BIOMARKER DETECTION FOR HEXACHLOROBENZENE TOXICITY USING GENETIC ALGORITHMS ON GENE EXPRESSION DATA

Cem Meydan[1], Alper Küçükural[2], Deniz Yörükoğlu[3], and O. Uğur Sezerman[4]

Biological Sciences and Bioengineering, Sabanci University
Sabancı Üniversitesi, Orhanlı-Tuzla, 34956 İstanbul, Türkiye
phone: + (90)2164839000, fax: + (90)2164839550,
email: {cemmeydan[1], kucukural[2], denizy[3]}@su.sabanciuniv.edu, [4]ugur@sabanciuniv.edu
web: http://www.sabanciuniv.edu

ABSTRACT

Discovering toxicity biomarkers is important in drug discovery to safely evaluate possible toxic effects of a substance in early phases. We tried evolutionary classification methods for selecting the important classifier genes in hexachlorobenzene toxicity using microarray data. Using genetic algorithms for selection of minimum number of features for classification of gene expression data, we discovered a number of gene sets of size 4 that were able to discriminate between the control and the hexachlorobenzene (HCB) exposed group of Brown-Norway rats with >99% accuracy in 5-fold cross-validation tests, whereas classification using all of the genes with SVM and other methods yielded results that vary between 48.48% to 81.81%. Making use of this small number of genes as biomarkers may allow us to detect toxicity of substances with mechanisms of toxicity similar to HCB in a fast and cost efficient manner when there are no emerging symptoms.

1. INTRODUCTION

Finding reliable toxicity biomarkers is important in toxicogenomics to safely evaluate possible toxic effects of a substance in early phases of drug discovery. Discovering the important mechanisms of toxicity for known toxic substances and developing biomarkers that detect these can lead to classification of new substances with respect to their toxicity in a cost-efficient manner.

Using microarray technology to evaluate the changes in gene expression data between control and experiment data sets, the significant set of genes that indicate the existence of the toxicity may be obtained. Discovering these genes that are correlated with the substance may point the mechanisms of toxicity and the effected pathways. These sets of genes may also be used for development of diagnostic kits that can be used to detect possible existence of toxicity of a substance on the test subjects, or on the early diagnosis of toxin exposure.

Minimizing the number of genes that are used as a biomarker, without affecting the accuracy of the prediction, is essential for practical purposes. Each redundant control will have a negative effect time-, complexity- and cost-wise, therefore finding the minimal set of genes with highest classification accuracy is in practical interest. Feature subset selection refers to this problem of selecting important set of attributes from a large set of redundant attributes that are uncorrelated with the class used in classification purposes.

The method proposed by Kucukural et al. successfully selects the minimum number of features for classification of gene expression data using evolutionary methods with low attribute counts and very high accuracy in cancer data sets compared to other methods [1]. The viability of this method on toxicity datasets was unexplored, and if any changes were necessary for adapting it to toxicogenomics. To compare its performance, we tried this evolutionary classification method and other classifiers for selecting the minimum number of important set of genes in hexachlorobenzene toxicity using microarray data.

Hexachlorobenzene (HCB) is an organochlorine fungicide with persistent environmental pollution effects, and has various toxic mechanisms in man. During the period 1955-1959, about 4000 people in southeast Anatolia in Turkey developed porphyria due to the ingestion of HCB that were used on wheat seedlings [2]. HCB has been also classified as a Group 2B carcinogen (possibly carcinogenic to humans) by the International Agency for Research on Cancer (IARC). Animal carcinogenicity data for hexachlorobenzene show increased incidences of liver, kidney (renal tubular tumours) and thyroid cancers [3].

Although HCB usage was banned in most areas, it is still generated as waste by-products of industrial processes. The pollution of the sea coasts and groundwater persists due to its stability, and HCB is still detectable in human milk and blood in some areas of the world.

In this study HCB data set is used because its effects, mechanisms of toxicity, the pathways affected and the toxicology data such as oral and inhalation dosage in mice, rats and humans are well documented, therefore the obtained set of classifier genes may be compared with the literature.

2. RELATED WORK

Studies for finding genomic markers for detection of changes in an organism are aimed at different reasons. One is mainly for practical diagnostic purposes, and other is for discovering the underlying mechanism in that change. Although both can be used for other purposes as well, the goal in finding diagnostic markers is to minimize the number of needed data without affecting accuracy.

If the toxin causes a response in gene expression level, microarray technology is very powerful for biomarker discovery [4-5]. The entire human genome can be contained on a single microchip, enabling us to generate complete profile of the response to toxicity [6-8]. Representational difference analysis (RDA) of tissue- or cell-specific arrays are used to find candidate biomarker proteins from protein-coding genes that have a specific change in expression, when control and experimental values are compared [9-10]. However, using only genomic data is insufficient, since it only measure changes in mRNA expression, but abundant quantities of mRNA does not necessarily equal abundant quantities of protein [5], thus semi-quantitative assays are required for checking the proteins. An example to this is the study from Ichimura et al. who identified the gene with most significant change in the postischemic rat kidney, KIM-1, and confirmed its viability for use as a biomarker by subsequent immunoblot, immunostaining, and RNA in situ hybridization [9]. Examples to other similar studies in which RDA in microarrays is used to find markers and changed gene expression levels for damage due to radiation toxicity [11], hemolytic anemia induced by drugs [12], nephrotoxicity induced by cisplatin [13], identification of glutathione depletion-responsive genes in rat liver [14], and many more. This approach is powerful, and pharmaceutical and biotechnology companies even created panels of biomarkers that detect drugs causing hepatotoxicity upon in vitro exposure to rat hepatocytes or in vivo dosing in rats [15].

For markers of toxicity, studies are mostly in mechanistic field. The work by Ezendam et al. [16], whose data set we also studied, tried to find the significant changes in gene expression due to hexachlorobenzene exposure, and found a total of 104 genes in different tissues that are affected by the HCB. Similar works concentrating on changes in expression levels in specific tissues or on the whole organism due to exposure to several chemicals are also numerous.

Other works in which the biomarker discovery by feature selection is studied are also present. Many studies concentrate on heuristic search and randomized population based techniques such as genetic algorithms. Early studies used known computational procedures such as greedy optimization, branch and bound, tabu search, simulated annealing, gibbs sampling, evolutionary programming, genetic algorithms, ant colony optimization and particle swarm optimization [17-24]. These perform differently in different conditions due to the heuristic methods; no best solution can be found due to the practically non-computable number of possible solutions, making exhaustive search impossible.

Recently, feature selection by coupling genetic algorithms with statistical classifiers have been studied. Alon et al. used clustering algorithms to find genes with correlated expression levels that can be used for diagnosis. They found a set of genes that can classify colon cancer by 90% accuracy [25]. On the same data set, Fröhlich et al. used genetic algorithms coupled with support vector machines to find minimum number of genes that can classify the data, which resulted in a set of 30 genes with 85% accuracy [26]. The work by Kucukural et al. that we used in this study concentrates on dynamic parent generation by fitness score of features using genetic algorithms, and was very efficient in finding a low number of highly accurate solutions, resulting in 98% accuracy using 12 genes in the same colon cancer data set and 100% accuracy with 12 genes in ovarian cancer data set [1].

3. METHODOLOGY

To discover the minimum number of features that can classify the data, we have to find a way represent these sets of genes. In the genetic algorithm, each individual in the population represents a candidate solution to the feature subset selection problem. The genetic code of a parent is boolean vector of size m, where m is the number of attributes. A value of 1 means the parent has that attribute, and 0, not. Since there are 2^m possible parents, exhaustive search is impossible with more than a handful of attributes, thus evolutionary algorithms or other heuristic methods are necessary.

The method is based on the selfish gene idea by Richard Dawkins, in which the individuals are only the carriers of genes, and the function of a parent is to leave the strong genes to next generation [27]. Thus, the main goal is the survival of the gene. If an individual has a good fitness score with fewer genes, the gene (i.e. feature) count can be decreased. Basically, the algorithm uses this concept to select features. Details of the genetic algorithm are given below.

The main base of the genetic algorithm uses standard mutation and cross-over algorithms, using support vector machine (SVM) learning to assign a fitness score. SVM is a very accurate supervised learning method, widely used in computational problems in biology. Score of an individual is calculated by the accuracy of classification by SVM in 5-fold cross-validation tests.

The genetic algorithm employs elitism in which low scoring children are replaced with best scoring parents, if the score of the parents are higher. Also a small number of "bad" parents are also selected, which keeps the algorithm from being stuck in local minima, thus acting as simulated annealing along with the roulette wheel method.

Used SVM parameters are given below. LibSVM library [28-29] was used in both the genetic algorithm and in the following tests.

In the first generation, parents are randomly generated with each having a set number of features and each feature

being covered a set amount, for reducing the chance of a good attribute being dropped due to redundant neighbours in that parent. Then, for a number of generations (given as non-reducing generations below) the genetic algorithm runs without trying to reduce the number of features. In this phase each feature is assigned the fitness score of its parent. After the first run for a set number of generations, average fitness score for each attribute is obtained by dividing the total fitness score by the number of times that feature was chosen in an individual.
After this first pass is completed, the reducing phase of the genetic algorithm with roulette wheel based selection strategy is used in which the number of features is reduced for filtering the redundant attributes. In roulette wheel the probability of selecting a feature is its fitness score, therefore high scoring attributes are selected more often. Each child carries a specific gene set generated by the roulette wheel selection, crossover and mutation steps, and when a gene is selected more than once for a specific set, the duplicates will be removed, consequently decreasing the total feature count. This way the number of features of a child decreases if the same attribute is selected more than once. This parent generation scheme, which focuses on "gene" instead of parent, allows the dynamic selection of optimal number of features. The effectiveness of this approach can be clearly seen in Figures 1 and 2.

The genetic algorithm uses the following parameters;

Number of Features: 8799
Feature Coverage (the number each feature is covered in the first generation): 2
Number of features in each parent in 1st generation: 30
Number of non-reducing generations: 50
Number of reducing generations (see above): 500
Population Size: 587 ((# of features)x(feature coverage) / (# of features in each parent))
Crossover Rate: 0.9
Mutation Rate: 0.1
Elite Parent Rate: 0.2
Bad Elitist Rate: 0.01

SVM Parameters:
Type: C_SVC
Kernel: Radial Basis Function
C: 100
Gamma: 1.0 / 8799
Coefficient0: 0
Epsilon: 0.001
P: 0.1
Using normalization and shrinking. Not using probability estimates.

The algorithm was run for 10 iterations for each 0-150, 0-450, and 150-450mg/kg comparison (see Sections 4.1 and 4.2). After the genes are selected iteratively, different SVM parameters are used to find the optimal SVM score. As kernel types both radial basis functions (RBF) and linear (LIN) functions are used, with cost parameters

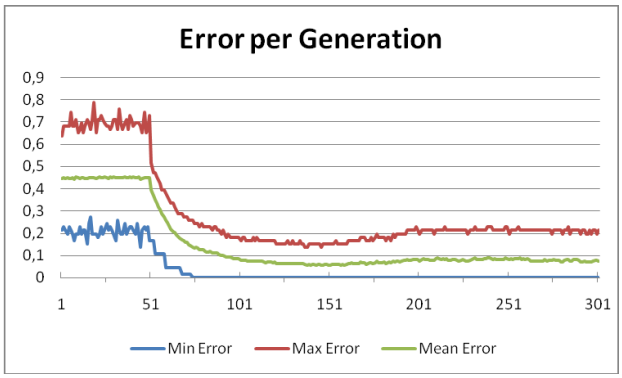

Figure 1: Min/Max/Mean errors in run of the algorithm with respect to generation in set 1. The given errors are the proportion of the false positives and false negatives in the whole test set. Notice that after 50 generations, algorithm changes its selection strategy from non-reducing to reducing (see text), and the min. error rate quickly converges to 0.

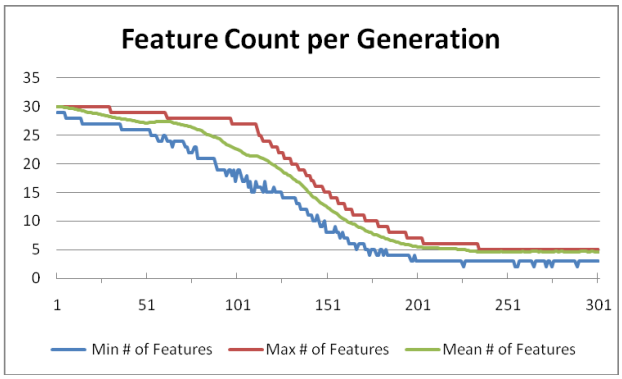

Figure 2: Number of features in the population with respect to generation. While the error rate becomes 0 at about 75 (see Figure 1 above), the feature count gradually decreases until it reaches a minimum of 4 in about 200th generation. Although there are some selected ones with 3 features, they are not able to classify with 100% accuracy and are dropped against the sets with 4 features.

100 and 10000. The results can be seen in Table 2. By optimization of the other parameters more accurate solutions can be found [28], however the results obtained were ~100%, thus no optimizations were necessary.

4. RESULTS

4.1. Data Set

The microarray data used is from the study Ezendam et al. in Dutch National Institute for Public Health [16]. Ezendam et al. fed Brown Norway rats with diets supplemented with 0, 150 and 450mg/kg HCB for 4 weeks, after which spleen, mesenteric lymph nodes (MLN), thymus, blood, liver, and kidney were collected and analyzed using the Affymetrix rat RGU-34A GeneChip microarray. Using 1 microchip per tissue per animal, they obtained a total of 96 hybridizations; 35 from untreated control group, 30 from 150mg/kg and 31 from 450mg/kg exposed group, each having 8799 genes of RGU-34A chip.

Method	Accuracy			
	Using all 8799 Genes in 0-450mg/kg		Using the 4 Genes in Set 3 (See Appendix)	
	5-fold cross-validation	Test on training data	5-fold cross-validation	Test on training data
SVM	53,03%	100%	100%	100%
Decision Table	71,21%	96,97%	77,27%	86,36%
Naive Bayesian	65,15%	100%	77,27%	78,78%
Naive Bayes Multinomial (updateable)	54,55%	100%	93,94%	93,94%
Multi-layer Perceptron	N/A	N/A	100%	100%
RBF Network	43,94%	100%	92,42%	100%
Simple Logistics	62,12%	100%	95,45%	100%
Random Forest	68,18%	100%	90,90%	100%
NBTree	75,76%	100%	78,79%	95,45%
C4.5 Decision Tree	66,67%	96,97%	93,94%	96,97%
IBk (k-nearest neighbour classification)	81,81%	100%	96,97%	100%
KStar	N/A	N/A	95,45%	100%
Classification via clustering (N=2)	48,48%	50%	45,45%	53,03%

Table 1: Accuracy of classification with different methods using both unfiltered 8799 features and the selected 4 features (set 3), in cases of 5-fold CV and testing on training data. Parameters are default values if not stated, except in SVM, which uses the parameters shown in section 3. N/A values were not classified due to high memory requirements. Post-selection accuracy values showing increase are indicated in green, showing decrease in red.

4.2. Experiments

HCB has relatively low acute toxicity, but it is cumulative in lipid tissues, and is persistent. According to Extension toxicology network data, oral toxic dosage of HCB is 10,000mg/kg [30] in rats. Rats exposed to 450mg/kg per day for 28 days accumulate about 12600mg/kg of HCB, thus showing the most toxic effects [30-32]. According to the previous works that studied the posology of HCB, the time of onset, severity and size of skin lesions, the increase in body, liver and spleen; in other words the toxic effects of HCB with respect to exposure, ideal dose of exposure in which pathological problems are seen in rats is thus 450mg/kg, so comparison between 0 and 450mg/kg is studied in most detail to find the markers of sub-chronic exposure. One rat in 450mg/kg exposed group even died after 25 days of exposure [16], thus we expect the changes in gene expression levels to differ more from the 0-150mg/kg comparison. However, for early prediction, the low-dose markers and studies for 0-150mg/kg and also the classification of dosage by 150-450mg/kg are also given, though not as detailed.

The found marker genes are then analysed and classified with other methods in the data mining environments WEKA (Waikato Environment for Knowledge Analysis from University of Waikato) [33] and RapidMiner (formerly known as YALE) [34]. The genes are also searched in Affymetrix NetAffx Analysis Center [35] and in Kyoto Encyclopedia of Genes and Genomes (KEGG) [36-38] for correlation between chip result and gene functions and pathways.

	Feature Set	SVM Accuracy			
		RBF, 100	RBF, 10000	LIN, 100	LIN, 10000
0-150	Set 1	0.991385	0.992	**0.999385**	0.998769
	Set 2	0.996154	0.996308	**0.987538**	0.986462
0-450	Set 3	0.993182	0.996515	0.997576	**0.998636**
	Set 4	0.991364	0.993636	0.988182	**0.995909**
150-450	Set 5	**0.997213**	0.991148	0.941148	0.938525
	Set 6	**0.990656**	0.984098	0.971475	0.97377

Table 2: Accuracy of SVM classification on highest scoring 2 sets for each comparison. Note that all sets have practically 100% classification accuracy in 5-fold cross-validation due to sample size of 66. Highest values for each set are shown in green.

4.3. Results

We discovered a number of gene sets of size 4 that were able to discriminate between the control and the 450mg/kg HCB exposed group of Brown-Norway rats with >99% accuracy by SVM classification in 5-fold cross-validation tests, whereas classification using all of the genes with the same methods such as SVM, Naïve Bayesian, C4.5 decision tree, RandomForest yielded results that vary between 48,48% to 81.81% (Table 1). Similarly, in 0-150mg/kg 7 genes, and in 150-450mg/kg 6 genes that increased the accuracy dramatically were discovered. Since the changes are more subtle, it is normal for them to have more features to classify correctly.

SVM scores for selected sets for all comparisons, with respect to different parameters may be seen in Table 2. While SVM scoring is given for the distance to the classifier vectors, for actual classifying purposes all of the data were classified correctly in 5-fold cross-validation tests, giving 100% accuracy for all sets using our data set.

Note that only the best scoring 2 sets are shown in Table 2. For 0-450, 8 of the iterations gave 4, 1 iteration gave 3 and 1 gave 5 genes, for an average of 4 genes per set. The distributions were similar for others, with few less than and few more than 6 and 7.

As seen in Table 1, since these genes are discriminative, any classifier can be used, not only SVM. Using multi-layer perceptron classification on set 3 also gave 100%, C4.5 gave 93.94%, k-nearest neighbour clustering gave 96.97% accuracy. Thus, without using SVM, accurate, simple, human readable decision trees of size 9 with 5 leaves can be built easily using these genes. One example is given for C4.5 decision tree in Figure 3, with accuracy of 96.97% in whole data. The confusion matrix of the tree is given in Table 4. It can be seen that for this feature set and data, all of the errors of the type I, i.e. false positives. Depending on the indicator being developed, the genetic algorithm or the decision tree algorithm may be modified to favour one type of error over the other when selecting classifier features.

Accession Number	Gene Symbol	Description	# of times selected		
			0-150	0-450	150-450
D00913_g_at	Icam1	intercellular adhesion molecule 1	1	8	0
AF093139_s_at	Nxf1	nuclear RNA export factor 1 (mRNA_processing_Reactome)	0	3	0
rc_AA892154_g_at	Mxd4_predicted	Max dimerization protein 4 (predicted)	0	3	0
rc_AA892325_at	Cept1	choline/ethanolamine phosphotransferase 1	0	2	0

Table 3 : Details of the genes in set 3 and cross selection of these genes in other sets.

An important finding is that, while filtering features (and thus decreasing the information content) mostly reduced the classification accuracy using all of the training data for testing, at the same time it increased the accuracy of 5-fold cross-validation dramatically. We can conclude that the 100% accuracy in testing by training data is due to overfitting, and when the redundant features (i.e. noise) are filtered, the classifiers actually work much better in real world conditions. Reducing the feature count not only decreases the test cost and time/memory requirements, but also increases the accuracy.

In 10 iterations, about 40 to 70 attributes are selected in total for each run. The selection of attributes shows half-normal distribution; a small number of genes appear in most of the sets while most of them only appear in one. It is possible that by running the algorithm for more iterations and selecting the most common elements may give more robust solutions for use in real world.

However, while intra-class gene counts are half-normally distributed, inter-class gene similarity is very low. As it can be seen in Table 3, in the genes selected for 0-450 separation are selected, just 1 of them is selected once in 0-150 and 150-450 classification. The results from other sets are similar. Although these occurrences are more frequent than randomly selecting the same genes from a pool of 8799 genes, they are still somewhat lower considering the selected genes are effected by HCB exposure.

Kucukural et al. compared the results on colon and ovarian cancer with other studies in terms of number of genes and accuracy. However, there are no studies in HCB toxicity focusing on minimum number of predictive genes. Study by Ezendam et al. focused on microarray data for important genes in various tissues that have role or affected by HCB toxicity, and although feature selection was done for selecting significant ($p < 0.001$) expression level changes, the number was not minimized. Nevertheless, 45 changed genes in spleen, 16 genes in MLN, 7 genes in thymus, 27 in blood and 19 in liver were detected. Of those, M63122, D00913, K00996, AA891209 and E00778 are also present in the genes we selected.

To compare the number of features given by the algorithms, we used the feature selection using genetic algorithm module in RapidMiner. The GA feature selection module was coupled with 5-fold validation test and SVM scoring method with the same parameters used in our algorithm. The module, however, was too slow and low generation counts were used. The number resulting selected features was about ~600, thus it only eliminated uncorrelated features and not tried to select the minimum number of genes that could classify the data.

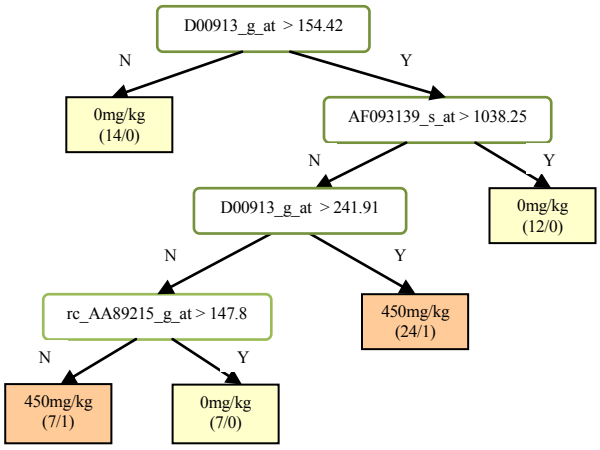

Figure 3: C4.5 decision tree built by the marker genes in set 3. Shown for demonstration purposes. The correct and erroneous decision counts are given in parentheses. Total accuracy is 96.97%, training and testing using all of the data

classified as			
0	**450**		
33	2	**0**	actual value
0	31	**450**	

Table 4: Confusion matrix of decision tree created by classification of set 3, shown in Figure 3.

The decision table method (results shown in Table 1), when trained on all of the genes, has built 10 rules based on 4 genes (selected genes shown in appendix 7.1), but there were no similarities between the genes selected by decision table and by our method, and the accuracy of these rules were lower than ours.

5. CONCLUSION AND DISCUSSION

The dynamic parent generation with emphasis on feature suitability for classification performs much better than the standard genetic algorithm in which the chance of good features coming together is dependent on crossover probabilities. In the standard approach heuristics are used only

to decrease the search space, while this method tries to select the optimal features. Along with parents (i.e. sets of features), features themselves get a score and these features are then put into survival. With a higher fitness score, a feature gets selected more often while others are not selected at all and eliminated, and better features have a chance to get selected more than once in an individual, thus decreasing the number of features in the parent.

Although no prior work on biomarker detection for HCB induced toxicity was done, we compared the method with other well known classifier algorithms, in which the selected gene count and accuracy was better in our case. To compare our accuracy, we classified the data without any feature selection. Classifying the data using all of the genes, without any selection, results in lower accuracy. Thus, low number of genes used as biomarkers favour both accuracy and financial/computational costs of the tests.

With this study, we concluded that biomarkers of toxicity can be discovered with the method by Kucukural et al, and possibly with even better results than cancer in this case of HCB. This method resulted in only 4 features out of 8799 that can classify HCB exposure in Brown Norway rats with 100% accuracy, which is higher than the results obtained in colon cancer (12 features, 98.38% accuracy) and ovarian cancer (12 features, 100% accuracy) [1]. Comparison of the selected genes with the literature [16, 34-37] showed that most of these genes are known for their functions in the pathological effects of toxicity induced by HCB. Using the low dose classifiers and other markers obtained by further study, toxin exposure can be detected when there are no emerging symptoms, which can be used in both medicine and experimental drug discovery. Studying only the changes in blood cells can lead to unintrusive markers that can be used detect the toxicity or disease from only blood samples.

Another point of use for these markers is that, by gathering data from various toxic studies and finding reliable biomarkers of toxicity pathways for different mechanisms of toxicity (e.g. in immunotoxicity; immunosuppression, immunostimulation, hypersensitivity reactions, autoimmune reactions, etc.) can allow us to generate toxicity assays that are more efficient and accurate than the ones used today, which will allow detection of toxicity in very early stages of drug discovery, thus saving much time, effort and money.

6. REFERENCES

[1] Küçükural, R. Yeniterzi, S. Yeniterzi, and O. U. Sezerman, "Evolutionary selection of minimum number of features for classification of gene expression data using genetic algorithms", in *Proc. of the 9th Annual Conference on Genetic and Evolutionary Computation* (London, England, July 07 - 11, 2007). *GECCO '07. ACM*, New York, NY, 401-406, 2007.

[2] Gocmen, H. A. Peters, D. J. Cripps, G. T. Bryan, C. R. Morris, "Hexachlorobenzene episode in Turkey.", *Biomed Environ Sci*, vol. 2(1) pp. 36-43, Mar. 1989.

[3] International Agency for Research on Cancer, *IARC Monographs on the Evaluation of Carcinogenic Risk to Humans*, World Health Organisation, vol.79, pp. 493-567, 2001.

[4] F. B. Collings, and V. S. Vaidya, "Novel technologies for the discovery and quantitation of biomarkers of toxicity", *Toxicology* (2008), doi:10.1016/j.tox.2007.11.020.

[5] J. D. Tugwood, L. E. Hollins, and M. J. Cockerill, "Genomics and the search for novel biomarkers in toxicology", *Biomarkers*, vol.8(2), pp.79-92. Review, Mar-Apr 2003.

[6] M. Arcellana-Panlilio, and S. M. Robbins, "Cutting-edge technology: I. Global gene expression profiling using DNA microarrays". *Am.J. Physiol. Gastrointest. Liver Physiol*. vol.282, pp.G397–G402, 2002.

[7] D. H. Geschwind, "DNA microarrays: translation of the genome from laboratory to clinic", *Lancet Neurol*. vol.2, pp.275–282, 2003.

[8] S. Ishkanian, C. A. Malloff, S.K. Watson, R. J. DeLeeuw, B. Chi, B. P. Coe, A. Snijders, D. G. Albertson, D. Pinkel, M. A. Marra, V. Ling, C. MacAulay, and W. L. Lam, "A tiling resolution DNA microarray with complete coverage of the human genome", *Nat.Genet.* vol.36, pp.299–303, 2004.

[9] T. Ichimura, J. V. Bonventre, V. Bailly, H. Wei, C. A. Hession, R. L. Cate, and M. Sanicola, "Kidney injury molecule-1 (KIM-1), a putative epithelial cell adhesion molecule containing a novel immunoglobulin domain, is up-regulated in renal cells after injury", *J. Biol. Chem.* vol.273, pp.4135–4142, 1998.

[10] M. Hubank, and D. G. Schatz, "Identifying differences in mRNA expression by representational difference analysis of cDNA", *Nucleic Acids Res.* vol.22, pp.5640–5648, 1994.

[11] J. Kruse, and F. A. Stewart, "Gene expression arrays as a tool to unravel mechanisms of normal tissue radiation injury and prediction of response", *World J Gastroenterol*, vol.13(19) pp.2669-2674, 2007.

[12] M. Rokushima, K. Omi, K. Imura, A. Araki, N. Furukawa, F. Itoh, M. Miyazaki, J. Yamamoto, M. Rokushima, M. Okada, M. Torii, I. Kato, and J. Ishizaki, "Toxicogenomics of drug-induced hemolytic anemia by analyzing gene expression profiles in the spleen", *Toxicol Sci,*. vol.100(1) pp.290-302, Nov 2007.

[13] Q. Huang, R. T. Dunn 2nd, S. Jayadev, O. DiSorbo, F. D. Pack, S. B. Farr, R. E. Stoll, and K. T. Blanchard, "Assessment of cisplatin-induced nephrotoxicity by microarray technol-

ogy", *Toxicol Sci.,* vol.63(2) pp.196-207, Oct 2001.

[14] Kiyosawa N, Uehara T, Gao W, Omura K, Hirode M, Shimizu T, Mizukawa Y, Ono A, Miyagishima T, Nagao T, Urushidani T, "Identification of glutathione depletion-responsive genes using phorone-treated rat liver", *J Toxicol Sci,.* vol.32(5) pp.469-86. Dec 2007.

[15] D. L. Mendrick, "Genomic and Genetic Biomarkers of Toxicity", *Toxicology* (2007), doi:10.1016/j.tox.2007.11.013.

[16] J. Ezendam, F. Staedtler, J. Pennings, R. J. Vandebriel, R. Pieters, P. Boffetta, J. H. Harleman, and J. G. Vos, "Toxicogenomics of subchronic hexachlorobenzene exposure in Brown Norway rats", *Environ Health Perspect*., vol.112(7), pp. 782–791, May 2004.

[17] Schölkopf, A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002

[18] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*, 2nd ed.John Wiley and Sons, New York, 2001.

[19] Webb, *Statistical Pattern Recognition*, Wiley, New York, 2002.

[20] V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New York, 1998.

[21] M. Raymer, W. Punch, E. Goodman, L.Kuhn, and A. Jain, "Dimensionality Reduction Using Genetic Algorithms", *IEEE Transactions on Evolutionary computing*, 2000.

[22] F. J. Ferri, V. Kadirkamanathan, and J. Kittler, "Feature Subset Search using Genetic Algorithms", *IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, Essex, 1993.

[23] M. Richeldi, P. Lanzi, "A Tool for Performing effective feature selection by investigating the deep structure of the data", in *Proc. of the International Conference on Tools with Artifcial Intelligence*, pp. 102 - 105, 1996.

[24] H. Witten, and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[25] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays", *Cell Biology*, 96:6745-6750, 1999.

[26] Holger Fröhlich, "Feature Selection for Support Vector Machines by Means of Genetic Algorithms", *Diploma Thesis in Computer Science*, University Marburg, 2002

[27] R. Dawkins, The Selfish Gene -- new edition, Oxford University Press, 1989.

[28] Chang, C. Lin, "LIBSVM : a library for support vector machines", 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[29] Y. EL-Manzalawy, and V. Honavar, "WLSVM : Integrating LibSVM into Weka Environment", 2005. Software available at http://www.cs.iastate.edu/~yasser/wlsvm

[30] Extoxnet. The Extension Toxicology Network. 1996a. P.I.P. Pesticide Information Profiles, “Hexachlorobenzene”. http://pmep.cce.cornell.edu/profiles/extoxnet/haloxyfop-methylparathion/ hexachlorobenzene-ext.html

[31] Michielsen, S. Zeamari, A. Leusink-Muis, J Vos, and N. Bloksma, "The environmental pollutant hexachlorobenzene causes eosinophilic and granulomatous inflammation and in vitro airways hyperreactivity in the Brown Norway rat". *Arch Toxicol*, vol. 76 pp. 236-247.

[32] N. Imai, T. Ichihara, K. Nabae, A. Hagiwara, S. Tamano, and T. Shirai Tomoyuki, "Dose Dependent Promoting Effects of Hexachlorobenzene on Hepatocarcinogenesis in a Rat Medium-Term Liver Bioassay", in *Proc. of the 32nd Annual Meeting of Carcinogenicity*.

[33] S. R. Garner, “WEKA: The waikato environment for knowledge analysis”, in *Proc. of the New Zealand Computer Science Research Students Conference*, pp.57-64, 1995.

[34] Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid Prototyping for Complex Data Mining Tasks", in *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)*, 2006.

[35] G. Liu, A. E. Loraine, R. Shigeta, M. Cline, J. Cheng, S. A. Chervitz, D. Kulp, and M. A. Siani-Rose, “NetAffx: affymetrix probeset annotations”. in *Proc. of the 2002 ACM Symposium on Applied Computing* (Madrid, Spain, March 11 - 14, 2002). *SAC '02. ACM*, New York, NY, 147-150, 2002.

[36] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, “KEGG for linking genomes to life and the environment”, *Nucleic Acids Res,* vol.36, pp. D480-D484, 2008.

[37] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama., M. Araki, and M. Hirakawa, “From genomics to chemical genomics: new developments in KEGG”, *Nucleic Acids Res*, vol. 34, pp. D354-357, 2006.

[38] M. Kanehisa, and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes”. *Nucleic Acids Res,* Vol. 28, pp. 27-30, 2000.

[39] J. R. Quinlan, *C4.5: Programs for Machine Learning*., Morgan Kaufmann Publishers, 1993.

[40] Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra, S.A. Sansone, "ArrayExpress--a public repository for microarray gene expression data at the EBI.", *Nucleic Acids Res*. vol. 1;31(1) pp. 68-71, Jan 2003.

7. APPENDIX

7.1. Marker Sets

0-150mg/kg

Set 1: M63122_at, S82489_s_at, AF016179_at, rc_AI639421_at, rc_AA964530_at, rc_AI059291_at, J05231_at

Set 2: AFFX-LysX-5_at, D28110_g_at, D89730_at, U57050_at, rc_AI639004_at, rc_AI639509_at, rc_AA874803_g_at

0-450mg/kg

Set 3: AF093139_s_at, D00913_g_at, rc_AA892154_g_at, rc_AA892325_at

Set 4: D00913_g_at, Z18877_s_at, rc_AI235585_s_at, rc_AA874969_at

150-450mg/kg

Set 5: rc_AI172097_g_at, U64030_at, AF030089UTR#1_at, U95368_at, rc_AI639113_at, D78303_at

Set 6: S82489_s_at, M15358cds_at, M15527_at, rc_AA875004_at, rc_H33001_at, rc_AA892333_at

0-150-450mg/kg

Set 7: K00996mRNA_s_at, Z19552cds_at, D00913_g_at, Z18877_s_at, rc_AI638945_at, rc_AI235585_s_at, AF065438_at, rc_AA866459_at, rc_AA891209_at, rc_AA800850_at

Set 8: E00778cds_s_at, D63761_at, AF083418_at, D00913_g_at, U94322_s_at, M23995_at, rc_AI639113_at, U65656_at, AF065438_at, U96490_at, rc_AA859719_at, rc_AA866459_at, rc_AA799762_g_at

Selected by Decision Table method on 0-450mg/kg

L14782_s_at, L03294_g_at, L00191cds#1_s_at, M61142_at