Bayesian Model Averaging (BMA)

When classifying samples on microarray data, our goal is to identify a small set of predictive genes. In the BMA framework, small sets of predictive genes are called "models". In microarray analysis, there are typically many "models" (or sets of predictive genes) that fit the data well. Bayesian Model Averaging (BMA) takes model uncertainty into consideration by computing the weighted average of the posterior probabilities that a test sample belongs to a given class over multiple "good" models.[1,2] In the weighted average calculations, the weight of each model is equal to the posterior probability of the model, and hence, proportional to the goodness of fit of the model. The set of "good" models (and hence, the number of predictive genes and models) used in the weighted average calculations is chosen by first applying the leaps and bounds algorithm, and then the Occam's window method. When the number of genes is small (< 40), the leaps and bounds algorithm [3] efficiently identifies a reduced set of "good" models M_k and returns the best *nbest* models of each size (*nbest* = 10 in this study).[1] The Occam's window method chooses a set of parsimonious and data-supported models by discarding models that are much less likely than the best model supported by the data.[4] We used logistic regression to compute the probability that a test sample belongs to the given class under each model M_k, and the Bayesian Information Criterion (BIC) to approximate the posterior probability of a model M_k.[1,5,6]

However, typical microarray analyses involve a large number of genes (>> 40) such that the leaps and bounds algorithm becomes inefficient, and the number of genes is usually much greater than the number of patient samples. Therefore, we developed the *iterative BMA algorithm* [7] in which we start by ranking the genes using a univariate measure such as the ratio of between-group to within-group sum of squares (BSS/WSS).[8] In this initial preprocessing step, genes with large BSS/WSS ratios (i.e., genes with relatively large variation between classes and relatively small variation within classes) receive high rankings. We then apply the leaps and bounds algorithm and Occam's window method to the top 30 ranked genes. Then genes that were assigned low posterior probabilities (< 5%) are removed. Suppose *m* genes are removed. The next *m* genes from the rank ordered BSS/WSS ratios are added back to the set of genes so that we maintain a window of 30 genes and apply leaps and bounds again. These steps of gene swaps and iterative applications of leaps and bounds are continued until all genes are subsequently considered. The iterative BMA algorithm is available as a bioconductor package called "iterative BMA" and part of MeV+R.[9] We showed that this iterative BMA method worked very well in predicting out-of-sample test cases for several datasets, and selected small numbers of genes in our previous work.[7]

In this study, we applied the iterative BMA algorithm to the 2,612 differentially expressed genes derived from ANOVA analysis using a posterior probability threshold of 5%. In the BMA framework, we distinguished between two different types of posterior probabilities: posterior probabilities of genes and posterior probabilities of models. Posterior probabilities of models are derived from the leaps and bounds and Occam's window steps. We define the posterior probabilities of a gene as the sum of the posterior probabilities of all the selected models containing the gene of interest. Using the 5% posterior probability threshold for genes, genes with posterior probabilities below 5% are removed in each of the iterative step. Since we threshold on the posterior probabilities of genes (but not of models), the posterior probabilities of the selected models are allowed to fall below 5%.

REFERENCES

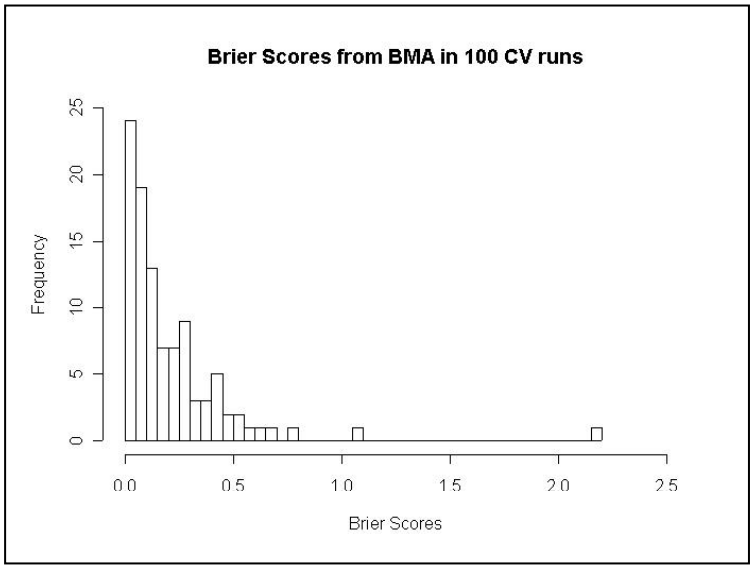
1. Raftery AE. Bayesian model selection in social research. In: Marsden PV, ed. Sociological Methodology. Vol 25. Cambridge, MA: Blackwells; 1995: 111–163.
2. Hoeting JA, Madigan D, Raferty AE, Volinksky C. Baysian model averaging: a tutorial. *Stat Sci*. 1999; 14:382–417.
3. Furnival GM, Wilson RW. Regression by leaps and bounds. *Technometrics*. 1974; 16:499–511.

4. Madigan D, Raftery A. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*. 1994; 89:1335–1346.
5. Schwarz G. Estimating the dimension of a model. *Annals of statistics*. 1978; 6:461–464.
6. Kass RE, Raftery AE. Bayes Factors. *Journal of the American Statistical Association*. 1995; 90:773–795.
7. Yeung KY, Bumgarner RE, Raftery AE. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*. 2005; 21(10):2394–2402.
8. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors in gene expression data. *Journal of the American Statistical Association*. 2002; 97:77–87.
9. Chu VT, Gottardo R, Raftery AE, Bumgarner RE, Yeung KY. MeV+R: using MeV as a graphical user interface for Bioconductor applications in microarray analysis. *Genome Biol*. 2008; 9(7):R118.

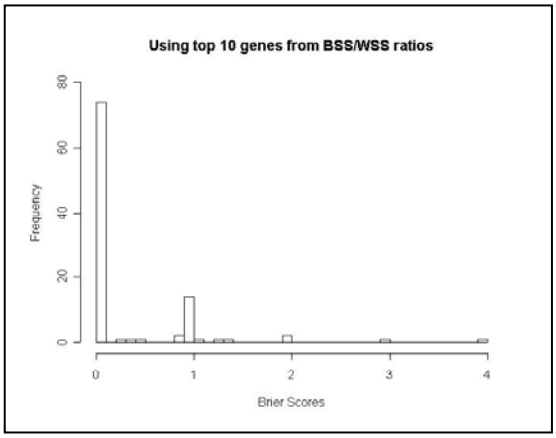
Comparison of BMA results to other methods

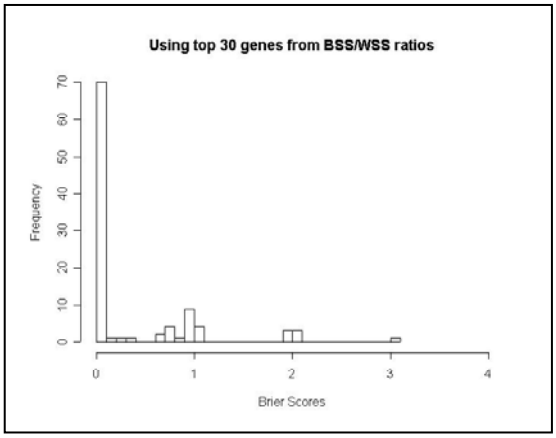
The iterative BMA algorithm selected 6 signature genes (*see Table 1 in the main manuscript*) to predict CML progression. Here, we compared the performance of iterative BMA to other widely-used computational methods. We evaluated different methods using 3-fold cross validation (CV) on the CML microarray data. Specifically, we used the same random splits (2/3 training and 1/3 testing) over 100 cross validation runs to evaluate performance of the following alternative methods to BMA:

1. Iterative BMA (our selected approach)
    - Approach: BMA selected 6 signature genes averaged over 21 models. The numbers of genes and models are determined automatically by the algorithm. See the main manuscript for more details.
    - Results: average brier score = 0.21 over 100 CV runs. In particular, 43 (out of 100 runs) have brier scores < 0.1 and only 2 CV runs have brier scores > 0.9.

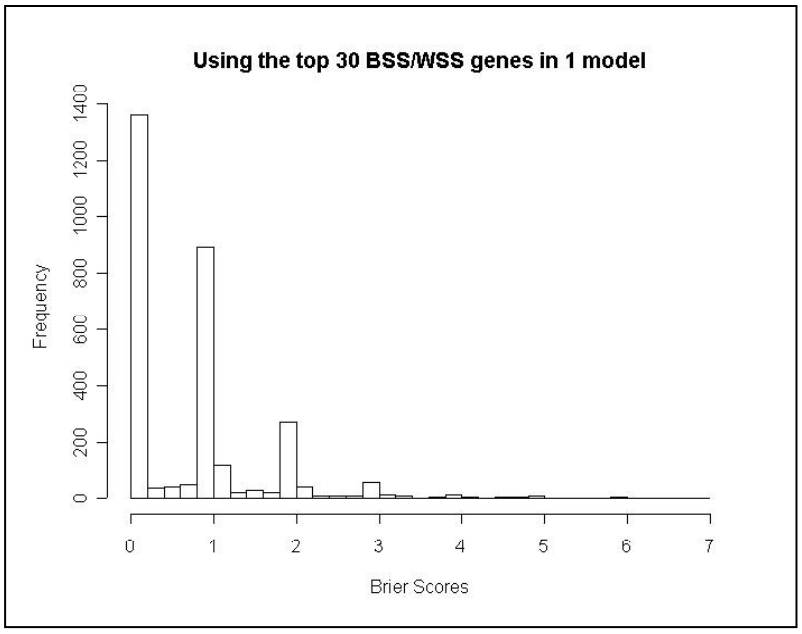


2. Logistic regression using the top P *BSS/WSS genes* in a single model
    - Approach: We first selected the top P genes (P=6, 10, 30) using the between-group to within group sum of squares (BSS/WSS) ratios, and then applied logistic regression using the P selected genes in a single model.
    - Results using P=6: average brier score = 0.375 over 100 CV runs. In particular, 68 (out of 100 runs) have brier scores < 0.1 and 24 CV runs have brier scores > 0.9. In general, this method gives rise to a more extreme distribution of brier scores than BMA.
    - Results using P=10: average brier score = 0.310 over 100 CV runs. In particular, 74 (out of 100 runs) have brier scores < 0.1 and 21 CV runs have brier scores > 0.9. In general, this method gives rise to a more extreme distribution of brier scores than BMA.
    - Results using P=30: average brier score = 0.339 over 100 CV runs. In particular, 70 (out of 100 runs) have brier scores < 0.1 and 20 CV runs have brier scores > 0.9. In general, this method gives rise to a more extreme distribution of brier scores than BMA.





3. Logistic regression using each of the top P *BSS/WSS genes* in single-gene models
    - Approach: We first selected the top 30 genes using the between-group to within group sum of squares (BSS/WSS) ratios, and then applied logistic regression to single-gene models consisting of each of these 30 genes.
    - Results: average brier score = 0.749 over 100 CV runs. Note that there are a total of 3000 single-gene models over 100 CV runs. In general, this method gives rise to a more extreme distribution of brier scores than BMA.



Positional Gene Enrichment (PGE) analysis using 2,612 genes derived from ANOVA analysis

We repeated the PGE analysis using the 2,612 differentially expressed genes derived from the ANOVA analysis, and obtained similar results to that reported in the main manuscript. As shown in **Table S1**, we observed chromosomal enrichment on the regions of chromosome 12 on which 3 of our 6 signature genes falls on (12p13, 12q23 and 12q24). Note that PGE does not exclude overlapping regions in its analyses.