A phase diagram for gene selection and disease classification

Hong-Dong Li[1*], Qing-Song Xu[2] , Yi-Zeng Liang[1*]

[1]*Research Center of Modernization of Traditional Chinese Medicines , College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P.R. China*
[2]*School of Mathematic Sciences, Central South University, Changsha 410083, P. R. China*

Abstract

Identifying a small subset of discriminate genes is important for predicting clinical outcomes and facilitating disease diagnosis. Based on the model population analysis framework, we present a method, called PHADIA, which is able to output a phase diagram displaying the predictive ability of each variable, which provides an intuitive way for selecting informative variables. Using two publicly available microarray datasets, it's demonstrated that our method can selects a few informative genes and achieves significantly better or comparable classification accuracy compared to the reported results in the literature. The source codes are freely available at: www.libpls.net.

Key words: *variable selection, gene selection, disease classification, model population analysis*

Introduction

High throughput experiments such as DNA microarray allow quantifying the expression levels of thousands of genes simultaneously and provide a large amount of data of potential clinical use such as disease risk prediction and classification [8,9,12,14,22,31]. The goal of disease/cancer classification includes, but is not limited to, predicting prognosis, proposing therapy according to the clinical situation, advancing therapeutic studies *etc*. Thus, it is critical for physicians/clinicians to establish the rule for accurate classification of tumor before any treatment is administered to the patient in order to avoid unnecessary treatment or propose the most appropriate therapies. The study of gene expression data based cancer classification has been extensively reported [21,25,26,29,30,34]

Due to many factors such as high cost, expression data are usually measured for a large number of genes on a small number of samples, giving rise to a measurement data matrix of a few rows and many columns. Predictive modeling and variable selection for such data is known as the a "large *p*, small *n*" problem[4,10,35], which is very challenging and has received a lot of attention in bioinformatics and statistics. So far, many methods have been proposed to identify potential genes which are relevant to cancer classification, *e.g.*, t-scores, class distinction correlation[12], support vector

• To whom correspondence should be addressed: lhdcsu@gmail.com or yizeng_liang@263.net

machines (SVM) [11,24], Gaussian process[7], sparse logistic regression[6], , regularized ROC method[22,23]， entropy-based method [20], graph based feature classification[13] and the MIA approach [18]. In predictive modeling, it is our opinion that a variable or gene is predictive performance should be assessed in terms of two aspects: (1) the expectation and (2) variance of predictive performance of a variable. The former shows how a variable improves a model's performance when included and the latter reflects how a variable affects the confidence interval (stability) of a model. These two factors are essential for quality assessment of variables but seemingly have not been addressed by existing methods.

In our previous work [18], we proposed the MIA (margin influence analysis) method that can computationally assess the expectation of predictive performance of variables using a population of sub-models [16,17,19]. However, the MIA method is only applicable to support vector machines [15] and does not consider the prediction variance of variables. Based on model population analysis [16,17,19], here we report a computational approach which can output a phase diagram displaying quality of variables in a two-dimensional plot. The algorithm was termed PHADIA. The phase diagram provides an intuitive tool for variable selection. Though our algorithm is implemented with PLS-LDA (Partial least squares - linear discriminant analysis), the PHADIA can be applied in combination with any other classifiers such as support vector machines and Bayesian network classifier. We applied PHADIA to two publicly available gene datasets and evaluated its performance for gene selection.

Methods

As mentioned above, PHADIA is proposed based on MPA[5,16], which is a general framework for developing new methods by analyzing the interesting model-related parameters, *e.g.* prediction errors and variable coefficients, of a number of sub-models based on Monte Carlo Sampling (MCS). An MPA algorithm consists mainly of three steps: (1) sampling N sub-datasets randomly, (2) building a sub-model using each sub-dataset, and (3) statistically analyzing some interesting output, *e.g.* prediction errors, of all the N sub-models. The third step is the key point of MPA. In the following section, the PHADIA algorithm will be described according to these three steps.

Sub-dataset sampling in the variable space

Given a dataset (**X**, **y**), where **X** of size n×p has n samples and p variables and **y** of size n×1 records the class label of each sample, valued 1 or -1 in the binary classification situation. The sub-dataset sampling in the variable space is described in the following (**Figure 1**). The number of Monte Carlo sampling (MCS) is set to N (*e.g.* 10,000). At each sampling, Q out of the p variables will be randomly sampled, giving rise to a sub-dataset of size n×Q. Repeating this procedure for N times, N

sub-datasets in total can be obtained, which are denoted $(\mathbf{X}_{sub}, \mathbf{y}_{sub})_i$, i = 1, 2, 3, …N. In **Figure 1**, the filled squares stand for the sampled variables.

(insert Figure 1)

Sub-model building using PLS-LDA

For each sub-dataset, a PLS-LDA model is built with the number of latent variables optimized by cross validation [2,32]. The reason to choose PLS-LDA is that (1) PLS has the potential to deal with high dimensional data and (2) linear classifier is easy to interpret compared to the nonlinear methods. N PLS-LDA classifiers are established in this step. 5-fold cross validation[3,28] is used to assess the performance of each sub-model. So each model is associated with a prediction error.

Statistical analysis of prediction errors for computing phase diagram

Without loss of generality, we take the *i*th variable for example to illustrate the procedure for computing a phase diagram.

First, we partition all the previously computed N PLS-LDA models into two groups, say Group A and Group B. Group A collects all the models including the *i*th variable. The remaining models not including the *i*th variable belong to Group B. Assuming that the numbers of the models in Group A and B are N_A and N_B, respectively, the sum of N_A and N_B is equal to N. So we have N_A prediction errors for Group A and N_B prediction errors for Group B. Then, the mean and standard deviation of the N_A and N_B prediction errors can hence be easily calculated, denoted $MEAN_A$, SD_A, $MEAN_B$ and SD_B, respectively. Finally, two statistics are can be computed for assessing the *i*th variable's prediction ability. The first is defined as the difference between $MEAN_A$ and $MEAN_B$, which is denoted and computed using the following formulae.

$$DMEAN_i = MEAN_{i,B} - MEAN_{i,A} \quad (1)$$

Clearly, $DMEAN_i$ measures the increment of prediction performances of model including the *i*th variable over models without it. So, if $DMEAN_i > 0$, one may infer that the prediction ability could be improved if a model contains the *i*th variable, and vice versa. To judge whether the mean errors of Group A and Group B are different, the nonparametric Mann-Whitney U test is employed to calculate a p-value which in combination with DMEAN is able to tell whether a variable can significantly improve prediction performance or not. In analogy to DMEAN, another statistic is defined as:

$$DSD_i = SD_{i,B} - SD_{i,A} \quad (2)$$

By definition, DSD_i can be thought as an criterion describing whether the inclusion of the *i*th variable will increase the variance of a model or not. It could be expected that the *i*th variable has the potential to stabilize a model if $DSD_i > 0$.

(insert Figure 2)

After computing DMEAN and DSD for each variable, one can plot DSD against DMEAN for all the variables. Such a plot is called a phase diagram in our work which intuitively displays the quality of all variables together. The phase diagram is illustrated in **Figure 2**. In this figure, all the variables are grouped into four regions. Phase 1: DMEAN>0, DSD>0, containing informative variables that can reduce prediction errors and also reduce prediction variance; Phase 2: DMEAN<0, DSD>0, housing variables which can will increase prediction variances; Phase 3: DMEAN<0, DSD<0, corresponding to variables that decrease model performance but can reduce prediction variances; Phase 4: DMEAN>0, DSD<0, including the "worst" variables which not only increase prediction errors but also increase prediction variances. Based on p-value of each variable, the shadow region with DMEAN close to 0 indicates variables which will not have significant influence on a model's prediction error. In general, variables in Phase 1 are best performing and those in Phase 2 can also be considered to be included into a model. However, the variables in Phase 3 and 4 are not suggested to be used for building a model.

Results and discussion

Colon data

This dataset contains the expression values of 6500 human genes measured on 40 tumor and 22 normal colon tissues using the Affymetrix gene chip. 2000 genes with the highest minimal intensity across samples were selected by Alon *et al.* [1]. Gene expression values were log2 transformed before modeling.

The parameter N, the number of Monte Carlo samplings, for PHADIA is fixed 10,000. Four Q values [20, 50, 100, 200] were tested. For each Q, we ran the PHADIA algorithm 10 times and calculated the predictive performance of PLS-LDA using cross validation (**Figure S1**). Considering both the mean and variance of prediction errors, Q=20 was chosen to be the optimal one. Although the sub-dataset is chosen randomly, we found that the top ranked informative variables identified by the PHADIA algorithm is highly reproducible (**Table S1**).

(insert Figure 3)

Figure 3 shows the phase diagram from one run of the PHADIA algorithm. This plot is divided into four phases, corresponding to the four types of genes as described previously. The genes (probe sets) in Phase 1 are most informative. Phase 2 houses less informative genes which can still be considered to be included in model. However, those genes in Phase 3 and 4 will most likely reduce a model's performance and therefore should be eliminated as interfering genes. We found that there are 878

genes with DMAN>0 (in Phase 1 or Phase 2), out of which 206 are significant with a p-value < 0.05.

We then from each Phase, manually picked one gene (Phase 1: index=493, Phase 2: index=633, Region 3: index=937 and Phase 4: index=261) and displayed their prediction error distributions in **Figure 4**. Take the 493th gene for example, the prediction error distribution of models including this variable shows significantly lower mean errors and smaller variance, representing the most informative variable. In contrast, the variables in Plot A and C are of poor predictive performances since including them into a model will on average reduce a model's performance.

To build a parsimonious PLS-LDA classifier, we ranked all the variables based on their DMEAN value and selected a subset of 200 genes using a forward strategy. The achieved average median leave-one-out-cross-validation (LOOCV) error of 10 replicate runs of the PHADIA algorithms is 0.0806±0.0221. For the same data, Furey *et al.* (2000) misclassified 6 samples using SVM based on LOOCV, leading to a prediction error=0.097. In Nguyen *et al.* (2002), their best result is obtained using PLS including 50 or 100 genes with the misclassification error = 0.065, which are slightly better than ours. Compared to the reported results, the proposed PHADIA algorithm shows competitive performances.

Estrogen data

This dataset consisted of the expression values of 7129 genes of 49 breast tumor samples, and presented by West *et al.* (2001) and Spang *et al.*(2001)[27]. There are 25 LN+ samples and 24 LN- samples. Before gene selection and classifier building, pretreatment is done on this data following the same methods described in Ma *et al.* (2005). 3333 genes were left and log2 transformed in our analysis.

(insert Figure 5)
(insert Figure 6)

For this data, N is also set to 10,000, and the optimal Q value is determined to be 20 using cross validation (**Figure S2**). The results of PHADIA are shown to be highly reproducible (**Table S1**). The phase diagram from one run of PHADIA with Q=20 is shown in **Figure 5**. In this plot, the 77th gene stands out, which should be of high predictive value based on our method. For illustration, we also selected one gene from each phase and their prediction error distributions are shown in Figure 6. Take the 77th gene as an example, it remarkably reduces the prediction error while simultaneously improve the predictive stability in terms of lower variance if included in a model. It may be inferred that the 77th gene is a key factor for underling the physiological state of estrogen. From Phase 1 and Phase 2, 312 genes were found to be informative with p<0.05.

Using the same method as applied to the colon data, we selected 25 genes and built a PLS-LDA classifier with a median LOOCV error 0.06±0.00 over 10 replicate runs of the PHADIA algorithm. For this dataset, based on their selected 100 genes, Dettling and Buhlmann (2003) yields classification errors 0.020 (LogitBoost, optimal), 0.06 (AdaBoost, 100 iterations) and 0.040 (CART). In the work of Ma *et al* (2005), their reported misclassification errors are 0.120±0.080. These results demonstrated that our method is very compelling.

Conclusions

Based on model population analysis, we in the present study introduced the PHADIA algorithm for variable selection. One unique feature of PHADIA is that it can output a phase diagram that describes the prediction ability of variables in terms of the expectation and variance of their prediction errors if included in a model. Based on the phase diagram, variables can also be classifier into informative, uninformative and interfering ones in a similar manner as in our previous work [18,19,33]. When applied to two gene expression datasets, competitive performances were achieved compared to the results reported in the literature. Our results indicate that PHADIA algorithm is a powerful tool for visualizing variables' prediction ability and identifying informative variables. It's expected that PHADIA will find more applications in other fields.

Acknowledgements

This work is financially supported by the National Nature Foundation Committee of P.R. China (Grants No. 21075138). The studies meet with the approval of the university's review board.

References

[1] Alon, U., N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96** (1999)6745-6750.

[2] Barker, M. and W. Rayens, 2003. Partial least squares for discrimination. *J. Chemometr.* **17** (2003)166-173.

[3] Baumann, K., 2003. Cross-validation as the objective function for variable-selection techniques. *TrAC* **22** (2003)395-406.

[4] Candes, E. and T. Tao, 2007. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35** (2007)2313-2351.

[5] Cao, D.S., Y.Z. Liang, Q.S. Xu, H.D. Li, and X. Chen, 2010. A New Strategy of Outlier

Detection for QSAR/QSPR. *J. Comput. Chem.* **31** (2010)592-602.

[6] Cawley, G.C. and N.L.C. Talbot, 2006. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics* **22** (2006)2348-2355.

[7] Chu, W., Z. Ghahramani, F. Falciani, and D.L. Wild, 2005. Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics* **21** (2005)3385-3393.

[8] Dettling, M. and P. Buhlmann, 2002. Supervised clustering of genes. *Genome Biology* **3** (2002)research0069.0061 - research0069.0015.

[9] Dhanasekaran, S.M., T.R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K.J. Pienta, M.A. Rubin, and A.M. Chinnaiyan, 2001. Delineation of prognostic biomarkers in prostate cancer. *Nature* **412** (2001)822-826.

[10] Fan, J. and R. Li, 2012. Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006*(2012).

[11] Furey, T.S., N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16** (2000)906-914.

[12] Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri et al., 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286** (1999)531-537.

[13] Hwang, T., H. Sicotte, Z. Tian, B. Wu, J.-P. Kocher, D.A. Wigle, V. Kumar, and R. Kuang, 2008. Robust and efficient identification of biomarkers by classifying features on graphs. *Bioinformatics* **24** (2008)2023-2029.

[14] Khan, J., J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson et al., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* **7** (2001)673-679.

[15] Li, H.-D., Y.-Z. Liang, and Q.-S. Xu, 2009. Support vector machines and its applications in chemistry. *Chemometr. Intell. Lab.* **95** (2009)188 -198.

[16] Li, H.-D., Y.-Z. Liang, Q.-S. Xu, and D.-S. Cao, 2009. Model population analysis for variable selection. *J. Chemometr.* **24** (2009)418-423

[17] Li, H.-D., Y.-Z. Liang, Q.-S. Xu, and D.-S. Cao, 2012. Model population analysis and its applications in chemical and biological modeling. *TrAC* **38** (2012)154-162.

[18] Li, H.-D., Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, B.-B. Tan, B.-C. Deng, and C.-C. Lin, 2011. Recipe for Uncovering Predictive Genes using Support Vector Machines based on Model Population Analysis. *IEEE/ACM T Comput Bi* **8** (2011)1633-1641.

[19] Li, H.-D., M.-M. Zeng, B.-B. Tan, Y.-Z. Liang, Q.-S. Xu, and D.-S. Cao, 2010. Recipe for revealing informative metabolites based on model population analysis. *Metabolomics* **6** (2010)353-361.

[20] Liu, J.J., G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, and X.B. Ling, 2005. Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics* **21** (2005)2691-2697.

[21] Lu, Y.-J., D. Williamson, J. Clark, R. Wang, N. Tiffin, L. Skelton, T. Gordon, R. Williams, B.

Allan, A. Jackman et al., 2001. Comparative expressed sequence hybridization to chromosomes for tumor classification and identification of genomic regions of differential gene expression. *Proc. Natl Acad. Sci. USA* **98** (2001)9197-9202.

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

[23] Ma, S., X. Song, and J. Huang, 2006. Regularized binormal ROC method in disease classification using microarray data. *Bmc Bioinformatics* **7** (2006)253.

[24] Pochet, N., F. De Smet, J.A.K. Suykens, and B.L.R. De Moor, 2004. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* **20** (2004)3185-3195.

[25] Qiu, P., Z.J. Wang, K.J.R. Liu, Z.-Z. Hu, and C.H. Wu, 2007. Dependence network modeling for biomarker identification. *Bioinformatics* **23** (2007)198-206.

[26] Shen, L., M. Toyota, Y. Kondo, E. Lin, L. Zhang, Y. Guo, N.S. Hernandez, X. Chen, S. Ahmed, K. Konishi et al., 2007. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc. Natl Acad. Sci. USA* **104** (2007)18654-18659.

[27] Spang, R., C. Blanchette, H. Zuzan, J. R. Marks, J. Nevins, and M. West. 2001. In proceedings of the German Conference on Bioinformatics GCB 2001.

[28] Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. B* **36** (1974)111-147.

[29] Virtanen, C., Y. Ishikawa, D. Honjoh, M. Kimura, M. Shimane, T. Miyoshi, H. Nomura, and M.H. Jones, 2002. Integrated classification of lung tumors and cell lines by expression profiling. *Proc. Natl Acad. Sci. USA* **99** (2002)12357-12362.

[30] Wang, L. and C. Aliferis, 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics* **9** (2008)319.

[31] West, M., C. Blanchette, H. Dressmna, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. Marks, and J. Nevins, 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA* **98** (2001)11462 - 11467.

[32] Yi, L.-Z., J. He, Y.-Z. Liang, D.-L. Yuan, and F.-T. Chau, 2006. Plasma fatty acid metabolic profiling and biomarkers of type 2 diabetes mellitus based on GC/MS and PLS-LDA. *FEBS Letters* **580** (2006)6837-6845.

[33] Yun, Y.-H., W.-T. Wang, M.-L. Tan, Y.-Z. Liang, H.-D. Li, D.-S. Cao, H.-M. Lu, and Q.-S. Xu, 2014. A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration. *Analytica Chimica Acta* **807** (2014)36-43.

[34] Zhang, L., W. Zhou, V.E. Velculescu, S.E. Kern, R.H. Hruban, S.R. Hamilton, B. Vogelstein, and K.W. Kinzler, 1997. Gene Expression Profiles in Normal and Cancer Cells. *Science* **276** (1997)1268-1272.

[35] Zou, H. and T. Hastie, 2005. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67** (2005)301-320.

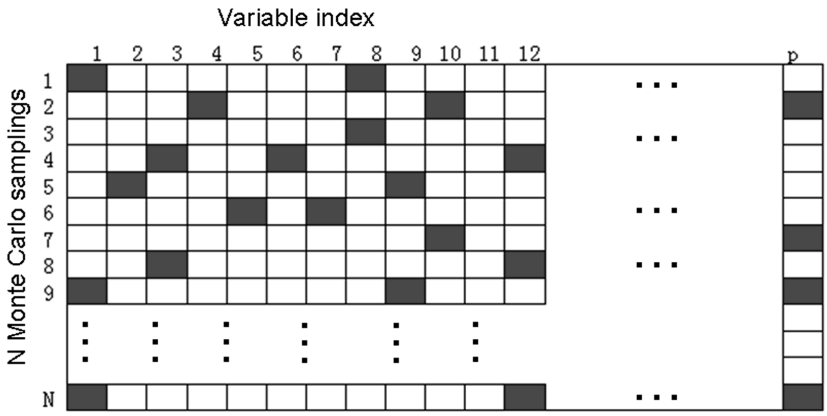

Figure 1. **Illustration of Monte Carlo sampling in the variable space.** In each sampling, a sub-dataset containing only a given number (Q, e.g. 10) of variables are randomly selected from the original data, denoted by filled square. Finally, N sub-datasets are generated and a sub-model will be built using each of the sub-dataset.

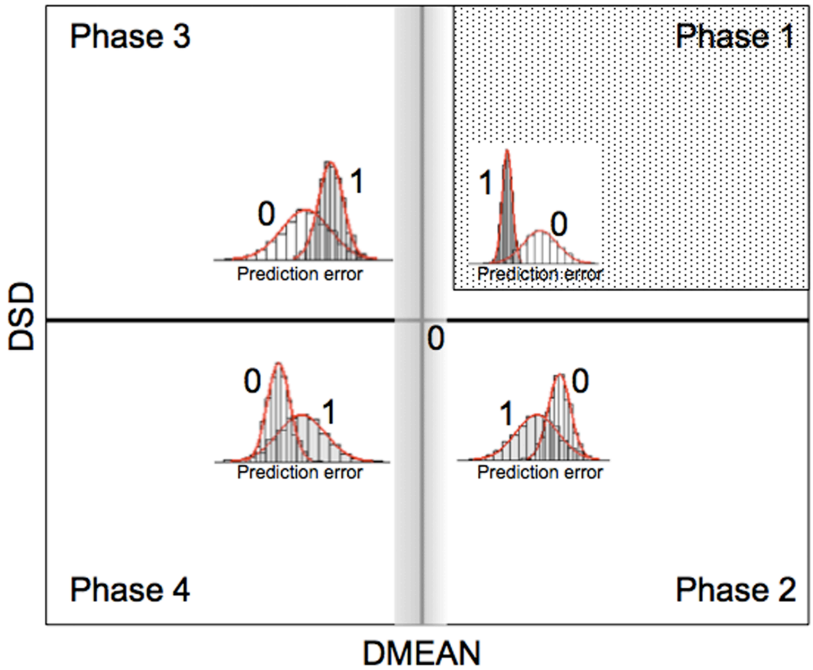

Figure 2. **A schematic of the phase diagram for variable selection.** Based on the DMEAN and DSD values (see main text for definition), variables fall into four main phases. The variables in Phase 1 can not only increase a model's performance (DMEAN>0) but also reduce prediction variances (DSD>0). In the inset, the peak denoted by "1" stands for the prediction error distribution of models including such a variable, while the peak denoted by "0" is the prediction error distribution of models that do not include this variable. This type of variables is called informative variables. The variables in Phase 2 can also increase performance but at the cost of increased variance, thus being less informative than those in Phase 1. In contrast, variables in Phase 3 and 4 will reduce the performance (DMEAN<0) of models, so they are suggested not to be used for modeling. These are called interfering variables. Specifically, including variables in Phase 4 into a model will even increase the prediction variance (DSD<0). In addition, variables at the boundary (the shadow region with DMEAN close to 0) can not significantly increase or decrease model performances, and are hence thought of as being uninformative.

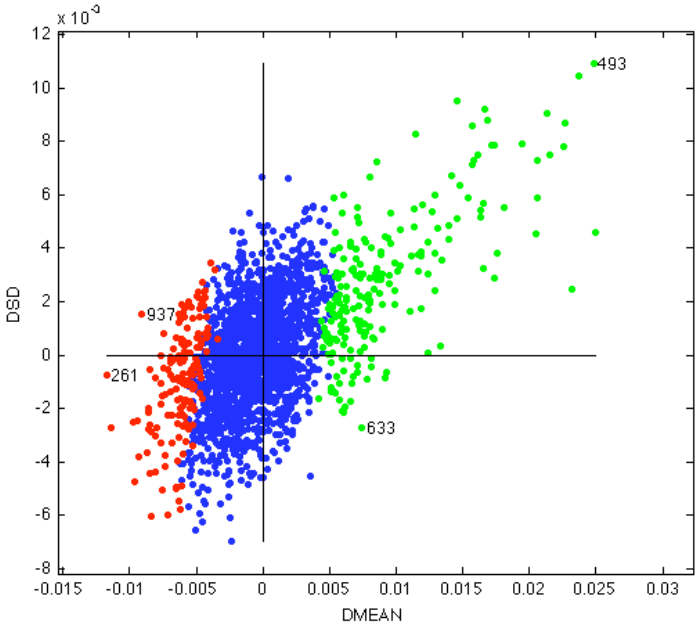

Figure 3. The phase diagram of the colon data. Red: informative variables (DMEAN>0, p<0.05); blue: uninformative variables (p>0.05); red: interfering variables (DMEAN<0, p<0.05).

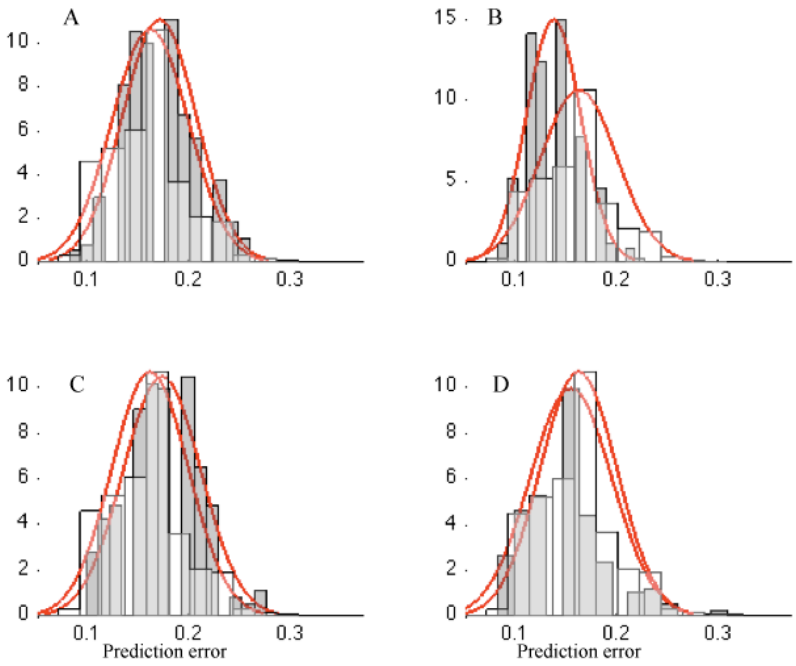

Figure 4. The prediction error distributions of four genes from each of the four phases for the colon data. They are picked from Phase 1 (plot B, index=493), Phase 2 (plot D, index=633), Phase 3 (plot A, index=937) and Phase 4 (plot C, index=261) in Figure 3, respectively.

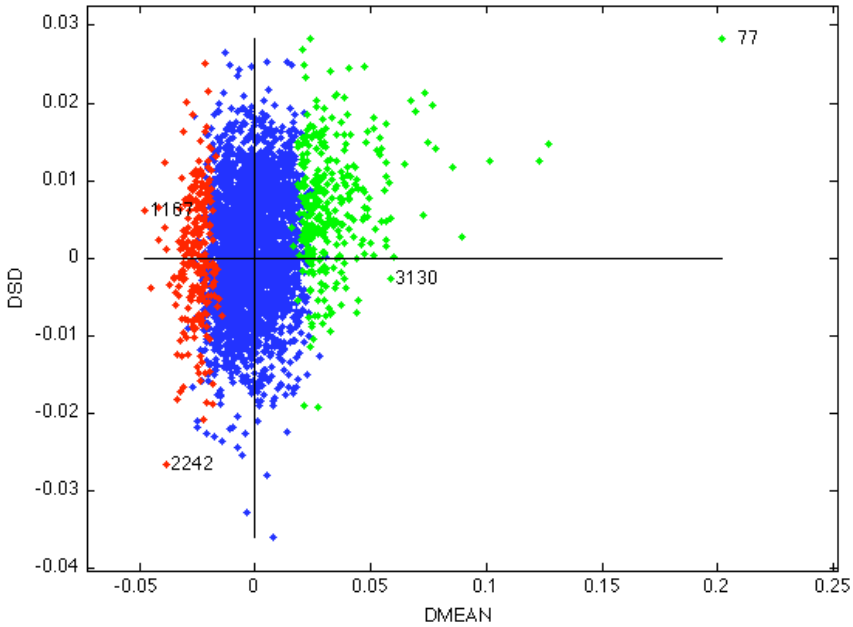

Figure 5. The phase diagram of the estrogen data. Red: informative variables (DMEAN>0, p<0.05); blue: uninformative variables (p>0.05); red: interfering variables (DMEAN<0, p<0.05).

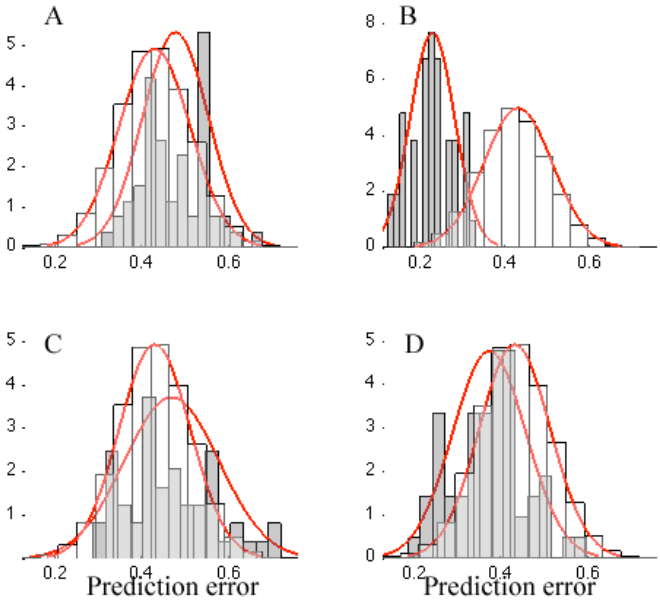

Figure 6. The prediction error distributions of four genes from each of the four phases for the estrogen data. They are picked from Phase 1 (plot B, index=77), Phase 2 (plot D, index=3130), Phase 3 (plot A, index=1167) and Phase 4 (plot C, index=2242) in Figure 5, respectively.

Supplementary materials

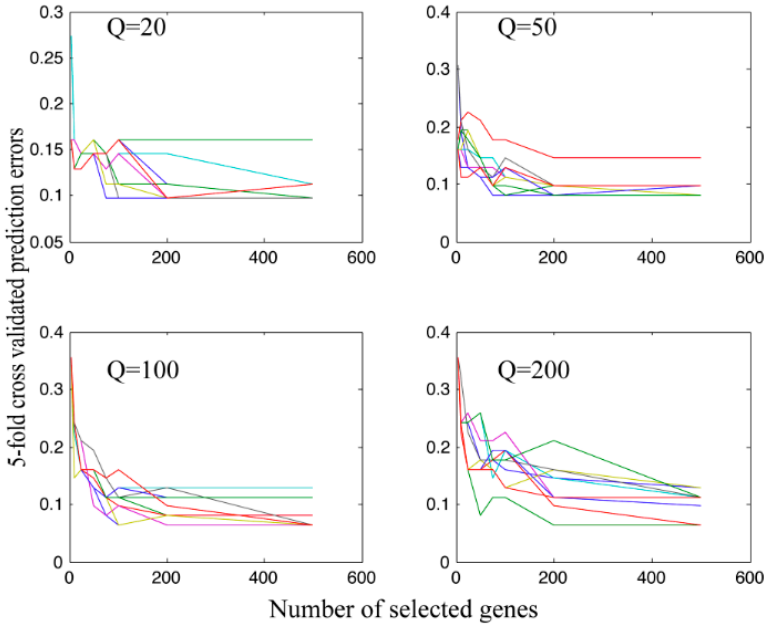

Figure S1. 5-fold cross validated prediction errors using different Q values for the colon data.

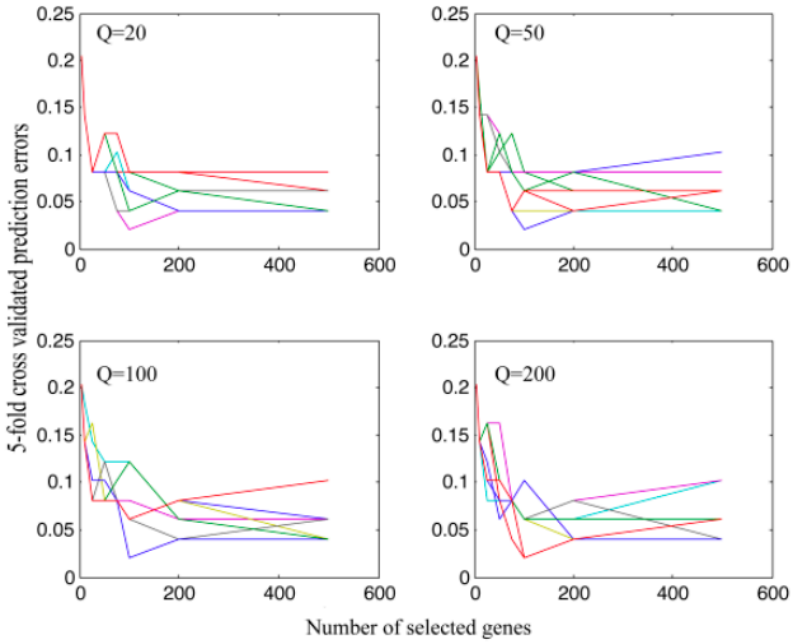

Figure S2. 5-fold cross validated prediction errors using different Q values for the estrogen data.