1199.

Fung, B. and Ng, V. (2004) Meta-classification of multi-type cancer gene expression data. *Proceeding of 4th Workshop on Data Mining in Bioinformatics*, 31-39.

Ghosh, D. and Choi, H. (2006) metaArray package for meta analysis of microarray data. *R package. http://www.r-project.org.*

Ghosh, D. and Chinnaiyan, A. (2004) Classification and selection of biomarkers in genomic data using LASSO. *Journal of Biomedicine and Biotechnology* **2**, 147-154.

Ghosh, D., Barette, T.R., Rhodes, D. and Chinnaiyan, A.M. (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional and Integrative Genomics* **3**, 180-188.

Grutzmann, R., Boriss, H., Ammerpoh, O., Luttges, J., Kalthoff, H., Schackert, H., Kloppel, G., Saeger, H. and Pilarsky, C. (2005) Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, 1-10.

Gui, J. and Li, H.Z. (2005a) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001-3008.

Gui, J. and Li, H. (2005b) Threshold gradient descent method for censored data regression, with applications in pharmacogenomics. *Pacific Symposium on Biocomputing* **10**, 272-283.

Hong, F., Breitling, R., McEntee, C.W., Witter, B.S., Nemhauser, J.L. and Chory, J. (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* **22**, 2825-2827.

Iacobuzio-Donahue, C.A., Ashfaq, R., Maitra, A., Adsay, N.V., Shen-Ong, G.L., Berg, K., Hollingsworth, M.A., Cameron, J.L., Yeo, C.J., Kern, S.E., Goggins, M. and Hruban, R.H. (2003) Highly expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization and comparison of the transcription profiles obtained from three major technologies. *Cancer Research* **63**, 8614-8622.

Jiang, H., Deng, Y., Chen, H., Tao, L., Sha, Q., Chen, J., Tsai, C. andZhang, S. (2004) Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* **5**: 81.

Jung, Y., Oh, M., Shin, D., Kang, S. and Oh, H. (2006) Identifying differentially expressed genes in meta-analysis via Bayesian model-based clustering. *Biometrical Journal* **48**, 435-450.

Kuo, W.P, Jenssen, T-K, Butte, A.J., Ohno-Machado, L. and Kohane, I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405412.

Li, H. (2007) Censored data regression in high-dimension and low sample size settings for genomic applications. *Statistical Advances in Biomedical Sciences: State of Art and Future Directions. Edited by A. Biswas, S. Datta, J. Fine and M. Segal, in press.*

Logsdon, C.D., Simeone, D.M., Binkley, C., Arumugam, T., Greenson, J., Giordano, T.J., Misek, D. and Hanash, S. (2003) Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer. *Cancer Research* **63**, 2649-2657.

Ma, S. (2006) Empirical study of supervised gene screening. *BMC Bioinformatics* **7**:537.

Ma, S. and Huang, J. (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21**, 4356-4362.

Ma, S. and Huang, J. (2007) Clustering threshold gradient descent regularization: with applications to microarray studies. *Bioinformatics* **23**, 466-472.

Ma, S., Song, X. and Huang, J. (2007) Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* **8**:60.

Stevens, J.R. and George, R.W. (2005) Meta-analysis combines Affymetrix microarray results across laboratories. *Comparative and Functional Genomics* **6**, 116-122.

Wang, J., Coombes, K.R., Highsmith, W.E., Keating, M.J. and Abruzzo, L.V. (2004) Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics* **17**, 3166-3178.

Warnet, P., Eils, R. and Brors, B. (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* **6**: 265.

Zhang, H., Ahn, J., Lin, X. and Park, C. (2006) Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22**, 88-95.