

Version 10/16/01

A method for developing toxicologically predictive gene sets using cDNA microarrays and Bayesian classification

Russell S. Thomas¹, David R. Rank², Sharron G. Penn², Mark W. Craven³, Norman R. Drinkwater¹, and Christopher A. Bradfield¹

¹McArdle Laboratory for Cancer Research, University of Wisconsin Medical School, Madison, Wisconsin 53706

²Aeomica, Sunnyvale, California 94085

³Department of Biostatistics and Medical Informatics, University of Wisconsin Medical School, Madison, Wisconsin, 53706

Running Title: Predictive gene sets

Correspondence should be addressed to:

Christopher A. Bradfield
McArdle Laboratory for Cancer Research
1400 University Avenue
Madison, WI 53706-1599
Ph: (608) 262-2024
Fax: (608) 262-2824
email: bradfield@oncology.wisc.edu

Introduction

The application of DNA microarray technology to the field toxicology has increased significantly. Recent applications include studies identifying gene expression changes related to dioxin exposure (Frueh *et al.*¹) and correlating carbon tetrachloride-induced hepatotoxicity with interleukin-8 expression (Holden *et al.*²). In most cases, the emphasis of the research has been to monitor global changes in gene expression in order to provide insight into the cellular mechanisms of toxicity. Although monitoring global gene expression changes may prove to be important when characterizing the action of a particular chemical, it is not necessarily predictive of how that chemical or any other chemical will behave toxicologically without constructing appropriate statistical models.

The potential applications of predictive statistical models in toxicology based on gene expression measurements are numerous. For example, short-term studies measuring gene expression could be used to predict long-term toxicity studies like those still performed by the National Toxicology Program (NTP). Given that chronic exposure studies can cost between 2- million dollars, the cost benefits of rapid predictive approaches are obvious. Other short-term gene expression studies could be used to predict which chemicals would be teratogenic or cause more subtle developmental changes after human prenatal exposure. In either case, the application of microarray analysis and predictive statistical models has the potential to be extremely useful from both economic and human health perspectives.

Background and Considerations in Deriving Predictive Gene Sets

The development of a toxicologically predictive statistical model based on gene expression profiles is similar to the development of other multivariate classification models. In this case, two or more predictor variables (i.e., expression measurements of multiple genes) are used along with a single criterion variable that is categorical with either binomial or multinomial classes (i.e., toxicological class or endpoint). The goal is to use the data vectors in the test set of chemicals, which are composed of the predictor variables, to build a model that correctly classifies the criterion variable and can eventually be used to predict other untested chemicals. As a result, the capability of the model to correctly classify the criterion variable is directly related to the amount of information the predictor variables contain about the criterion variable.

Not only is the ultimate success of a model dependent on the information contained within the predictor variables, the removal of uninformative predictor variables and the selection of a diagnostic subset of genes is also important for both biological and statistical reasons. In the biological sense, the classification of a set of chemicals into a toxicological class or endpoint based on gene expression is difficult due to the variety of potential mechanisms that underlie the toxicity of these chemicals. For example, both 2,3,7,8-tetrachlorodibenzo-p-dioxin and *N,N*-dimethyl-4-aminoazobenzene are carcinogenic in laboratory animals. However, they produce the same toxic endpoint, cancer, through different mechanisms. Dioxin is a nongenotoxic carcinogen that acts through a ligand-activated transcription factor known as the Ah receptor while DAB is metabolized to an active form that is mutagenic. Even though the toxicological endpoint may be the same, these two chemicals will undoubtedly activate

and repress different biochemical and molecular pathways leading to different global patterns in gene expression. Therefore, in order to construct a predictive model for chemically induced cancer, the global gene expression dataset must be reduced to a subset of genes that are diagnostic across all chemicals.

From a statistical perspective, selection of a subset of diagnostic genes is also important. Statistical models with few predictors relative to the overall sample size typically yield more accurate (i.e., less biased) and more precise estimators (Hora and Wilcox³; Huberty⁴). For Bayesian models, underlying assumptions are made about the dependency of the variables. If the assumptions are violated in the full data set, better classification results can be obtained by removing uninformative predictor variables making the dependency assumptions less severe.

Although the method presented in this paper can be used on any microarray dataset related to toxicology and virtually any toxicological endpoint, the original application and development was performed on a microarray study to classify 24 prototype chemical treatments into five well-characterized toxicological classes. A more detailed description of the treatments and toxicological classes are provided in the original study (Thomas *et al.*⁵).

Microarray Construction and Image Analysis

Gene expression for the original set of toxicological treatments was measured using a microarray that contained approximately 1,200 minimally redundant cDNAs from both an internal expressed sequence tag (EST) project and the public EST effort. In the internal EST project, mice were treated with various hepatotoxicants and ESTs were

sequenced from the livers of control and treated mice. As a result, the microarrays contained a high percentage of genes that are both expressed in the liver and induced following chemical treatment. The ESTs supplemented from the public effort were genes believed to be toxicologically important, but missing from our EST collection.

To construct cDNA microarrays, an aliquot from each glycerol stock is transferred to a 96-well PCR plate and lysed at 95 °C in 1X PCR buffer. The resulting lysate is amplified using PCR. PCR reactions are purified using Millipore glass filter plates (MAFB N0B, Bedford, MA). Briefly, 200 µl of binding buffer (150 mM potassium acetate, pH 4.9; 5.3 M guanidine-HCl) is added to the PCR reactions and passed over the glass filters using vacuum filtration. The filters are washed 4X with 80% EtOH and eluted with 65 µl distilled water. The purified PCR products are dissolved in 5 M NaSCN to create the final spotting solution. Prior to spotting, cleaned microscope slides (VWR, West Chester, PA) are exposed for 2 h to the vapors of 3-aminopropyltrimethoxysilane (United Chemical Technologies, Bristol, PA) in a stream of nitrogen. The slides are washed with water, cured overnight at 100 °C, and spotted in replicate with the purified PCR products in NaSCN. After spotting, the slides are baked at 80 °C for 2 h. Prior to hybridization the slides are gently agitated for 10 min in isopropyl alcohol and boiled in water for 5 min.

Labeled cDNA probe is produced from 1 µg poly-A RNA by incorporation of Cy-dCTP (Amersham Pharmacia, Piscataway, NJ) during a standard reverse transcriptase reaction primed with oligo dT₍₁₂₋₁₈₎ and random nonomers (Life Technologies, Gaithersburg, MD). The poly-A RNA and oligonucleotides are heated at 70 °C for 10 min and transferred to ice. The reverse transcription is performed in a solution

containing 1X Superscript II buffer (Life Technologies), 10 mM DTT, 100 mM dATP, 100 mM dTTP, 100 mM dGTP, 50 mM dCTP, 50 mM Cy- dCTP, and SuperScript II enzyme. The reaction is incubated at 42 °C for 2 h. Following incubation, 1 µl RNase H (Life Technologies) is added to the reaction and the reaction incubated at 37 °C for 30 min. Purification of labeled cDNA is carried out using Qiaquick PCR purification columns (Qiagen). Fifty picomoles of each dye is dried down in a SpeedVac and resuspended with 40µl of hybridization solution containing a mixture of 50 % formamide, 4 X SSC (0.15 M sodium citrate, 15 mM sodium chloride), 0.1 % SDS, 50 ug/mL human COT-1 DNA (Life Technologies), and 50 ug/ul polyA₈₀ DNA (Amersham Pharmacia). The hybridization solution is added to the slide, a cover slip added, and the hybridization performed for 18 h at 42 °C. Following hybridization, the slides are washed for 5 min in 2X SSC/0.1 % SDS at 42 °C, 5 min in 1X SSC/0.1 % SDS at 25 °C, 5 min in 0.1X SSC at 25 °C, and a brief rinse in deionized water. The slides are scanned using a microarray scanner at an excitation wavelength of 532 nm (Cy3) or 632 nm (Cy5).

For each hybridization, the fluorescence data from the replicate spots corresponding to each cDNA are averaged and normalized. To eliminate any dye bias, all samples are typically analyzed at least twice, with one experiment using Cy3 to label the control mRNA and Cy5 to label the treated mRNA and in the replicate experiment Cy5 to label the control and Cy3 to label the treated. The results from the replicate hybridizations are averaged.

Data Reduction

Prior to statistical analysis, the genes (i.e., predictor variables) are first collapsed into a fully nonredundant set by averaging the results from multiple copies of the same gene. After collapsing, the dataset is screened for genes that do not respond to any of the treatments used in the study and do not contribute significantly to the classification. A threshold of 2-fold change in gene expression is typically used as the cut-off value and is similar to a standard threshold level used in other studies (e.g., Schena *et al.*⁶).

In datasets with time course or dose response experiments, additional screening is typically performed to identify a subset of genes with a stable expression over all time points and doses. This provides computational savings by eliminating variables that would have little predictive value and is achieved by collapsing all time points or doses into a single average value representing the average change in that gene. After collapsing the data, only genes that showed greater than a 2-fold change in more than one treatment are selected for further analysis. After this data reduction step, the individual time points or doses are analyzed separately in the classification analysis. Collapsing the data is only used as a data screening tool and not for the statistical analysis. Finally, the gene expression values are discretized such that genes upregulated greater than 2-fold were given a value of one, genes downregulated greater than 2-fold were given a value of minus one, and genes with less than a 2-fold change were given a value of zero. A flow chart describing the data reduction is provided in Figure 1.

Statistical Classification Analysis

The type of classification model applied to this problem can be one of a number of alternatives. We have chosen to use the Naïve Bayes model structure based on previous work by Kontkanen and colleagues⁷, which has been shown to perform well in comparison with other approaches (Langley *et al.*⁸, Friedman *et al.*⁹). In the Naïve Bayes method, the predictor variables X_1, \dots, X_k are assumed to be independent of each other when conditioned on the class variable C . Our model M is constructed by the joint probability distribution for a data vector $(x, c) = (X_1 = x_1, \dots, X_k = x_k, C = c)$ and can be written as follows:

$$P(x, c) = P(C = c) \prod_{i=1}^k P(X_i = x_i | C = c) \quad (1)$$

Prior to incorporating any data, we also assume that all classes are equally probable (i.e., the probability that a toxicological treatment will belong to a certain class or endpoint is the same for each class) and that within each class the gene expression values of each gene are also equally probable. Given these assumptions, we can use Bayesian probability theory to calculate the conditional predictive distribution for the class c given x and the data set D by:

$$P(c | x, D) = \frac{P(c, x | D)}{P(x | D)} \quad (2)$$

where the numerator is calculated as:

$$P(c, x | D) = \frac{t_c + 1}{N_t + NC} \prod_{i=1}^k \frac{f_{cx_i} + 1}{F_c + V_{x_i}} \quad (3)$$

where t_c is the number of toxicological treatments in class c , N_t is the total number of toxicological treatments, NC is the total number of classes, f_{cx_i} is the number of cases in class c having a value equal to x_i , F_c is the number of toxicological treatments in class c ,

and V_{x_i} is the number of possible values of x_i . The denominator is the same for all c and is calculated as:

$$P(x | D) = \sum_{c'} P(c', x | D) \quad (4)$$

The result of this conditional predictive distribution is then used to classify the data vector (i.e., the combination of all the gene expression measurements for that toxicological treatment).

Parameter Selection

For the parameter or gene selection, a modification of forward parameter selection process outlined in Huberty⁴ is employed. Specifically, an iterative process is used where genes are run individually using the Naïve Bayes model and the gene with the best internal classification rate and highest confidence (represented by the sum of all probabilities for correctly classified treatments) is selected. The internal classification rate is defined as the number of data vectors that are classified correctly with the model without performing cross-validation divided by the total number of data vectors. In the subsequent round, the selected gene is fixed and the remaining parameters are added individually to find which gene, along with the first selected gene, produces the highest internal classification rate and confidence. This process is repeated until all genes that pass the data reduction process are added to the model in the order of their internal classification rate. This type of selection ranks the genes in the order of their estimated predictive value and sequentially adds them to the model. It should be noted that the forward selection approach does not, necessarily, yield the best set or even the smallest set. In addition, genes with similar expression profiles may also be added to the set if

they significantly increase the accuracy or confidence. The approach is simply a heuristic to look for a diagnostic set of predictors with a high accuracy.

To estimate the predictive accuracy of this approach, the process of parameter selection is integrated with leave-one-out cross-validation where one of the treatments is removed from the analysis, the model constructed and then used to predict the left-out treatment. The predictive accuracy is then assessed after each parameter is added and the number of genes in the 'diagnostic set' is chosen based on the peak predictive accuracy and confidence of the model (Fig. 2). The final 'diagnostic set' is derived by following the same procedure on the complete dataset (i.e., no treatment left out). A flow chart describing the classification analysis and parameter selection is outlined with the data reduction steps in Figure 1. All statistical analyses are scripted in Perl and the general code is available via download (<http://edge.oncology.wisc.edu/>) or by emailing the corresponding author (bradfield@oncology.wisc.edu).

Conclusions

The application of predictive statistical models to chemically induced gene expression is the next logical step in the developing field of toxicogenomics. The development of these models will eventually open the door to a new era of toxicological testing where relatively short and inexpensive microarray studies will allow the assessment of the human health risks associated with a previously untested chemical. This would mean significant savings in both animal usage and financial resources while also reducing the disparity between the number of tested and untested chemicals in commerce today. However, the accuracy and applicability of these models are highly

dependent of the quality of the training sets used in their development. As the public gene expression database grows, more chemicals can be added to training the models and the more predictive those models will become.

Acknowledgements

This work was supported by The Burroughs Wellcome Foundation, National Institutes of Health (Grants ES05703, T32CA09681, CA07175, and GM23750), and a postdoctoral fellowship co-sponsored by the Society of Toxicology and the Colgate-Palmolive Corporation.

References

1. F.W. Frueh, K.C. Hayashibara, P.O. Brown, and J.P. Whitlock Jr, *Toxicol. Lett.* **122**, 189 (2001).
2. P.R. Holden, N.H. James, A.N. Brooks, R.A. Roberts, I. Kimber, W.D. Pennie, *J. Biochem. Mol. Toxicol.* **14**, 283 (2001).
3. S.C. Hora and J.B. Wilcox, *J. Marketing Res.* **19**, 57 (1982).
4. C.J. Huberty, "Applied Discriminant Analysis." Wiley, New York. 1994.
5. R.S. Thomas, D.R. Rank, S.G. Penn, G.M. Zastrow, K.R. Hayes, K. Pande, E. Glover, T. Silander, M.W. Craven, J.K. Reddy, S.B. Jovanovich, and C.A. Bradfield, *Mol. Pharmacol.* **60**, (In Press).
6. M. Schena, D. Shalon, R. Heller, A. Chai, P.O. Brown, and R.W. Davis, *Proc. Natl. Acad. Sci.* **93**, 10614 (1996).
7. P. Kontkanen, P. Myllymaki, T. Silander, and H. Tirri, in "Proceedings of the Fourth International Conference on Knowledge Discovery & Data Mining", p. 254. AAAI Press, Menlo Park, CA, 1998.

8. P. Langley, W. Iba, and K. Thompson, in "Proceedings of the Tenth National Conference on Artificial Intelligence", p. 223. AAAI, San Jose, CA, 1992.

9. N. Friedman, D. Geiger, and M. Goldszmidt, *Machine Learning* **29**, 131 (1997).

Figure Legends

Figure 1. A flow-chart outlining the method for data reduction and classification analysis leading to the diagnostic set of genes.

Figure 2. An example of the estimated accuracy and relative confidence of the classification model for predicting a toxicological endpoint as a function of genes added to the model. The genes are added to the model using a forward selection scheme and the predictive accuracy is estimated using leave-one-out cross-validation. The diagnostic set of genes is highlighted by the hatched area.

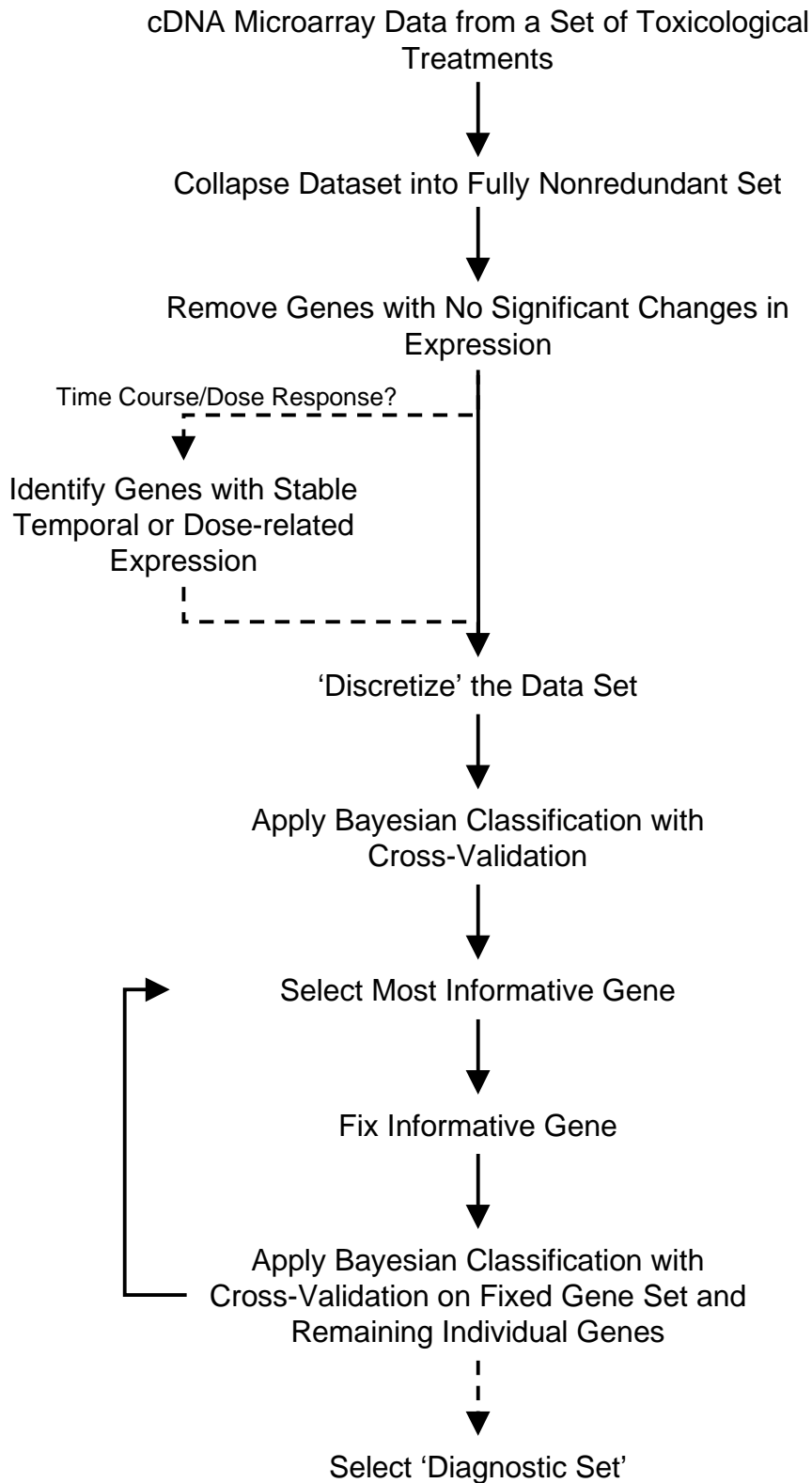


Fig. 1

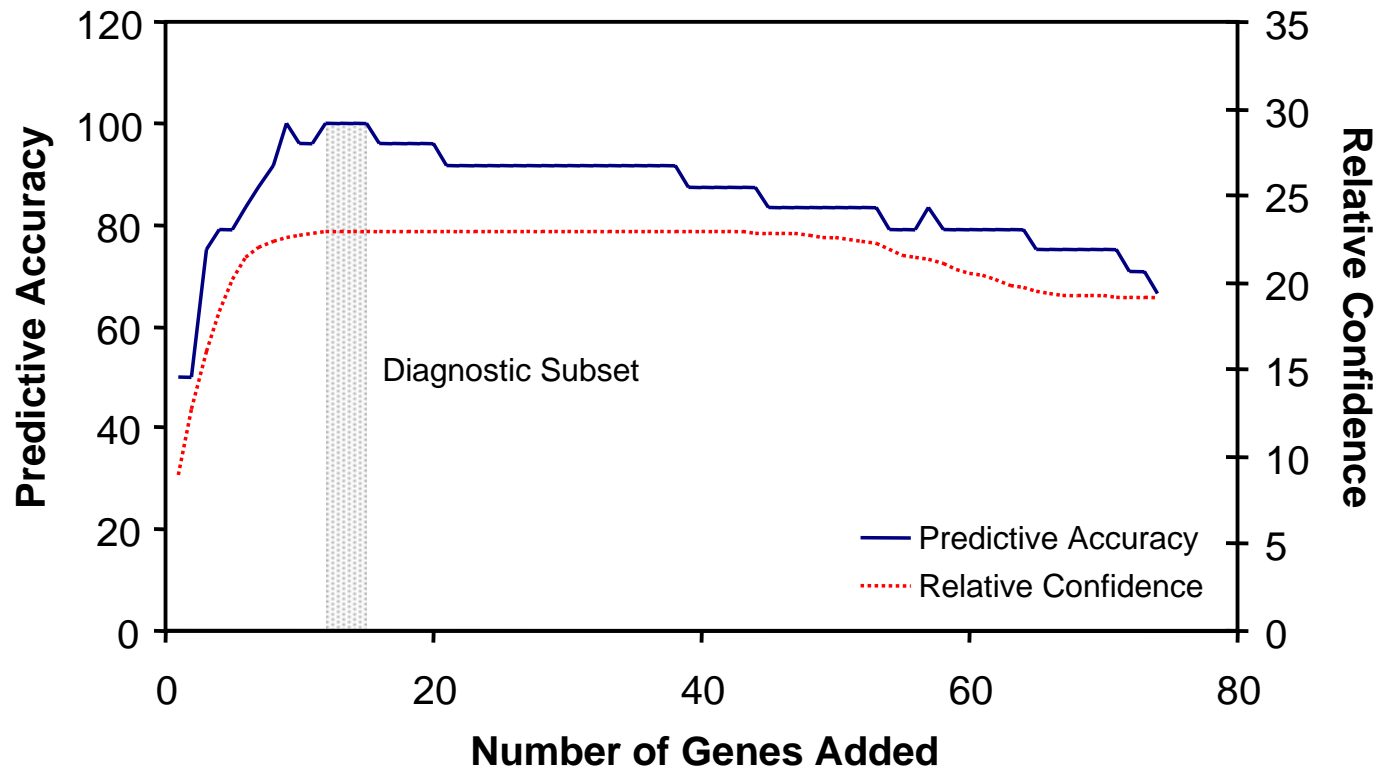


Fig. 2