BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 10 2012, pages 1368–1375
doi:10.1093/bioinformatics/bts145

Genetics and population analysis Advance Access publication March 30, 2012

Penalized logistic regression for high-dimensional DNA methylation data with case-control studies

Hokeun Sun and Shuang Wang*

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: DNA methylation is a molecular modification of DNA that plays crucial roles in regulation of gene expression. Particularly, CpG rich regions are frequently hypermethylated in cancer tissues, but not methylated in normal tissues. However, there are not many methodological literatures of case-control association studies for high-dimensional DNA methylation data, compared with those of microarray gene expression. One key feature of DNA methylation data is a grouped structure among CpG sites from a gene that are possibly highly correlated. In this article, we proposed a penalized logistic regression model for correlated DNA methylation CpG sites within genes from high-dimensional array data. Our regularization procedure is based on a combination of the l_1 penalty and squared l_2 penalty on degree-scaled differences of coefficients of CpG sites within one gene, so it induces both sparsity and smoothness with respect to the correlated regression coefficients. We combined the penalized procedure with a stability selection procedure such that a selection probability of each regression coefficient was provided which helps us make a stable and confident selection of methylation CpG sites that are possibly truly associated with the outcome.

Results: Using simulation studies we demonstrated that the proposed procedure outperforms existing main-stream regularization methods such as lasso and elastic-net when data is correlated within a group. We also applied our method to identify important CpG sites and corresponding genes for ovarian cancer from over 20 000 CpGs generated from Illumina Infinium HumanMethylation27K Beadchip. Some genes identified are potentially associated with cancers.

Contact: sw2206@columbia.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 12, 2011; revised on February 29, 2012; accepted on March 22, 2012

1 INTRODUCTION

DNA methylation, which is the addition of a methyl group to the 5′ position of cytosine in the context of a CpG dinucleotide, is a molecular modification of DNA that plays crucial roles in regulation of gene expression. Particularly, CpG rich regions are frequently hypermethylated in cancer tissues, but not methylated in normal tissues. Tremendous amounts of DNA methylation data have recently been generated from high-throughput DNA methylation platforms. For example, the Illumina GoldenGate array, the Illumina Infinium HumanMethylation27K array and the most recent Illumina Infinium HumanMethylation450K array are popularly used. These platforms are based on genotying bisulfite converted DNA. The results of the array, the methylation status of the interrogated CpG site are a sequence of β-values, one for each locus, calculated as the average of approximately 30 replicates of the quantity (Bibikova *et al.*, 2006),

$$\beta = \frac{\max(M,0)}{\max(U,0)+\max(M,0)+100},$$

where U is the fluorescent signal from an unmethylated allele on a single bead, and M is that from a methylated allele. A maximum between signal intensity and 0 is chosen to compensate for negative signals due to background subtraction. The constant 100 is to regularize β-values where both M and U values are small. This β-value ranges continuously from 0 (unmethylated) to 1 (completely methylated) and reflects the methylation level of each CpG site.

Many researchers have applied statistical classification methods to select differently methylated loci (Houseman *et al.*, 2008; Kuan *et al.*, 2010; Siegmund *et al.*, 2004). Statistical approaches developed for gene expression data may not be applied directly to methylation data since many genes are methylated, while only a few genes are differently expressed. However, disease related CpG regions should still be sparse, in which case the problem is equivalent to identify a few relevant genes from high-dimensional gene expression data. Compared with the case-control studies of gene expression data, there are currently not many methodological developments for methylation data. Wang (2011) has recently proposed a likelihood-based uniform–normal mixture model to select differently methylated loci between case and control groups. One difference of methylation data from gene expression is that the former ranges between 0 and 1. But, this is not an issue in regression frameworks. Another key feature of DNA methylation data that has not been fully utilized is the group structure within a gene. With the Illumina HumanMethylation27K array, there are about 1–22 CpG sites per gene where methylation levels of CpG sites within a gene are usually correlated. Unlike genotype data with single nucleotide polymorphisms (SNPs), β-values $\in (0,1)$ are continuous thus correlations among them can be observed more precisely. Note that although CpG sites within a gene are correlated, some CpG sites might be neutral while some CpG sites might be causal. Therefore, considering these features of DNA methylation data, in this article we proposed a penalized logistic regression model for correlated CpG sites within a gene as predictors of a disease status. The proposed method can select CpG sites individually that associated with a disease while grouping effects are encouraged.

The penalized logistic regression has been recently used to select SNPs or genes associated with a disease in high-dimensional

*To whom correspondence should be addressed.

data settings. It has great advantage in performing consistent variable selection even when the number of predictors are much larger than number of samples, a typical setting in genetics and genomics. Hierarchically penalized selection method was first proposed by Breheny and Huang (2009). They developed a selection procedure taking into account both between-group and within-group features along with different penalties. Wu *et al.* (2009) evaluated the performance of lasso penalized logistic regression in case-control disease gene mapping with SNP data. Zhou *et al.* (2010) proposed a model that uses the mixture of group and lasso penalties in logistic regression to identify common and rare variants in genome-wide association studies (GWAS). Their approaches basically combine the l_1 norm penalty of Tibshirani (1996) and the l_2 norm group penalty of Meier *et al.* (2008). These penalties induce the overall sparsity and group sparsity, respectively. Consequently, predictors with stronger effects on responses are more likely to be selected into the model. However, none of the methods above are designed to encourage selection effects for highly correlated predictors.

For the analysis of correlated microarray gene expression data, several variable selection methods have been developed in the past decade. The fused lasso (Tibshirani *et al.*, 2005) imposes the l_1 penalty on the absolute differences of the regression coefficients in order to account for some smoothness of the coefficients. The elastic-net (Enet) procedure (Zou and Hastie, 2005), which is a compromise between a ridge regression penalty and a lasso penalty, encourages a grouping effect of highly correlated variables. When genetic network information such as a genetic pathway is available, Li and Li (2008, 2010) proposed a graph-constrained regularization procedure. In their method, a Laplacian matrix representing a graph structure was imposed on a ridge-regression penalty of the Enet procedure such that linked genes have a smoothness penalty on their regression coefficients. The authors demonstrated that when the information of gene networks is incorporated into the regularization procedure, it can select more relevant genes than the lasso and Enet procedures.

Motivated by the graph-constrained procedure by Li and Li (2008, 2010), we extended it to the logistic regression model for the analysis of case-control DNA methylation data, where our Laplacian matrix represents CpG sites clustered within genes. The rest of this article is organized as follows. In Section 2, we described our statistical model and regularization procedure for grouped and correlated predictors (CpG sites) together with the computational algorithm and the stability selection procedure. We then presented the simulation results in Section 3, where existing main-stream variable selection methods were compared with the proposed method when predictors within a group are correlated. Section 4 analyzed a real DNA methylation data in a case-control study of ovarian cancer. Finally, we gave a brief discussion of the method and future research in Section 5.

2 METHODS

2.1 Penalized logistic regression

Let us denote the methylation β-values of the i-th individual by $x_i = (x_{i1},\ldots,x_{ip})^{\mathrm{T}}$, $i=1,\ldots,n$, and p is the total number of CpG sites considered in the analysis. The penalized logistic regression is defined as

$$-\frac{1}{n}\sum_{i=1}^{n}[y_i \log p(x_i)+(1-y_i)\log(1-p(x_i)]+P(\theta), \qquad (1)$$

where $P(\theta)$ is a penalty function, and the response y_i is 0 for controls and 1 for cases. The probability that the i-th individual is a case based on his/her DNA methylation information is denoted as

$$p(x_i)=\frac{\exp(\theta_0+x_i^{\mathrm{T}}\theta)}{1+\exp(\theta_0+x_i^{\mathrm{T}}\theta)}.$$

The intercept parameter θ_0 and regression coefficients $\theta=(\theta_1,\ldots,\theta_p)^{\mathrm{T}}$ can be estimated by minimizing the objective function (1).

In the work of Li and Li (2008, 2010), the p-dimensional Laplacian matrix $L=\{l_{uv}\}$ was used to represent a graph structure when the network information of predictors is provided. It is defined as

$$l_{uv}=\begin{cases}1 & \text{if} \quad u=v \text{ and } d_u\neq 0\\ -(d_u d_v)^{-\frac{1}{2}} & \text{if} \quad u \text{ and } v \text{ are linked with each other}\\ 0 & \text{otherwise,}\end{cases}$$

where d_u is the total number of links of the u-th predictor, and it is often called a degree of the vertex u in graph theory. Their penalty function is

$$\begin{aligned}P(\theta)&=\lambda_1\|\theta\|_1+\lambda_2\theta^{\mathrm{T}}L\theta\\ &=\lambda_1\sum_{j=1}^{p}|\theta_j|+\lambda_2\sum_{u=1}^{p}\sum_{u\sim v}\left(\frac{\theta_u}{\sqrt{d_u}}-\frac{\theta_v}{\sqrt{d_v}}\right)^2\end{aligned} \qquad (2)$$

where $\|\cdot\|_1$ is a l_1 norm, and $u\sim v$ indicates the index set of all linked variables to the u-th predictor. The tuning parameters λ_1 and λ_2 control the amount of regularization for sparsity and smoothness, respectively. When $\lambda_2=0$, this penalty simply reduces to the lasso (Tibshirani, 1996), and if the Laplacian matrix L is replaced by an identity matrix I, the penalty becomes the Enet penalty (Zou and Hastie, 2005). Also, the estimates of ridge regression for logistic regression can be obtained when $\lambda_1=0$ and $L=I$.

This penalty is defined as a combination of the l_1 penalty and squared l_2 penalty on degree-scaled differences of coefficients between linked predictors. It induces both sparsity and smoothness with respect to the correlated or linked structure of the regression coefficients. Li and Li (2008, 2010) have shown that a desirable grouping effect can be reached by specifying links among regression coefficients in the model. However, their approach is limited to the ordinary regression model.

Here we extended it to the penalized logistic regression for the analysis of case-control methylation data. First, we need to specify a network structure that describes correlation patterns for methylation measures of CpG sites within a gene. There are two networks prevalently used in graph theory that fit our situation, namely, the ring network and the fully connected (F.con) network. Figure 1 depicts these network topologies in the scenario when there are 6 genes and each of which consists of 1, 2, 3, 4, 6 and 9 CpG sites. In the ring network, we assumed that the first and last CpG sites from a gene are connected with each other so that all CpG sites within the same gene have the same number of links. In contrast, a F.con network assumes that all CpG sites within the same gene are connected with each other. In the penalized regression model, both networks specify a group structure of predictors so that the coefficients of correlated predictors within the same group can shrink toward each other, allowing them to borrow information

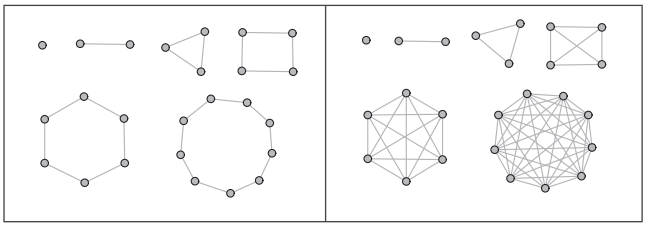

Fig. 1. The ring network (left) and F.con network (right) are shown when there are 6 genes each of which consists of 1, 2, 3, 4, 6 and 9 CpG sites.

from each other. Unlike the group penalized method of Meier *et al.* (2008), our model still performs individual selection, so neutral CpG sites within significant genes are not forced to remain in the final model.

To compare the biological basis of the ring and the F.con networks, although it depends on the underlying true correlation patterns of CpG sites within genes, the later might be more appropriate for DNA methylation data since one CpG site might be correlated with the rest sites within the same gene. The ring network assumes correlation of just flanking sites except for the first and last sites. Note that the link between the first and the last sites in the ring network is less possible biologically but is imposed to apply the ring network. In terms of variable selection, the ring network will induce a mild groping effect on a large group since the number of links per CpG site is fixed at 2 for all genes that have >2 CpG sites. In contrast, the total number of links per CpG site in the F.con network model is increased by the number of CpG sites within a gene. So, the fully connected network usually produces a strong grouping effect since variables with more links are more likely to be selected in the network-based regularization procedure. This is also discussed by Li and Li (2010). If all CpG sites in a gene are true signals, the F.con network are desirable. But, if there are both neutral sites and causal sites in a gene, the ring network might be preferable to weaken the grouping effect. However, in simulation study and real data application we showed that the selection results using the two networks are almost identical. Also note that the Laplacian matrix in our model (2) has the form of a blockwise diagonal matrix for both network models.

2.2 Computational algorithms

Li and Li (2010) has applied the coordinate descent algorithm of Friedman *et al.* (2007) to obtain the minimizer of the function (1) when a response variable follows a Gaussian distribution. But this cannot be directly applied to the logistic regression. Moreover, recent publication of Friedman *et al.* (2010) has drastically improved the computational efficiency, and a `R` package `glmnet` was developed for the Enet regularization procedure of the logistic regression. Since the Enet method simply replaces the Laplacian matrix by an identity matrix in the penalty function (2), we could impose the Laplacian matrix in the `glmnet` code and be benefited from the efficient computation algorithm already implemented. In this section, we briefly explained how the Enet procedure and our method differ in terms of algorithm.

Let us denote the negative log likelihood function of the logistic regression model by $-l(\theta_0,\theta)$, which is equivalent to the first term of the function (1). Our objective function is then

$$Q(\theta_0,\theta)=-l(\theta_0,\theta)+P(\theta), \tag{3}$$

where

$$P(\theta)=\lambda\alpha\sum_{i=1}^{p}|\theta_i|+\frac{1}{2}\lambda(1-\alpha)\sum_{u=1}^{p}\sum_{u\sim v}\left(\frac{\theta_u}{\sqrt{d_u}}-\frac{\theta_v}{\sqrt{d_v}}\right)^2, \tag{4}$$

and

$$\lambda=\lambda_1+2\lambda_2 \qquad \text{and} \qquad \alpha=\frac{\lambda_1}{\lambda_1+2\lambda_2}.$$

As $l(\theta_0,\theta)$ is approximated by a second-order Taylor series expansion at current estimates (θ_0^*,θ^*), the function (3) can be re-written as

$$Q^*(\theta_0,\theta)=-\frac{1}{2n}\sum_{i=1}^{n}w_i(z_i-\theta_0-x_i^{\mathrm{T}}\theta)^2+P(\theta),$$

where

$$z_i=\theta_0^*+x_i^{\mathrm{T}}\theta^*+w_i^{-1}(y_i-p^*(x_i)),$$
$$w_i=p^*(x_i)(1-p^*(x_i)),$$
$$p^*(x_i)=1-[1+\exp(\theta_0^*+x_i^{\mathrm{T}}\theta^*)]^{-1}.$$

We refer the readers to Friedman *et al.* (2010) for the details of this derivation.

Next, we can compute the gradient at $\theta_u=\theta_u^*$ while the other estimates θ_v^* for all $v\neq u$ are fixed. By setting the gradient to 0 and solve for θ_u, we can get the current estimate of θ_u^* using the following formula,

$$\theta_u^*=\frac{S\left(n^{-1}\sum_{i=1}^{n}w_ix_{iu}(z_i-\tilde{z}_i^{(u)})+\lambda(1-\alpha)g(u),\lambda\alpha\right)}{n^{-1}\sum_{i=1}^{n}w_ix_{iu}^2+\lambda(1-\alpha)},$$

where $\tilde{z}_i^{(u)}=\theta_0^*+\sum_{j\neq u}x_{ij}\theta_j^*$,

$$g(u)=\sum_{u\sim v}\frac{\theta_v^*}{\sqrt{d_ud_v}}, \tag{5}$$

and $S(z,\gamma)$ is a soft thresholding operator with value

$$\mathrm{sign}(z)\left(|z|-\gamma\right)_+=\begin{cases}z-\gamma & \text{if } z>0 \text{ and } \gamma<|z|\\ z+\gamma & \text{if } z<0 \text{ and } \gamma<|z|\\ 0 & \text{otherwise.}\end{cases}$$

If the u-th predictor is a unique member of a group, i.e. no links to any other predictors, the function $g(u)$ in (5) is equal to 0. If all predictors are isolated, the penalty function (4) is reduced to

$$P(\theta)=\lambda\alpha\sum_{i=1}^{p}|\theta_i|+\frac{1}{2}\lambda(1-\alpha)\sum_{u=1}^{p}\theta_u^2,$$

and thus $g(u)=0$ for all $u=1,\ldots,p$. In this case the solution becomes exactly the same as that of the Enet procedure. In other words, we only need to include the term $\lambda(1-\alpha)g(u)$ in the formula of the Enet when we create a group of CpG sites by genes.

2.3 Selection probabilities

In our penalty function (4) we have two tuning parameters to be selected. The parameter $\alpha\in(0,1)$ induces the model to have the estimates between ridge ($\alpha=0$) and lasso ($\alpha=1$). As α increases from 0 to 1, the sparsity solution to (3) is obtained for each λ, but the model near $\alpha=1$ becomes indifferent to correlated predictors. The Enet procedure usually uses a small value of α to overcome collinearity problems in regressions with high-dimensional predictors. However, results of variable selection are very similar with a small perturbation of α. In contrast, the tuning parameter λ forces the model to select more variables as λ decreases with a fixed α. Since the value of λ directly determines the total number of relevant predictors, the choice of λ is crucial in variable selection problems.

Cross-validation (CV) is commonly used to find the optimal value of λ in variable selection literatures, although it is known that for high-dimensional data CV generally selects too many variables including some truly unrelated variables. Meinshausen and Bühlmann (2010) in their recent work proposed to compute selection probabilities for all variables, and to include only the variables with high selection probabilities in the model. Their method is based on resampling, and provides a more stable selection for high-dimensional data compared with CV. Recently, Alexander and Lange (2011) have applied it to select relevant SNPs in GWAS. We employed their method to determine important CpG sites in our regularization procedure.

Let us denote Λ as a regularization parameter space, e.g. $\lambda\in\Lambda$. Let I_k be the k-th random subsample of $\{1,\ldots,n\}$ of size $\lfloor n/2\rfloor$ without replacement, where $\lfloor x\rfloor$ is the largest integer not greater than x. The selection probability of the u-th predictor is then defined as

$$\mathrm{SP}(u)=\max_{\lambda\in\Lambda}\frac{1}{K}\#\{k\leq K:\hat{\theta}_u^{\lambda}(I_k)\neq 0\},$$

where $\hat{\theta}_u^{\lambda}(I_k)$ is the estimator of θ_u using a regularization procedure based on the subsample I_k given λ, and K is the total number of resampling. Then, the variables whose selection probabilities are greater than some cutoff value, say π, are selected in the model. According to Meinshausen and Bühlmann (2010), selection results do not rely much on the choice of Λ, and tend to be very similar with different values of the cutoff π. They also provided the formula to select both Λ and π so that the expected number of falsely

selected variables can be controlled when the exchangeability assumption of predictors are met. However, this assumption is a weaker version of independence and is hard to be satisfied with highly correlated genomic data. We thus focused on identifying the CpG sites which have high selection probabilities in the analysis of methylation data.

2.4 Adaptive regularization procedure

Li and Li (2010) have pointed out that their procedure does not perform well when two predictors that are linked with each other are negatively correlated with the response because in this situation the corresponding regression coefficients have different signs, so they are not expected to be locally smooth. Their solution to this problem is to estimate the signs of coefficients first, and refit the model with the estimated signs. This change only affects the Laplacian matrix of the penalty function, so we can simply modify it in the following way,

$$l_{uv} = \begin{cases} 1 & \text{if} \quad u = v \text{ and } d_u \neq 0 \\ -s_u s_v (d_u d_v)^{-\frac{1}{2}} & \text{if} \quad u \text{ and } v \text{ are linked each other} \\ 0 & \text{otherwise,} \end{cases}$$

where s_u is the estimated sign of the u-th predictor, which can be obtained by ordinary regression estimates when $p < n$, and ridge estimates for $p \geq n$. Then, the function $g(u)$ in (5) should be updated in the formula,

$$g(u) = \sum_{u \sim v} \frac{s_u s_v \theta_v^*}{\sqrt{d_u d_v}}.$$

Li and Li (2010) have demonstrated that this adaptive regularization procedure leads to improved estimates and selection results compared with the ordinary procedure in their simulation studies. Therefore, we also have included the sign estimates in our regularized estimates for both simulation studies and the real data application.

3 SIMULATION STUDIES

We conducted some simulation studies to compare the performance of our proposed method to existing main-stream regularization procedures. In our simulation models, we have 600 genes which consist of 1–9 CpG sites such that 100 genes have 1 CpG site, 150 genes have 2 CpG sites and each 50 genes of the others have 3–9 CpG sites, respectively. Therefore, a total of 2500 CpG sites were simulated.

To mimic real DNA methylation data, we need to generate group correlated variables within the $[0, 1]$ range. To do so, we employed the inverse logit transformation of multivariate normal random variables. Thus, the methylation β-values of the g-th gene for the i-th individual is calculated as

$$x_{i,g} = \frac{\exp(t_{i,g})}{1 + \exp(t_{i,g})}, \quad \text{and} \quad t_{i,g} \sim \sqrt{s} N_{p_g}(\mu, \Sigma),$$

where s is a scale parameter and p_g is the size of the g-th gene, i.e. $1 \leq p_g \leq 9$. In this study, we set $\mu = (-0.1, \ldots, -0.1)^{\mathrm{T}}$ and $s = 4$ so that the distribution of the methylation values have an enriched '0' (unmethylated) and enriched '1' (completely methylated) as previously observed (Wang, 2011).

Next, we specified the true regression coefficients θ similar as Li and Li (2010). Let us first denote the coefficients of the g-th gene by $\theta_{(g)} = (\theta_{1,g}, \ldots, \theta_{p_g,g})^{\mathrm{T}}$. Since 600 genes have 1 to 9 CpG sites, we grouped the genes by the number of CpG sites. For example, the first gene group has only one CpG site, and the second group has only two CpG sites, and so on. Then, we selected one gene from each of the 9 different gene groups, which leads to a total of 45 CpG sites. We denoted the regression coefficients of these 45 CpG sites as

$$\theta_{k,g} = (-1)^{p_g + 1} \frac{\delta}{\sqrt{p_g}}, \quad \text{for all } k = 1, \ldots, p_g,$$

for $g = 1, \ldots, 9$ and referred them as CpG set-1. Similarly, we selected another nine genes from each of the nine different gene groups, and denoted their coefficients as

$$\theta_{k,g} = (-1)^{p_g} \frac{\delta}{\sqrt{p_g}}, \quad \text{for all } k = 1, \ldots, \lceil \frac{p_g}{2} \rceil,$$

where $\lceil x \rceil$ is the smallest integer not less than x. We referred them as CpG set-2. All the other θ's are set to 0. Therefore, in all simulation settings, all CpG sites in CpG set-1 are disease related sites, while only half of the CpG sites in CpG set-2 are disease related sites. Thus, there are a total of 70 disease related CpG sites out of 2500 total CpG sites in the simulation models.

Finally, the corresponding responses y_i was simulated according to a Bernoulli distribution with the following model-based probabilities,

$$y_i \sim \text{Bernoulli}\big(p(x_i)\big), \qquad p(x_i) = \frac{\exp(x_i^{\mathrm{T}} \theta)}{1 + \exp(x_i^{\mathrm{T}} \theta)},$$

where $x_i = (x_{i,1}^{\mathrm{T}}, \ldots, x_{i,600}^{\mathrm{T}})^{\mathrm{T}}$ and $\theta = (\theta_{(1)}^{\mathrm{T}}, \ldots, \theta_{(600)}^{\mathrm{T}})^{\mathrm{T}}$. For each simulation set, samples were generated until we have 200 cases. We then randomly selected 200 controls from the control pool already generated.

We considered nine different simulation models, differing the strength of the true signals δ and the covariance matrix Σ within genes. First, three models are defined in the following way,

(1) $\delta = 1$, and $\Sigma_{uv} = \rho^{|u-v|}$

(2) $\delta = 2$, and $\Sigma_{uv} = \rho^{|u-v|}$

(3) $\delta = 2$, and $\Sigma_{uv} = \rho$ for $u \neq v$ and $\Sigma_{uv} = 1$ for $u = v$,

where the correlation of the first two models is AR(1), and the third model has compound symmetric correlation structure. We then simulated the data with different correlation coefficient $\rho = 0.2, 0.5$ and 0.7 for all three models. For each model we repeated simulations 100 times, and selection probabilities for each simulation set was computed based on 100 resamplings.

We compared the performance of the proposed ring and F.con network-based method to that of the group MCP (gMCP) procedure using the `R` package `grpreg` (Breheny and Huang, 2009), the lasso and Enet procedures using `glmnet` (Friedman *et al.*, 2010). Figure 2 shows the box plots of averaged selection probabilities of 70 true signals in 9 different simulation models. It appears that the lasso has the lowest selection probabilities, and the Enet and our methods have similar selection probabilities through all simulation models. The selection probabilities of gMCP are slightly lower than that of the Enet but much higher than that of the lasso. However, the model with the highest selection probabilities does not always lead to the best model, when the model might have selected too many false positives if the selection probabilities of the disease unrelated predictors are also high. Thus, we computed both true positive rates and false positive rates of each procedure while varying the cutoff π of selection probabilities from 0 to 1.

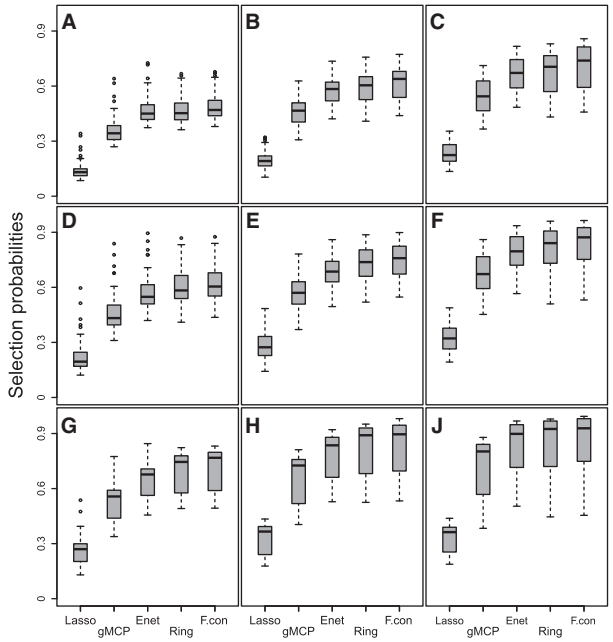

Fig. 2. The averaged selection probabilities of 70 true signals via Lasso, gMCP, Enet, and the ring and F.con network-based procedures are present. The signal strength is set at $\delta = 1$ in **A–C**, and $\delta = 2$ in **D–J**. The AR(1) covariance is used in **A–F**, and the compound symmetric correlation for **G–J**. The correlation is set at $\rho = 0.2$ for **A, D** and **G**, $\rho = 0.5$ for **B, E** and **H**, and $\rho = 0.7$ for the others.

Table 1. The area under the averaged ROC curves of Lasso, gMCP, Enet, and the ring and F.con network-based procedures along with the different signal strength δ, covariance Σ and correlation coefficient ρ

δ	Σ	ρ	Lasso	gMCP	Enet	Ring	F.con
1	AR(1)	0.2	0.6896	0.6980	0.7000	0.7472	0.7470
1	AR(1)	0.5	0.7707	0.7963	0.8019	0.8538	0.8562
1	AR(1)	0.7	0.8202	0.8632	0.8728	0.9083	0.9127
2	AR(1)	0.2	0.7943	0.8012	0.8023	0.8574	0.8582
2	AR(1)	0.5	0.8619	0.8831	0.8862	0.9310	0.9321
2	AR(1)	0.7	0.9037	0.9360	0.9410	0.9655	0.9669
2	CS[a]	0.2	0.8468	0.8592	0.8615	0.9113	0.9124
2	CS	0.5	0.9010	0.9279	0.9328	0.9583	0.9587
2	CS	0.7	0.9088	0.9490	0.9549	0.9700	0.9707

[a]Compound symmetry covariance.

The averaged receiver operating characteristic (ROC) curves of selection results of the procedures are given in Supplementary Materials (Fig. S1). The corresponding area under the ROC curves (AUC) are shown in Table 1. It is obvious that our proposed network-based methods outperform other procedures in all simulation scenarios. Surprisingly, both versions of the proposed network-based methods (Ring and F.con) have very similar AUCs and almost identical ROC curves in all simulation models. This may suggest that the penalty of misidentifying correlation structures of CpG sites within a gene is negligible, but instead the group selection effects of clustering CpG sites is strong enough to enhance overall selection performance and overwhelm the others.

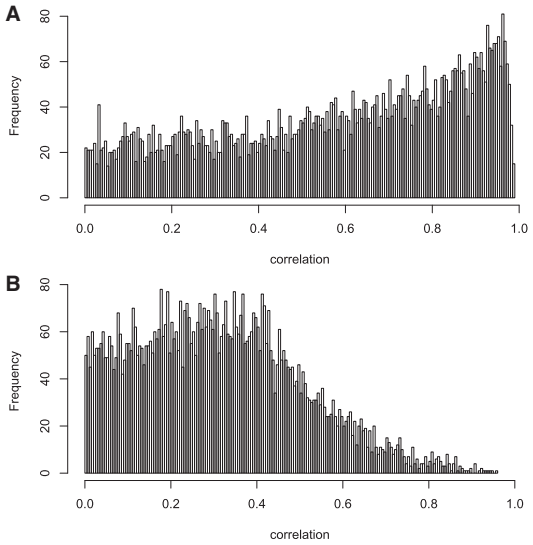

Fig. 3. The histograms of maximum sample correlation between CpG sites within 6936 genes for (**A**) pre-treatment case and healthy control group combined, and (**B**) post-treatment case and healthy control group combined

The null simulation (with $\delta = 0$) were also conducted for these five procedures for validation. The results are provided in Supplementary Materials (Figs S2 and S3), where averaged selection probabilities over 100 simulation replications are displayed for all 2500 CpG sites. It can be seen that all procedures produced roughly uniform selection probabilities across all CpG sites. Note that the comparison of selection probabilities across the five procedures in the null simulation is not meaningful, but we should compare selection probabilities across CpG sites.

Since the lasso procedure handles neither group structure nor correlated predictors, their selection performance is always poor for group correlated predictors. The gMCP is better than the lasso as it specifies grouped variables but does not differentiate their correlations. In contrast, the Enet is designed to have grouping effects for correlated predictors without specifying defined groups. Only our proposed methods account for both group structure and correlated variables thus lead to a superior selection performance over the other regularization procedures.

4 DATA ANALYSIS

We applied our proposed method to select differentially methylated CpG sites between ovarian cancer cases and healthy controls using the DNA methylation data generated from Illumina Infinium HumanMethylation27K Beadchip (Teschendorff *et al.*, 2010). The data is available at the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/).

We first performed the quality control procedure of the methylation data similar as Teschendorff *et al.* (2010) and Wang (2011), which include the removal of samples with either a low BS conversion efficiency or low CpG coverage. We also left out CpG sites with any missing β-values. Since our procedure assumes that every CpG site belongs to a single gene, CpG sites without corresponding gene information were removed. We ended up with 20 461 CpG sites from 12 770 genes where we have 152 controls, 119

Table 2. The CpG sites and corresponding genes with top 20 selection probabilities identified by Enet, and the ring and F.con network-based procedures from the comparison between pre-treatment cases and normal controls

	Enet			Ring			F.con		
	Prob	IlmnID	Gene	Prob	IlmnID	Gene	Prob	IlmnID	Gene
1	0.990	cg11009736	(MARCO)[a]	0.980	cg20792833	(PTPRCAP)[a]	0.980	cg20792833	(PTPRCAP)[a]
2	0.990	cg04988978	(MPO)[a]	0.925	cg04988978	(MPO)[a]	0.925	cg04988978	(MPO)[a]
3	0.985	cg20792833	(PTPRCAP)[a]	0.895	cg09964921	(KCNE1)[a]	0.895	cg09964921	(KCNE1)[a]
4	0.950	cg09964921	(KCNE1)[a]	0.890	cg11009736	(MARCO)[a]	0.890	cg11009736	(MARCO)[a]
5	0.920	cg06521852	(HRIHFB2122)[a]	0.810	cg14360917	(SP2)[a]	0.810	cg14360917	(SP2)[a]
6	0.920	cg00134539	(UBASH3A)[a]	0.790	cg06521852	(HRIHFB2122)[a]	0.795	cg03801286	(KCNE1)
7	0.890	cg14360917	(SP2)[a]	0.790	cg03801286	(KCNE1)	0.790	cg06521852	(HRIHFB2122)[a]
8	0.885	cg21932814	(CSTA)	0.790	cg21517055	(MGC11271)[a]	0.790	cg21517055	(MGC11271)[a]
9	0.885	cg00974864	(FCGR3B)	0.745	cg00201234	(FBLN2)[a]	0.745	cg00201234	(FBLN2)[a]
10	0.885	cg21517055	(MGC11271)[a]	0.740	cg00134539	(UBASH3A)[a]	0.740	cg00134539	(UBASH3A)[a]
11	0.885	cg02374486	(PRF1)	0.730	cg15616083	(KCNQ2)	0.730	cg15616083	(KCNQ2)
12	0.885	cg20070090	(S100A8)	0.725	cg27303882	(PAGE2)	0.725	cg27303882	(PAGE2)
13	0.845	cg27067618	(CYP4F3)	0.720	cg05294455	(MYL4)[a]	0.715	cg05294455	(MYL4)[a]
14	0.845	cg04353769	(MS4A6A)	0.690	cg11412582	(HERC2)	0.685	cg13626881	(ADORA1)
15	0.830	cg02240622	(PLCB2)	0.680	cg13626881	(ADORA1)	0.685	cg11412582	(HERC2)
16	0.830	cg06196379	(TREM1)	0.680	cg01405107	(HOXB5)	0.680	cg01405107	(HOXB5)
17	0.825	cg21126943	(CEACAM6)	0.670	cg04398282	(BRDG1)	0.665	cg10494770	(IGLL1)
18	0.820	cg00201234	(FBLN2)[a]	0.670	cg24993443	(SNRPN)	0.665	cg24993443	(SNRPN)
19	0.820	cg27461196	(FXYD1)	0.665	cg10494770	(IGLL1)	0.660	cg04398282	(BRDG1)
20	0.815	cg05294455	(MYL4)[a]	0.650	cg27067618	(CYP4F3)	0.650	cg06409153	(ABCA5)
20	–	–	–	–	–	–	0.650	cg27067618	(CYP4F3)

[a]Indicates the overlapped genes in the top 20 lists of all 5 procedures.

pre-treatment cases, and 123 post-treatment cases. We chose to work with more homogenous comparisons by separating pre-treatment and post-treatment cases. Among 12 770 genes, 5834 genes have one CpG site, and 6744 genes have two CpG sites. Only 169 genes include 3 to 9 CpG sites, and 23 genes have 10 to 22 CpG sites.

Next, we investigated the sample correlations between CpG sites of 6936 genes that have at least 2 CpG sites. The first comparisons is between 152 controls and 119 pre-treatment cases, and second comparison is between 152 controls and 123 post-treatment cases. The sample correlation matrix was computed gene by gene. We took the maximum correlation from all pairwise correlations among CpG sites on genes with >2 CpG sites and plotted the histograms of these correlations over 6936 genes for comparisons (Fig. 3). Although the distributions of the correlations are quite different between two comparisons, it is clear that methylation β-values within genes are highly correlated. The correlation histograms of separate pre-treatment cases, post-treatment cases, and controls are also given in Supplementary Materials (Fig. S4).

To identify the important CpG sites and corresponding genes for ovarian cancer, we applied the proposed two versions of the regularization procedures (Ring and F.con) and three existing ones (Lasso, gMCP and Enet). The selection results of the pre-treatment and post-treatment comparisons are reported in Tables 2 and 3, respectively. In each table, the CpG sites and genes with top 20 highest selection probabilities for the Enet, the ring and F.con network-based procedures were listed. The selection results of Lasso and gMCP are given in Supplementary Materials (Tables S1 and S2). The selection probabilities of each procedure were computed based on 200 resampled subsets of individuals.

It appears that although the five procedures selected somewhat different CpG sites in their top 20 list, around 10 CpG sites overlapped. It is also expected that the ring and F.con network-based methods identified almost the same lists of CpG sites as simulation studies suggested. In Table 2 when comparing pre-treatment cases and controls, gene HOXB5, which was selected by the proposed procedure was previously identified to be differentially methylated between liver cancer tumor tissues and adjacent normal tissues (Shen *et al.*, 2011), but Enet failed to select HOXB5 into its top 20 list. The gene MPO is known to be related to lung cancer risk (London *et al.*, 1997) and all procedures identified one CpG site from this gene. In Table 3 when comparing post-treatment cases and controls, DNA methylation level of the gene EGF was previously reported to be related to head and neck cancer (Marsit *et al.*, 2009) and EGF was identified by all five procedures. Gene TNFAIP8, which is known to act as a negative mediator of apoptosis and may play a role in tumor progression (http://www.ncbi.nlm.nih.gov/gene/25816) was only selected by the proposed procedures but not the Enet procedure.

5 DISCUSSION

In this article, we proposed a penalized logistic regression model for correlated predictors within a group and applied it to high-dimensional methylation data. In simulation studies we have demonstrated that the proposed procedure outperforms existing main-stream regularization methods such as lasso and Enet when data is correlated within a group. We also identified important CpG sites and corresponding genes for ovarian cancer from over 20 000 CpGs using the Illumina Infinium Human Methylation27K

Table 3. The CpG sites and corresponding genes with top 20 selection probabilities identified by Enet, and the ring and F.con network-based procedures from the comparison between post-treatment cases and normal controls

	Enet			Ring			F.con		
	Prob	IlmnID	Gene	Prob	IlmnID	Gene	Prob	IlmnID	Gene
1	1.000	cg23580000	(ADCY7)	1.000	cg06653796	(LIME1)[a]	1.000	cg06653796	(LIME1)[a]
2	1.000	cg06653796	(LIME1)[a]	1.000	cg10986043	(TCAP)[a]	1.000	cg10986043	(TCAP)[a]
3	1.000	cg10986043	(TCAP)[a]	0.975	cg23580000	(ADCY7)	0.975	cg23580000	(ADCY7)
4	0.980	cg13379236	(EGF)[a]	0.950	cg13379236	(EGF)[a]	0.955	cg13379236	(EGF)[a]
5	0.980	cg03547797	(GAS2)[a]	0.940	cg03547797	(GAS2)[a]	0.940	cg03547797	(GAS2)[a]
6	0.970	cg05135288	(RHOT2)[a]	0.935	cg05135288	(RHOT2)[a]	0.935	cg05135288	(RHOT2)[a]
7	0.965	cg20357806	(PPBP)[a]	0.870	cg12006284	(WT1)[a]	0.870	cg12006284	(WT1)[a]
8	0.965	cg12006284	(WT1)[a]	0.865	cg20357806	(PPBP)[a]	0.860	cg20357806	(PPBP)[a]
9	0.905	cg21640749	(CD300LF)[a]	0.840	cg24335895	(COX7A1)[a]	0.840	cg24335895	(COX7A1)[a]
10	0.900	cg12243271	(CFI)	0.810	cg21640749	(CD300LF)[a]	0.815	cg21640749	(CD300LF)[a]
11	0.890	cg09626634	(EBI2)	0.805	cg12243271	(CFI)	0.815	cg12243271	(CFI)
12	0.885	cg22988566	(WFDC10B)	0.800	cg19573166	(SLC22A17)	0.805	cg10467098	(Bles03)[a]
13	0.880	cg24335895	(COX7A1)[a]	0.795	cg10467098	(Bles03)[a]	0.800	cg19573166	(SLC22A17)
14	0.880	cg19573166	(SLC22A17)	0.785	cg15096140	(MYO1B)[a]	0.785	cg15096140	(MYO1B)[a]
15	0.875	cg15096140	(MYO1B)[a]	0.760	cg05767404	(C1orf150)	0.760	cg05767404	(C1orf150)
16	0.850	cg13745870	(SPATA12)	0.760	cg23506842	(PTPN7)	0.755	cg05004940	(C20orf195)
17	0.850	cg00134539	(UBASH3A)	0.755	cg05004940	(C20orf195)	0.755	cg23506842	(PTPN7)
18	0.840	cg16853982	(ACTN2)	0.750	cg13745870	(SPATA12)	0.750	cg13745870	(SPATA12)
19	0.840	cg10467098	(Bles03)[a]	0.735	cg13247990	(MLCK)	0.735	cg09626634	(EBI2)
20	0.835	cg13247990	(MLCK)	0.735	cg23917399	(TNFAIP8)	0.735	cg13247990	(MLCK)
20	–	–	–	–	–	–	0.735	cg23917399	(TNFAIP8)

[a]Indicates the overlapped genes in the top 20 lists of all 5 procedures.

Beadchip. Some genes in our findings are known to be associated with other types of cancers.

Since we mainly focused on identifying most likely associated CpG sites through selection probabilities, we did not discuss about estimation performance of our proposed method. In the case that prediction is desirable, we strongly recommend to reobtain unpenalized likelihood estimates based on selected predictors. This unpenalized MLE is often adopted for better prediction (Meier *et al.*, 2008; Wu *et al.*, 2009).

Our proposed selection method provides a way to potentially better select CpG sites that are truly related to the outcomes among tens of thousands of CpG sites who are correlated within genes. However, underlying true correlation structure of methylation data is much more complicated than what we assumed. Particularly, genes from a pathway may also be correlated with each other (Zhang and Wiemann, 2000). To incorporate a prior knowledge of genetic pathways may further improve the selection accuracy. We are currently pursuing the usage of genetic pathways information for the DNA methylation data analysis.

ACKNOWLEDGEMENTS

The authors want to thank the Yale High Performance Computing Center for providing computing facility (supported by NIH grant RR19895) and also want to thank Dr Yuanjia Wang for helpful discussions.

Funding: NCI (grant R03 CA150140) and NIH (grant R01 ES005116).

Conflict of Interest: none declared.

REFERENCES

Alexander,D. and Lange,K. (2011) Stability selection for genome-wide association. *Genet. Epidemiol.*, **35**, 722–728.

Bibikova,M. *et al*. (2006) High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.*, **16**, 383–393.

Breheny,P. and Huang,J. (2009) Penalized methods for bi-level variable selction. *Stat. Interface*, **2**, 369–380.

Friedman,J. *et al*. (2007) Pathwise coordinate optimization. *Ann. Appl. Stat.*, **1**, 302–332.

Friedman,J. *et al*. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Houseman,E. *et al*. (2008) Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, **9**, 365.

Kuan,P. *et al*. (2010) A statistical framework for illumina DNA methylation arrays. *Bioinformatics*, **26**, 2849–2855.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Li,C. and Li,H. (2010) Variable selection and regression analysis for covariates with a graphical structure with an application to genomics. *Ann. Appl. Stat.*, **4**, 1498–1516.

London,S.J. *et al*. (1997) Myeloperoxidase genetic polymorphism and lung cancer risk. *Cancer Res.*, **57**, 5001–5003.

Marsit,C.J. *et al*. (2009) Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma. *Carcinogenesis*, **30**, 416–422.

Meier,L. *et al*. (2008) The group lasso for logistic regression. *J. Roy. Stat. Soc. B*, **70**, 53–71.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. Roy. Stat. Soc. B*, **72**, 417–473.

Shen,J. *et al*. (2011) Genome-wide DNA methylation profiles in hepatocellular carcinoma. *Hepatology*, (in press).

Siegmund,K. *et al*. (2004) A comparision of cluster analysis methods using DNA methylation data. *Bioinformatics*, **20**, 1896–1904.

Teschendorff,A. *et al*. (2010) Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.*, **20**, 332–340.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B*, **58**, 267–288.

Tibshirani,R. *et al*. (2005) Sparsity and smoothness via the fused lasso. *J. Roy. Stat. Soc. B*, **67**, 91–108.

Wang,S. (2011) Method to detect diffentially methylated loci with case-contol designs using illumina arrays. *Genet. Epidemiol.*, **35**, 686–694.

Wu,T. *et al*. (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Zhang,J. and Wiemann,S. (2000) Kegggraph: a graph approach to KEGG pathway in r and bioconductor. *Bioinformatics*, **25**, 1470–1471.

Zhou,H. *et al*. (2010) Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, **26**, 2375–2382.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B*, **67**, 301–320.