Joint Genomic and Metabolomic Analysis of Toxic Dose-Response Experiments

Gary L. Jahns[1], Nicholas DelRaso[2], Mark P. Westrick[2], Victor Chan[3], Nicholas V. Reo[4] and Timothy R. Zacharewski[5]

[1] BAE Systems Advanced Information Technologies, 9655 Granite Ridge Dr., ste. 245, San Diego, CA 92123 (gary.jahns@baesystems.com). [2] Human Effectiveness Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH 45433 (Nicholas.DelRaso@wpafb.af.mil; Mark.Westrick@wpafb.af.mil). [3]Alion Science and Technology, Wright-Patterson Air Force Base, OH 45433 (Victor.Chan@wpafb.af.mil). [4]Dept. of Biochemistry and Molecular Biology, Wright State University, Dayton, OH 45435 (nicholas.reo@wright.edu). [5] Dept. of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824 (tzachare@msu.edu).

Abstract

A methodology has been implemented for analyzing microarray and NMR spectral data obtained from the same set of toxic-exposure dose-response experiments. The NMR spectra additionally track the time course of exposure. Analyses consist of screening the data to eliminate variates with insignificant signal, normalization appropriate to the experimental design, Principal Components Analysis, and nonlinear classification using a Support Vector Machine. It is found that exposure at subtoxic levels can be detected.

1. Experiments

The goal of the study reported here was to develop tools to characterize complex, time-evolving changes in gene and metabolite expression patterns as a first step in demonstrating the feasibility of detecting toxic-substance exposure at very low, subtoxic concentrations. The experiments, performed by the Air Force Research Laboratory (AFRL), Wright-Patterson AFB, consisted of orally dosing rats with the liver toxin α-naphthylisothiocyanate (ANIT) over a dose range of 0.1 to 100 mg/kg. Urine samples were collected pre-dose and daily for 4 days for Nuclear Magnetic Resonance (NMR) spectral analysis. Microarray analysis was performed at day-4 post-exposure on liver tissue from 29 of the 39 experimental animals.

The genomic data consist of the results from Affymetrix RAE230A microarrays, one microarray per animal. Each microarray gives 15,866 signal amplitudes and associated p-values corresponding to probe sets associated with specific genes [1]. The metabolomic data specifically are 600 MHz proton (^{1}H) NMR spectra in which amplitudes are summed into 255 bins and normalized by total intensity to account for sample-to-sample density variations. The water peak was experimentally reduced and the residual signal at 4.7-4.9 parts/million was zeroed out. The data for each animal is provided as a single file containing 5 spectra from urine samples taken pre-dose and every 24 hours for 4 days.

2. Analysis

Our analysis procedures consist of (1) screening the data to eliminate signals with no significant signal to reduce dimensionality, (2) normalizing the data [2,3], (3) performing Principal Components Analysis (PCA) [4] to characterize the data in terms of an orthogonal basis determined by the data covariance, and (4) nonlinear classification using a Support Vector Machine (SVM) [5] methodology [6,7].

In the case of genomic data, we found that the first two principal components are sufficient to separate the data. A projection plot of the principal components is shown in Figure 1 (left). For classification, we associated the controls and seven dosage levels into 3 groups: controls, low dose, and high dose (high dose comprising the two highest dosages, which are known to have significant histopathological effects). Applying an SVM classifier with a radial basis

function kernel to this two-dimensional data, we found the decision boundaries shown in the Figure when the one obvious low-dosage outlier was included in the control group. To summarize our methodology, 15,866 gene probe sets across 29 microarrays (replicates of vehicle-only controls and 7 nonzero dosages) were reduced to 7844 probe sets with "Signal Present," then screened to the 100 most significant probe sets, and finally reduced to two principal components for which intuitively natural separation boundaries between control, low-level doses, and high-level doses were found, with misclassification of one 0.1mg/kg case.

Metabolomic normalized signals were concatenated by spectral bins (255) and post-dose days (4) across all replicates to form a 1020-by-39 data matrix, which was then screened down to the 100 most significant bins/times. We performed a modified analysis on just the controls and the low-dose cases, keeping the 100 most significant spectral bins/times according to consistency between 10 and 20 mg/kg doses. The result is shown in Figure 1 (right), again a scatterplot of the first two principal components, where the SVM decision boundary separates the controls and low-dosage points except for two control and one 0.1 mg/kg cases that are misclassified.

In summary, as a first step in taking a systems-biology approach to toxic exposure, we have shown that roughly equivalent results can be obtained by applying the same analysis procedures to data from the genetic and organismic levels of the experimental subjects. We have also developed a normalization methodology for combined time-course and dose-response experiments. Regarding the specific goal of the study undertaken, we have shown that signatures of low-level toxic exposure can be observed.

We gratefully acknowledge help from Jessica Young, Andrew Neuforth, and Michael Ferguson.

3. References

[1] Affymetrix GeneChip Operating Software manual, Appendix E, available at http://www.affymetrix.com /products/software/specific/gcos.affx.

[2] Eckel, J. E., Gennings, C., Therneau, T. M., Burgoon, L. D., Boverhof, D. R., and Zacharewski, T. R., Normalization of two-channel microarray experiments: a semiparametric approach,. Bioinformatics, *21:* 1078-1083 (2005).

[3] Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc., *96:* 1151-1160 (2001).

[4] I. T. Jolliffe, *Principal Component Analysis*, 2nd edition, Springer-Verlag (2002).

[5] V. Vapnik, *Statistical Learning Theory*. Wiley (1998).

[6] S. Rüping, *mySVM-Manual*, University of Dortmund, Lehrstuhl Informatik 8, http://www-ai.cs.uni-dortmund .de/SOFTWARE/MYSVM/ (2000).

[7] LIBSVM—A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

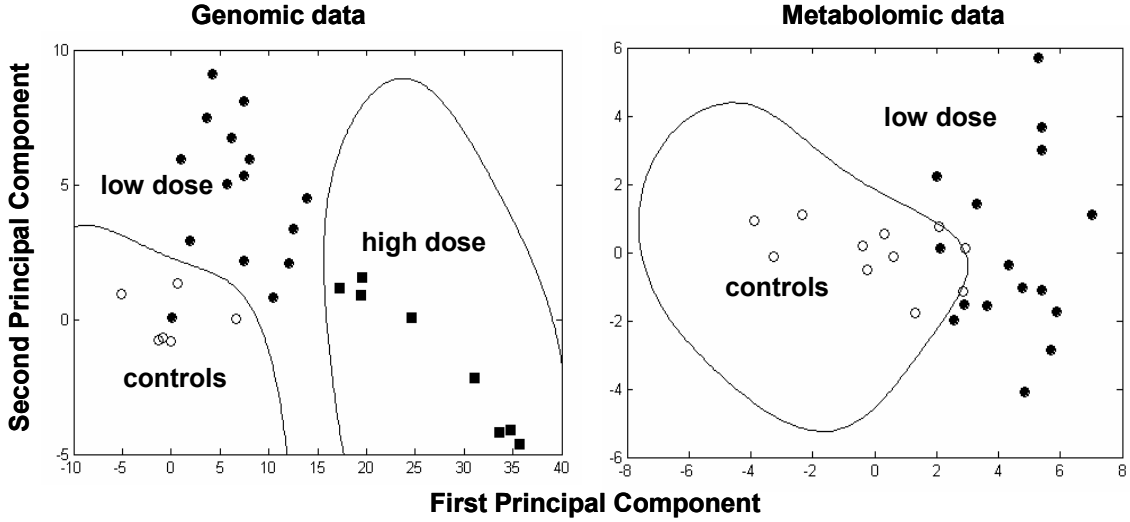

Figure 1. Data from microarray analysis of liver tissue (left panel) and NMR spectra of urine (right panel) reduced to the first two principal components. Contours are decision boundaries between controls, low doses (0.1-20 mg/kg) and high doses (50, 100 mg) determined by a nonlinear classifier.