CYCLOSARIN EXPOSURE DETECTION VIA MICROARRAY DATA ANALYSIS

Andrzej K. Brodzik, John Dileo, Andrea Jensenius, Taehwan Kim, and Olivia Peters
The MITRE Corporation, McLean VA 22102

ABSTRACT

The purpose of this work is to establish feasibility of cyclosarin exposure detection via microarray data analysis and to evaluate the sensitivity of several detection methods to exposure level. First, we test the methods on the Golub leukemia data, and then we apply the methods to multi-level cyclosarin exposure data. Initial results of the investigation suggest that relatively low error rates in detection of a low dose sarin exposure can be obtained using either Bayes classifier, neural networks, or simple class signature matching.

1. INTRODUCTION

While the effects of exposure to high level dose of chemical nerve agents have been widely studied since World War I, so far there has been relatively little attention paid to the effects and detectibility of the low level exposure [6]. The latter became of interest since the reporting of Gulf War syndrome, possibly attributed to the exposure to the cyclosarin vapor. While the low level chemical agent exposure usually does not manifest itself with easily diagnosable pathologies, subtle changes in gene expression levels can occur, which, if confirmed, might be used to design an effective early warning system.

The goal of our work is to investigate the feasibility of such an early detection system and to evaluate its potential sensitivity to exposure level. In this paper, which summarizes the first phase of our investigation, we perform a comparative study of the efficacy of several well known classification methods and their refinements: correlation, principal component analysis, independent component analysis, Bayes, and neural networks. The first and the last two methods appear to be the most effective, with error rates in the range of 4%-20%. We discuss the advantages and disadvantages of all the methods and suggest future improvements. The issue of a possible diagnostic gene set has not been as yet answered satisfactorily and will need to be further explored in future research with the aid of higher dimensional data.

2. APPROACH

The processing included the pre-processing stage, the feature extraction stage, and the classification stage. Three different feature extraction methods and five different classification algorithms were used. The objective of classification was to identify either the correct leucemia type (Golub data) or the correct exposure level (cyclosarin data).

2.1. Data Pre-Processing

Data pre-processing included normalization and gene pre-selection. Normalization was performed across sample number. Genes containing outliers and weakly expressed genes were removed. Outliers were identified by computing for each gene the z-score,

$$z = \max_i \left\{ \frac{x_i - \bar{x}}{s} \right\}, \tag{1}$$

where x_i is the i-th sample normalized gene expression level, $\bar{x}$ is the class mean, and s is the class standard deviation. Gene discriminating power was evaluated by the Welch's modified t-test

$$t = (\bar{x}_1 - \bar{x}_2) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\,, \tag{2}$$

where n_1 and n_2 denote the number of samples, $\bar{x}_1$ and $\bar{x}_2$ denote the means, and s_1 and s_2 denote the standard deviations of groups 1 and 2, respectively. Subsequently, the Empirical Bayes approach [2] was used to estimate *a posteriori* probability of gene expression from the mixture model of affected and unaffected gene probability densities. Genes which were determined not to be differentially expressed, either through the Welch's modified t-test or the Empirical Bayes approach, were removed.

2.2. Feature Extraction

Three different methods were used to identify class differentiating genes. The first method selected genes which had the highest values of the modified t-test (MTT),

$$t_i' = \frac{|m_i^{(A)} - m_i^{(B)}|}{\sigma_i^{(A)} + \sigma_i^{(B)}} \tag{3}$$

where $m_i^{(A)}$, $m_i^{(B)}$, $\sigma_i^{(A)}$, $\sigma_i^{(B)}$ are the median values and standard variations of the training data two classes gene expression values, and i is the gene number.

The second method relied on gene expression profile correlation (GEC) [4]. An ideal profile was given by a binary step function, where 0 was assigned to samples of class 1, and 1 was assigned to samples of class 2. The correlation with the ideal profile was then calculated for each gene, and genes with the highest absolute value of correlation were selected as biomarkers.

The maximum likelihood (ML) method selected genes using the relative differences in their log likelihoods. Call M_i^g the class i Gaussian distribution for gene g. In the two class case, the goal is to find M_1^g more likely than M_2^g given a sample of class 1, and M_2^g more likely than M_1^g given a sample of class 2 using relative log likelihood scores for gene g,

$$\mathrm{LIK}_{1\to 2} = \log p(M_1^g|X_1) - \log p(M_2^g|X_1), \tag{4}$$
$$\mathrm{LIK}_{2\to 1} = \log p(M_2^g|X_2) - \log p(M_1^g|X_2), \tag{5}$$

where X_1 are samples of class 1 and X_2 are samples of class 2. The ideal gene would have both LIK scores much greater than 0. Genes with the highest value of $LIK_{1\to 2} + LIK_{2\to 1}$ are selected (only those with both $LIK_{1\to 2}$ and $LIK_{2\to 1}$ greater than 0 are considered).

The cardinality of the class differentiating gene lists produced by the three methods for both the Golub and the cyclosarin data ranged from 50 to 80.

2.3. Classification

Five different classification algorithms were used. The first method, the Class Signature Matched Filter Contest (CSMFC), relied on class model estimates obtained from processing the training set. This is a version of one of the simplest and most widely used classification methods and is realized by computing cross-correlation between the class model and the unknown sample.

The next two methods, the PCA and ICA Projection Methods (PCAPM and ICAPM), rely on projections of the classification data matrix onto subspaces spanned by subsets of the principal or independent components of the matrix of training and classification data. The main motivation for use of PCA and ICA is noise reduction and de-coupling of the disease-induced gene expression pattern from competing gene expression patterns, respectively.

The fourth classification method used artificial neural networks (ANN). A simple multi-layer perceptron with one hidden layer was used to classify samples. The advantage of the ANN method is robustness to nonlinearities, noise and missing samples.

The last method tested was the naïve Bayes classifier (NBC) [5]. NBC classifies a test sample by determining the class model for which the sample has the largest *a posteriori* probability. A Gaussian sample distribution has been assumed in this method.

3. GOLUB LEUKEMIA DATA

The feature extraction and classification methods described in the previous section were applied to the Golub leukemia data to discriminate the AML and ALL leukemia types [4]. The training set consisted of 38 samples, 27 ALL and 11 AML, while the independent testing set was comprised of 34 samples, 20 ALL and 14 AML.

The detection rates (defined as the number of correct class assignments divided by the number of samples) ranged from 76.5% to 97.1% (Table 1). The best results were obtained using MTT/ CSMFC (97.1%) and GEC/ANN (95.6%) feature selection/ classification approaches. These results are similar to the results cited in the literature (97.1-88.2% with self-organizing maps [4], 88.2-82.4% with between-group analysis [1], and 91.2% with bagged clustering procedures [2]).

Method	Detection rate [%]
MTT/CSMFC	97.1
MTT/PCAPM	91.2
MTT/ICAPM	76.5
ML/NBC	88.2
GEC/ANN	95.6

Table 1: Classification results for the Golub leukemia data. Method refers to the combination of different feature extraction and classification algorithms.

4. CYCLOSARIN DATA

Cyclosarin is a colorless liquid organophosphate nerve agent. Its primary mechanism of action is to inhibit Acetylcholinesterase (AChE), causing accumulation of Acetylcholine (ACh) at the synapses. This results in hyperstimulation of the muscarinic and nicotinic ACh receptors.

The data was collected from low-level whole body inhalation exposure of rats to cyclosarin vapor. Three different exposure levels (0.004, 0.0134 and 0.0251 mg/m^3) were chosen to investigate the effects of low-level chemical nerve agent exposure. After exposure, the rat brain tissues were collected and snap frozen, the mRNA was extracted, and following the process described in [6] mRNA expression levels for 8799 genes and 90 samples were collected from the Affymetrix microarray data.

The data was split into independent training and testing sets, with each set containing 15 control samples and 10 samples from each of the exposures. All groups were evenly split between male and female samples. The training set was used to extract differentiating genes, the test set was used in the classification stage. Table 2 summarizes results of processing.

Method	Detection rate [%]		
	0.004 mg/m^3	0.0134 mg/m^3	0.0251 mg/m^3
MTT/CSMFC	92.0	84.0	84.0
MTT/PCAPM	64.0	60.0	84.0
MTT/ICAPM	72.0	80.0	88.0
ML/NBC	84.0	80.0	88.0
GEC/NN	96.0	84.0	88.0

Table 2: Classification results for the cyclosarin data.

5. CONCLUSIONS

We have compared three feature extraction methods and five classification methods for cyclosarin exposure detection efficacy. Among those, three approaches (MTT/CSMFC, ML/NBC and GEC/ANN) performed consistently well. The error rates obtained with the Golub leukemia data (2.9-11.8 $\%$) are comparable to the results published in literature obtained by use of more sophisticated methods. The error rates obtained with the cyclosarin data are markedly higher (4-20 $\%$), which might be due in part to low data dimensionality and a relatively low level of exposure. Further refinements of the gene selection procedure, the classification algorithms, and, most of all, analysis of larger data sets will need to be undertaken to arrive at final determination of feasibility of a practical cyclosarin exposure early detection system.

6. REFERENCES

[1] Culhane, A.C., et al. "Between-group analysis of microarray data," *Bioinformatics* 18(12): 1600-1608, 2002.

[2] Dudoit, S., and Fridlyand, F. "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics* 19(9): 1090-1099, 2003.

[3] Efron, R., Tibshirani, R., Storey, J., and Tusher, V. "Emperical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association* 96(456): 1151-1160.

[4] Golub, T.R., et al, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science* 286: 531-537, 1999.

[5] Keller, A., Hood, L., Schummer, M., and Ruzzo, W. "Bayesian classification of DNA expression data," Department of Computer Science and Engineering, University of Washington Seattle, 2000.

[6] Sekowski, J.W., et al. "Gene expression changes following low level exposure to sarin vapor," 23rd Army Science Conference, December 2-5, 2002, Orlando, Florida.