Biomarkers, March–June 2005; 10(2–3): 153–172

Taylor & Francis
Taylor & Francis Group

Data-driven analysis approach for biomarker discovery using molecular-profiling technologies

T. WEI[1], B. LIAO[1], B. L. ACKERMANN[2], R. A. JOLLY[3], J. A. ECKSTEIN[2], N. H. KULKARNI[4], L. M. HELVERING[4], K. M. GOLDSTEIN[1], J. SHOU[1], S. T. ESTREM[1], T. P. RYAN[3], J.-M. COLET[3], C. E. THOMAS[3], J. L. STEVENS[3], & J. E. ONYIA[1]

[1]*Integrative Biology, Lilly Research Laboratories, Greenfield, IN, USA,* [2]*ADME, Research Laboratories, Greenfield, IN, USA,* [3]*TOX, Lilly Research Laboratories, Greenfield, IN, USA,* [4]*Bone and Inflammation, Lilly Research Laboratories, Indianapolis, IN, USA*

Abstract

High-throughput molecular-profiling technologies provide rapid, efficient and systematic approaches to search for biomarkers. Supervised learning algorithms are naturally suited to analyse a large amount of data generated using these technologies in biomarker discovery efforts. The study demonstrates with two examples a data-driven analysis approach to analysis of large complicated datasets collected in high-throughput technologies in the context of biomarker discovery. The approach consists of two analytic steps: an initial unsupervised analysis to obtain accurate knowledge about sample clustering, followed by a second supervised analysis to identify a small set of putative biomarkers for further experimental characterization. By comparing the most widely applied clustering algorithms using a leukaemia DNA microarray dataset, it was established that principal component analysis-assisted projections of samples from a high-dimensional molecular feature space into a few low dimensional subspaces provides a more effective and accurate way to explore visually and identify data structures that confirm intended experimental effects based on expected group membership. A supervised analysis method, shrunken centroid algorithm, was chosen to take knowledge of sample clustering gained or confirmed by the first step of the analysis to identify a small set of molecules as candidate biomarkers for further experimentation. The approach was applied to two molecular-profiling studies. In the first study, PCA-assisted analysis of DNA microarray data revealed that discrete data structures exist in rat liver gene expression and correlated with blood clinical chemistry and liver pathological damage in response to a chemical toxicant diethylhexylphthalate, a peroxisome-proliferator-activator receptor agonist. Sixteen genes were then identified by shrunken centroid algorithm as the best candidate biomarkers for liver damage. Functional annotations of these genes revealed roles in acute phase response, lipid and fatty acid metabolism and they are functionally relevant to the observed toxicities. In the second study, 26 urine ions identified from a GC/MS spectrum, two of which were glucose fragment ions included as positive controls, showed robust changes with the development of diabetes in Zucker diabetic fatty rats. Further experiments are needed to define their chemical identities and establish functional relevancy to disease development.

Keywords: *Biomarker, data-driven, supervised, unsupervised, molecular profiling*

(Received 8 November 2004; accepted 14 March 2005)

Correspondence to: Jude E. Onyia, Integrative Biology, Lilly Research Laboratories, Indianapolis, Greenfield, IN 46140, USA. Tel: 1-317-277-1267. Fax: 1-317-277-2934. E-mail: jeo@lilly.com

ISSN 1354-750X print/ISSN 1366-5804 online © 2005 Taylor & Francis Group Ltd
DOI: 10.1080/13547500500107430

Introduction

The discovery and application of biomarkers is a promising strategy to reduce the cost, time and attrition of development of new therapies for diseases. In preclinical studies, biomarkers help to validate targets, select appropriate animal models and differentiate lead compounds. In clinical studies they help to confirm efficacy, safety and mechanism of action, guide protocol design, and aid in patient selection. However, identification and development of biomarkers is a substantial undertaking (Frank & Hargreaves, 2003). High-throughput technologies such as DNA microarray developed for functional genomics, technologies developed for proteomics and metabonomics are used to interrogate thousands of molecules (genes, proteins or metabolites) in biological samples simultaneously, thus provide rapid, efficient and systematic approaches to search for appropriate biomarkers (Timbrell 1998, Robertson et al. 2001, Pang et al. 2002, He & Chiu 2003, Ilyin et al. 2004).

A major challenge in biomarker discovery is that it is necessary to analyse millions of data points collected from a study using high-throughput technologies to identify suitable candidate biomarkers. Many computational algorithms, either supervised or unsupervised (Butte 2002), can be applied to data analysis with three distinctive objectives: class comparison, class prediction and class discovery (Simon et al. 2003). Supervised methods are analytic procedures that require sample grouping information (Duda et al. 2001) such as diseased versus healthy or toxic versus non-toxic. The first application of these methods is class comparison, i.e. to identify differentially expressed molecules in two or more predefined experimental conditions or two or more diseased stages or types, e.g. statistical t- or F-test. The second application is class prediction that usually includes two interdependent steps: first, to identify informative molecules (feature selection) and, second, to develop a multivariate function (the classifier or predictor) that accurately predicts the class membership of a new sample based on its measured values of selected molecules, a process called model training or machine learning. In class discovery, unsupervised methods attempt to identify internal data structures or relationships in a data set without requiring a prior knowledge of sample grouping. A number of techniques are available including hierarchical clustering, self-organizing maps, network determination and cluster correlation (Ross et al. 2000) to integrate two or more data sets of distinct types providing complementary information.

Many experiments designed for biomarker discovery involve using high-throughput technologies to profile samples from two or more predetermined groups or treatments such as healthy versus diseased tissues. Supervised algorithms use sample grouping information to guide selection of informative molecules as well as construction of a predictive classifier. Thus, supervised methods are particularly suitable for the purpose of biomarker discovery using high-throughput technologies. A variety of supervised algorithms have been successfully applied including neighbourhood analysis to classify human acute leukaemias and its subtypes (Golub et al. 1999), K-nearest neighbours combined with a genetic algorithm to classify toxicants (Hamadeh et al. 2002), support vector machine (SVM) to classify two or more cancer classes (Furey et al. 2000, Valentini 2002, Lee & Lee 2003) and tree-based classification of breast cancer (Zhang & Yu 2002) to name a few.

While supervised methods were shown to be effective to construct a molecular classifier from a high-throughput study, its successful application depends on accurate sample grouping knowledge or some sort of cost functions (Duda et al. 2001). Thus, it becomes of paramount importance to confirm experimental effects based on expected group membership. In addition, the molecular classifiers developed usually involve many dozens of genes. Given the current state-of-the-art of multiplex technologies, such as quantitative real-time PCR (QRT-PCR), it is difficult to apply these classifiers directly to screen thousands of compounds required in early drug development. Ideal molecular biomarkers for pre-clinical development would be a few genes or proteins whose expression is a direct measurement of biological actions of compounds interacting with intended target on which efficacy depends or targets that cause toxicity in an *in vivo* or *in vitro* model. With these issues and goals in mind, a data-driven analysis approach was devised that combines both unsupervised and supervised analysis techniques to facilitate the identification of a small number of candidate biomarkers using high-throughput technologies for further experimental characterization. In the present report we began by comparing several popular clustering algorithms and principal component analysis (PCA)-assisted visual clustering method for their effectiveness of discovery of data structures that confirm intended experimental effects using a well-known leukaemia DNA microarray data set (Golub et al. 1999). Two real examples are then presented to illustrate this data-driven analysis approach to identify candidate biomarkers.

Materials and methods

Analysis overview

Our data-driven approach consists of two analysis steps. In the first, an unsupervised analysis, in particular PCA, was employed to project samples into a series of two- or three-dimensional projections in the first *m* selected principal components (PCs), which allowed us to explore data structures visually to identify outlier samples and confirm intended experimental effects based on expected group membership. In the second step, one or more filtering criteria were applied to identify a small number of molecular features such as candidate biomarkers for further experimental characterization. In particular, a supervised learning algorithm, shrunken centroid analysis (Tibshirani et al. 2002), was performed to identify top-ranked discriminant molecular features. Other biological information, wherever available, such as functional relevancy as well as subcellular location was employed as the secondary filters.

Comparison of unsupervised data analysis algorithms

Popular clustering algorithms were compared for their effectiveness of disclosing data structures that confirm intended experimental effects using a well-known leukaemia DNA microarray data published by Golub et al. (1999). Clustering algorithms compared include one agglomerative hierarchical clustering algorithm (HCA) (Eisen et al. 1998), two divisive clustering algorithms, *K*-means (Soukas et al. 2000) and partitioning around medoids (PAM) (Bozinov & Rahnenfuhrer 2002), a recently described bagging clustering algorithm (Dudoit & Fridlyand 2003) and our visual clustering based on PCA-assisted sample projections in low dimensional subspaces. The leukaemia DNA microarray dataset was generated using a Affymetrix DNA chip HuGeneFL from 72 leukaemia cancer samples, 25 acute myeloid leukaemia (AML), nine acute lymphoblastic leukaemia T-cell (ALT), and 38 acute lymphoblastic leukaemia B-cell (ALB). Based on the known sample classification, all genes on the chips were ranked according to their discriminant power using the shrunken centroid algorithm (see below for details). The top ranked 100, 1000, 2000 and 3000 genes were selected for sample clustering in the comparison. Clustering results were compared with the known sample classifications to obtain clustering error rates, which is defined as:

$$\text{Error rate} = \text{number of clustering errors/sample size}$$

HCA, K*-mean, PAM and bagging PAM*

All clustering algorithms were run in R statistical computing environment version 1.9.0 (Maindonald & Braun 2003). HCA was performed using the hclust function in Euclidean distance and the complete linkage method. Two divisive clustering algorithms, K-means and PAM, were carried out with $K=3$ representing the three known subtypes of leukaemia. Recent efforts were made to improve clustering accuracy by using a resampling procedure called bagging (Breiman 1996). A resampling clustering program in R was written based on the algorithm described by Dudoit & Fridlyand (2003). Briefly, in the bagging clustering a selected clustering algorithm (PAM in this study) was repeatedly applied to each bootstrap sample to form a new dissimilarity matrix by recording for each pair of samples the proportion of time they were clustered together in the bootstrap clusters. The new dissimilarity matrix is then used as an input to PAM that assigns each sample into one of the K clusters.

PCA-assisted visual clustering

PCA (Jolliffe 2002) takes advantage of linear correlations among a large number of molecular features measured from each sample and finds a few orthogonal linear combinations of them that capture the majority of variations or information. Thus, PCA reduces high dimensionalities of original datasets so as to project the samples in a high-dimensional molecular feature space to lower dimensional subspaces such that one can explore the data structures visually to identify outlier samples and to confirm intended experimental effects based on expected group membership. To avoid genes with large variances dominating the variance–covariance structure, expressions for each gene were converted to z-scores with zero mean and unit variance. The first m PCs were selected such that:

$$t_m = 100/p \sum_{k=1}^{m} l_k > 60\%,$$

where p is the number of genes and l_k is the variance of the kth PC. The 60% variances, smaller than the recommended 70–90% (Jolliffe 2002), was due to a large number of genes p in the dataset.

PCA-assisted visual clustering of samples proceeded heuristically as described below. For the first m PCs selected, a series of two- or three-dimensional projections were generated. Each projection was examined visually for any discrete clusters of samples relevant to experimental conditions. Cluster boundaries in a linear form were positioned visually such that they minimize the grouping errors when compared with expected group membership under the experimental design.

Shrunken centroid algorithm

The detailed algorithm was described by Tibshirani et al. (2002) and briefly accounted here. Let x_{ij} be a measure for features $i = 1, 2, \ldots, p$ and samples $j = 1, 2, \ldots, N$. There are K classes in N samples identified from a previous analysis step with our PCA-assisted visual clustering procedure. Let C_k be indices of the N_k samples in class k ($k = 1, 2, \ldots, K$). The centroid for the kth class ($k = 1, 2, \ldots, K$) is defined by p components as $\bar{x}_{ik} = \Sigma_{j \subset Ck} x_{ij}/N_k$, i.e. the class average of the pth feature. The overall data centroid is defined by p components as an average over all samples, $\bar{x}_i = \Sigma_{j \subset N} x_{ij}/N$. The nearest-centroid classification algorithm will assign the jth sample ($x_{1j}, x_{2j}, \ldots, x_{pj}$) to be class k if its squared distance from the centroid of the kth class is the smallest one. In other words, it has a shorter distance to the kth centroid than to any other $k - 1$ classes. The shrunken centroid algorithm takes the idea of the nearest-centroid algorithm by integrating a feature selection procedure into its model building process. Each feature i is evaluated by a metric defined as:

$$d_{ik} = (\bar{x}_{ik} - \bar{x}_i)/(\text{pooled within class SD}).$$

Thus, features with large changes and small pooled within-class variations will be weighted higher than those with small changes and large pooled within-class variations. Importantly this concept is very consistent with the characteristics of a desired biomarker, i.e. large yet stable changes in response to a treatment or pathophysiological alterations. Using d_{ik} to select molecular features as putative biomarkers is a matter of establishing a threshold Δ. If $d'_{ik} = |d_{ik}| - \Delta <= 0$, the ith feature is considered to be non-informative and excluded from the model. To establish Δ, a ten-fold cross-validation procedure is executed and the appropriate Δ is defined as the one that produces a model consisting of features with $d_{ik}{}' > 0$ that results in the minimum error rate in class prediction. An implementation of the algorithm was made available to the public by the authors in an R package pamr. Minor modifications were made to generate scatter plots for selected molecules.

Diethylhexylphthalate (DEHP) study for liver toxic biomarkers

Male Sprague–Dawley (SD) rats ($n = 3$–5) were given a single dose of DEHP at 2 g kg^{-1} for low dose and 20 g kg^{-1} for high dose with a water vehicle group and sacrificed at 4, 24, 48 and 168 h post dosing, as described (Lindon et al. 2003). The dose was set based on the dose needed to obtain acute treatment toxicity as determined by a range finding study (data not shown). To assess liver damage, harvested livers were weighed and processed through to paraffin wax, sectioned, stained with haematoxylin and eosin, and examined microscopically. Clinical pathology assessment was performed on plasma from blood samples collected at necropsy. Approximately 0.80 ml blood per sample were placed into glass micro-centrifuge tubes following retro-orbital bleed. Fresh plasma samples were analysed for clinical pathology parameters. The following parameters were measured: alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase, gamma-lutamyl transferase, blood urea nitrogen, protein, albumin, inorganic phosphorus, bilirubin, creatinine, glucose, sodium, potassium and calcium.

To determine induced gene expression changes using DNA chip technology, total RNA from liver was isolated with RNA STAT-60 (Tel-Test, Friends Wood, TX, USA) according to the manufacturer's protocol. A total of 10 μg total RNA were labelled and hybridized to the RG-U34A chip according to the Affymetrix (2002) protocol. Signal intensities were generated from Mircoarray Suite version 5.0 (MAS5) using the default setting, and a global scaling set to 1500.

GC/MS metabonomic study of disease progression in the Zucker diabetic fatty rat

Six Zucker diabetic fatty (ZDF) and six lean Zucker (LZ) rats were housed in metabolism cages and fed a diet of Purina 5008 rat chow. The rats were given water *ad libitum*. Urine was collected over 24 h at ages 5, 8 and 12 weeks. In addition to mass spectral analysis, various clinical chemistry assays were conducted on the samples including the determination of serum creatinine. Urine aliquots (100 μl) taken for GC/MS analysis were treated with urease followed by ethanol precipitation (100 μl) to remove the enzyme. The supernatant was dried and incubated with β-glucuronidase/sulfatase (Sigma Chemical Co., St Louis, MO, USA) to remove conjugates. The incubation was terminated by the addition of 100 μl ethanol containing the internal standard d_3-stearic acid. The sample was dried and reacted with methoxyl amine in pyridine (μg/100 μl) before silylation using N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) containing 1% (v/v) trimethylchlorosilane (TMCS). The samples were diluted in hexane (1:5, v/v) and 1 μl was injected for GC/MS analysis. Samples were analysed using a Voyager single quadrupole (ThermoElectron, San Jose, CA, USA) in positive-ion mode using electron-impact ionization (70 eV). Mass spectra were acquired under unit mass resolution by scanning the quadrupole from *m/z* 50 to 650 in 1 s. GC separation occurred using a DB-5 column (0.25 mm × 30 m) obtained from J&W Scientific (Folsom, CA, USA) using a temperature gradient from 40 to 300°C in 30 min. Each urine sample was analysed in duplicate resulting in 72 data files.

Mass spectral peak detection and initial data processing occurred using the resident software provided by the vendor (Xcalibur™ v. 1.2). Using the software provided in Xcalibur, each mass spectrum was converted into netCDF format (ANDII standard) to facilitate further processing by PCA, as described below.

Results

Improved accuracy with PCA-assisted visual clustering procedure

Several clustering algorithms including HCA, *K*-mean, PAM and bagging clustering were tested against the leukaemia DNA microarray dataset with known leukaemia types. Clustering error rates obtained are shown in Table I. Samples were also clustered visually after projection into two-dimensional spaces as defined by the first three principal components from PCA analysis. As shown in Table I, HCA was the poorest performer at all levels of gene selection. Bagging clustering also yielded high error rates, which conflicts with results reported by Dudoit & Fridlyand (2003). Although *K*-means was the best performer when the top 100 genes were used in the analysis, its accuracies degraded quickly as the number of less discriminant genes in the analysis increased. Visual clustering based on PCA projections (Figure 1) outperformed all distance metrics-based algorithms. For example, using 2000 genes, three leukaemia subtypes could be easily identified from the two projections: one

Table I. Error rates[1] of clustering leukaemia samples using different unsupervised algorithms.

Number of probes[2]	HCA[3]	*K*-mean[4]	PAM[5]	Bagging[6] PAM	PCA[7]
100	0.11	0.056	0.153	0.194	0.056
1000	0.125	0.069	0.153	0.208	0.056
2000	0.167	0.264	0.319	0.361	0.069
3000	0.29	>0.5	>0.5	>0.5	0.069

[1] Error rate = number of clustering errors/sample size.

[2] Topmost discriminant Affymetrix probes identified by shrunken centroid method.

[3] Hierarchical clustering analysis was done in R using an hclust function with Eucledian distance and complete linkage.

[4] K-mean was done in R using kmeans function with $K=3$.

[5] PAM was done in R using pam function with $K=3$.

[6] Bagging PAM was written based on the algorithm described in Dudoit & Fridlyand (2003).

[7] PCA was done in R using the prcomp function.

(Figure 1C) defined by PC1 versus PC2, which separates AML from ALT and ALB, and the other (Figure 1D) defined by PC1 versus PC3, which separates ALT from ALM and ALB. Even with 3000 genes, two main leukaemia subtypes could still be identified, while all other algorithms failed completely. Thus, it was established that PCA-assisted visual clustering is the most effective method to disclose data structures that can be used to confirm intended experimental effects and thus was chosen to use in the first step of our data-driven analysis approach.

Identification of genes as candidate liver toxic biomarkers for liver injury in a DNA microarray experiment

DNA microarray data were generated from livers of rats at four different time points after a single exposure to vehicle, low and high doses of DEHP. Data were first analysed by PCA with all 8799 probe sets on RG-U34A chip (Figure 2A). Three visible clusters of samples formed based on global changes in gene expression. Cluster 1 consisted of high dosed samples at 24 h; cluster 2 consisted of high dosed samples at 48 h; and cluster 3 of all remaining 33 samples.

Clinical data showed that the high dose of DEHP caused significant changes of six and all nine measured parameters at 24 and 48 h respectively, while the low dose did not cause any of them to change significantly, and only protein and albumin were significantly changed by high dose at 168 h. Histopathological data indicated liver damages (slightly enlarged and more demarcated hepatocytes with eosinophilic granular cytoplasm while the glycogen content decreased as well as minimum-to-slight hepatocellular proliferation) induced by DEHP occurred when observed at 48 h. High-dose samples at 4 h did not show any clinical pathology and its global expression profiles were similar to those of vehicle and low-dose samples and joined the same cluster. Similar results were also obtained for high-dose samples at 168 h. Thus, the PCA analysis revealed there is a time-dependent trajectory in which the largest separation between vehicle and toxicant-treated animals occurs at which liver injury is clearly detectable; the same time points at which the maximal numbers of gene expression changes occurred.

Since most of the cluster 3 in Figure 2 was comprised of samples from rats treated with vehicle or low dose of the toxicant and showed no visible liver damage, this cluster of samples was called the 'non-toxic' group. Since the main objective of the

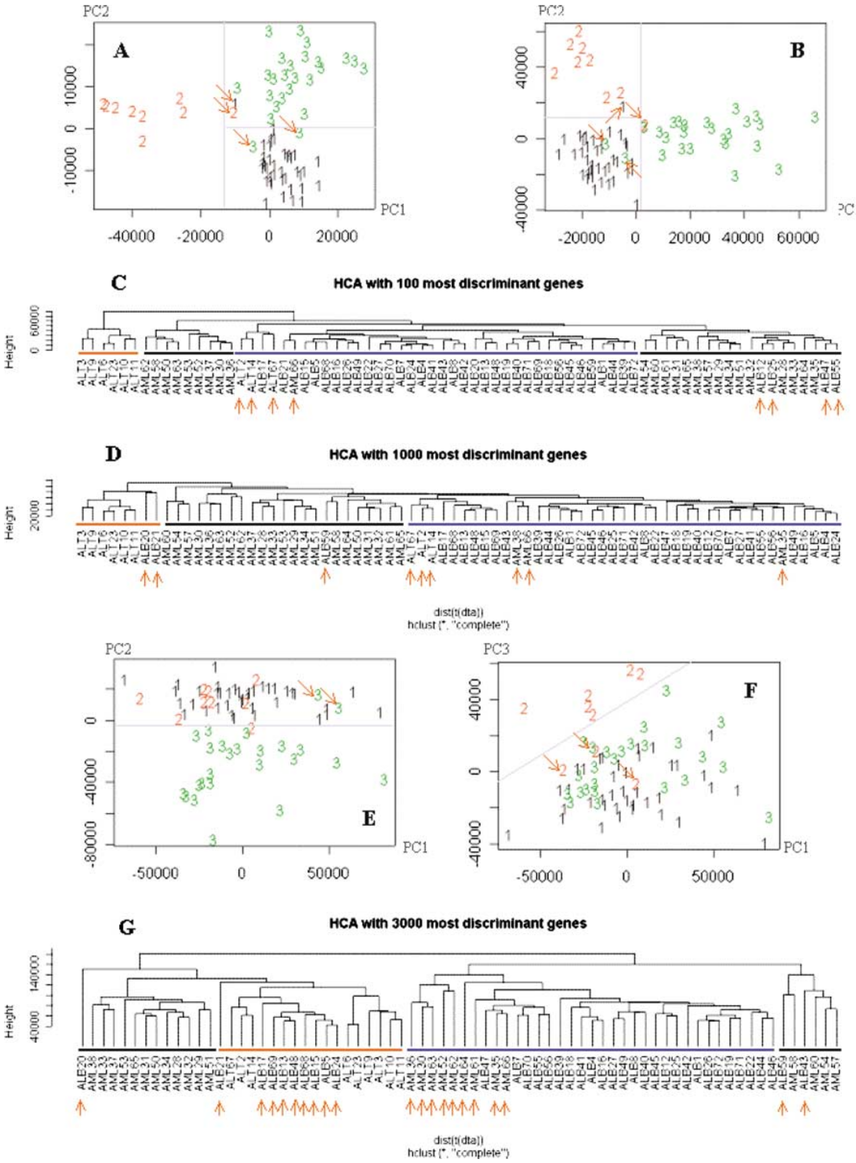

Figure 1. Comparison of HCA and PCA-assisted visual clustering. 1, ALB; 2, ALT; 3, AML. Light grey lines show cluster boundaries that produce the minimum clustering errors. A red horizontal bar represents for ALT-type leukaemia, a black horizontal bar for AML-type leukaemia and blue horizontal bar for ABL-type leukaemia. Red arrows indicate incorrectly clustered samples. (A) Two-dimensional projection of 72 samples in the first two PCs resulting from PCA of the top 100 discriminant genes, which may be compared with (C) generated by HCA over the same 100 genes. (B) Two-dimensional projection of 72 samples in the first two PCs resulting from PCA of the top 1000 discriminant genes, which should be compared with (D) generated by HCA over the same 1000 genes. (E, F) Two two-dimensional projections of 72 samples in the first three PCs together defining three distinct subtypes of leukaemia resulting from PCA of the top 3000 genes. Panel (E) separates 3 (AML) from 1 (ALB) and 2 (ALT), while (F) separates 2 from 1 and 3. Both (E) and (F) should be compared with (G) generated by HCA over the same 3000 genes.

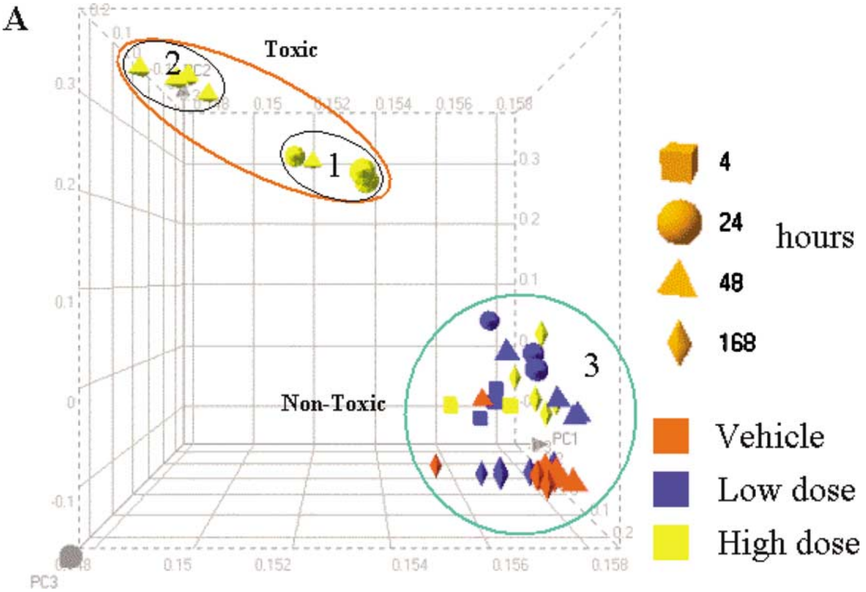

Figure 2A. Three-dimensional projection of global gene expression of liver samples collected at different time points from rats treated with vehicle, low and high dose DEPH. PCA was performed over all 8799 probe sets on a RG-U34A chip. The first three PCs with an accumulated 60% of the total variances were selected to generate the projection.

analysis was to identify candidate hepatotoxicity biomarkers and the fact that clusters 1 and 2 of high-dose samples are close (Figure 2A) with the observable liver damage, clusters 1 and 2 were merged into a single 'toxic' cluster. Such a grouping is necessary

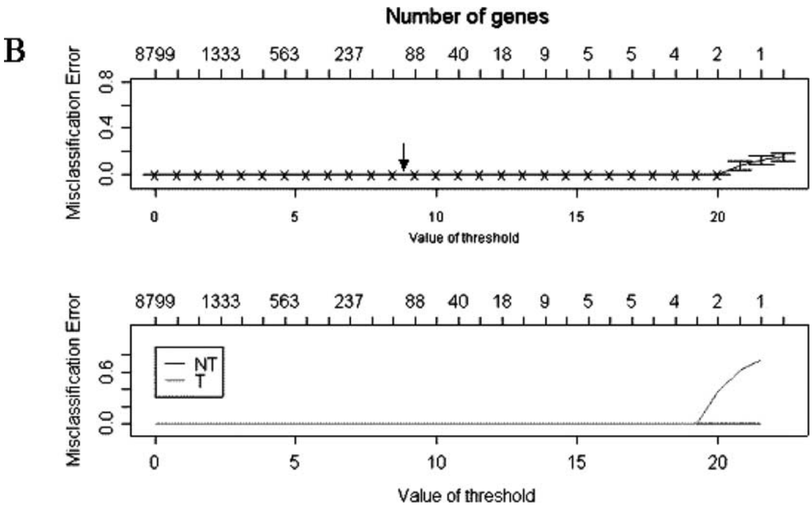

Figure 2B. Misclassification errors at different levels of shrinkage generated from a ten-fold cross-validation. The first graph is total misclassification errors; the second graph is classification errors for each clusters. An arrow indicates the selected shrinkage at which the minimum misclassification errors were achieved, and the top 100 genes were selected to examine their mechanistic relevance based on available functional annotation. NT, non-toxic cluster; T, toxic cluster.

for the shrunken centroid algorithm to identify a group of genes with similar expression response to the toxicant at both time points.

The second step in our data-driven analysis procedure was to apply the shrunken centroid algorithm to the dataset, with sample grouping information taken from the previous step, to identify a small set of putative liver toxicity biomarkers. Figure 2B shows error rates of the cross-validation at various threshold levels. From Figure 2B, $\Delta \leq 20$ resulted in zero error rate in the cross-validation. $\Delta = 8$ was chosen such that the top-ranked 100 of probe sets were obtained and examined for their functional relevancy. DEHP is a known agonist of the peroxisome proliferator-activated receptor (PPAR). Sixteen genes of the 100 probes listed in Table II show that they function in acute-phase response, lipid or fatty acid metabolisms. Figure 3 shows expression scatter plots of 16 selected genes. As shown, each gene demonstrates robust (large yet stable) expression changes between the toxic and non-toxic clusters of samples, thus constituting a group of good candidate liver toxicity biomarkers. For example, acyl-CoA hydrolase has been reported to be inducible by DEHP (Yamada et al. 1998) and was identified by the algorithm twice, showing a repeatable expression profile (compare panels 2 and 3 in the first row).

Identification of urinary markers correlated to disease progression in Zucker diabetic fatty rats using GC/MS

A metabonomic study was conducted to investigate disease progression in the ZDF rat, a widely used pharmacology model used in the study of type II diabetes. This model has been profiled by Etgen & Oldham (2000), who monitored disease progression using common markers of diabetes. ZDF rats appear phenotypically normal at 5 weeks of age, but experience a significant decline in insulin production by 8 weeks and are overtly diabetic by 12 weeks. In the present study, urine collected for ZDF rats and LZ control animals at these three time points was analysed by GC/MS to identify metabolic differences and to look for markers of disease progression. During GC/MS data acquisition, a full-mass spectrum is acquired each 1 s during the chromatographic profile. This generates a relatively large data set consisting of three dimensions (retention time, *m/z*, peak intensity). This data set was transformed for PCA analysis as described below.

The starting point for transformation was a tabular representation of each mass spectrum consisting of increasing nominal *m/z* values and their corresponding peak intensities. Since nominal values were used, the bin width was 1 Da. An in-house script was written to transform the MS data into a format amenable for PCA analysis. Thus, each mass spectrum was divided into 600 bins to accommodate the scan range used (*m/z* 50–650) and the ‘features’ for PCA corresponded to pairs of *m/z* and intensity. Expressing GC/MS data files in this two-dimensional format meant that the dimension of retention time could not be encoded. Hence, each data file was represented as a composite mass spectrum created by averaging the data for the entire chromatogram into a single spectrum.

Before conducting PCA analysis, two additional data transformations were performed. The first step was to remove a series of 16 *m/z* values corresponding to prominent ions in the reference mass spectrum of glucose. This action was taken to remove the dominant effect of glucose and reduced the number of bins used for PCA to 584. The second step involved normalization of each MS data file according to the

Table II. Functional annotation of 16 genes selected from the 100 top ranked probes.

Rank	Genbank Accession number	Gene symbol and description	Biological process
50	D00752	Spin2a: serine protease inhibitor, a putative contrapsin-like serine protease inhibitor that can function in the inflammatory response, member of the class II acute phase protein family	acute-phase response
11	Y09332	Bach: acyl-CoA hydrolase, peroxisome proliferator-induced acyl-CoA thioesterase (liver acyl-CoA hydrolase) that hydrolyses palmitoyl-CoA	acyl-CoA metabolism
39	D88890	Bach: acyl-CoA hydrolase, peroxisome proliferator-induced acyl-CoA thioesterase (liver acyl-CoA hydrolase) that hydrolyses palmitoyl-CoA	acyl-CoA metabolism
8	D17349	Cyp2b15: cytochrome P450 2b19, member of the cytochrome P450 monooxygenase family, converts arachidonic acid to 11,12-epoxyeicosatrienoic acid, involved in keratinocyte differentiation and intracellular signalling within differentiated keratinocytes	epoxygenase P450 pathway
47	AI232087	Hao3: hydroxyacid oxidase 3 (long-chain L-alpha-hydroxy acid oxidase), a member of the FMN-dependent alpha-hydroxy acid-oxidizing family, may function in lipid metabolism	fatty acid alpha-oxidation
77	AF080468	G6pt1: glucose 6-phosphate translocase, component of glucose 6-phosphatase enzyme complex, inhibited by chlorogenic acid and its synthetic derivatives, upregulated by insulin deficiency and hyperglycaemia in streptozotocin-induced diabetes	fatty acid metabolism
94	AA800120	Slc25a20: solute carrier family 25 member 20 (carnitine-acylcarnitine translocase), mitochondrial carnitine transporter that may function in fatty acid beta-oxidation; mutation of the human SLC25A20 gene results in hypoketotic hypoglycaemia and cardiac abn	fatty acid oxidation
17	M16235	Lipc: hepatic triglyceride lipase, member of the triacylglycerol lipase family that metabolizes high-density lipoproteins; mutation of the human LIPC gene may be associated with familial combined hyperlipidaemia	fatty acid transport
2	M11251	LOC292728: cytochrome P450 CYP2B21, a putative cytochrome p450 that may play a role in xenobiotic metabolism in the oesophagus	icosanoid metabolism
91	M94548	Cyp4f14: cytochrome P450 subfamily IVF, a leukotriene B4 omega-hydroxylase, which also catalyses omega hydroxylation of six *trans* leukotriene B4, lipoxin A4, prostaglandin A1 and several hydroxyeicosatetraenoic acids, but not lipoxin B4, laurate and arachi	leukotriene metabolism
70	U08976	Ech1: enoyl hydratase-like protein (peroxisomal), putative enzyme that may play a role in peroxisomal beta-oxidation	lipid metabolism
67	AB010429	Mte1: mitochondrial acyl-CoA thioesterase I (very long-chain acyl-CoA thioester hydrolase), plays important roles in lipid metabolism; subject to dietary regulation, strongly upregulated by peroxisome proliferators including clofibrate, and by fasting	lipid metabolism
23	Y09333	Mte1: mitochondrial acyl-CoA thioesterase I (very long-chain acyl-CoA thioester hydrolase), plays important roles in lipid metabolism; subject to dietary regulation, strongly upregulated by peroxisome proliferators including clofibrate, and by fasting	lipid metabolism

Table II (*Continued*)

Rank	Genbank Accession number	Gene symbol and description	Biological process
7	AB010428	Cte1: cytosolic acyl-CoA thioesterase I (long-chain acyl-CoA thioester hydrolase), mainly active toward fatty acyl CoAs with chain-lengths of C12–C16, may play a role in lipid metabolism; inducible by peroxisome proliferators, and subject to dietary regulation	lipid metabolism
20	D38065	Ugt1a1: UDP glycosyltransferase 1 family polypeptide A1, catalyses the transfer of glucuronic acid to a variety of substrates, can detoxify xenobiotics and drugs; mutations in human UGT1A1 cause Gilbert and Crigler–Najjar syndromes	response to toxin
89	AA859529	Dgat1: diacylglycerol *O*-acyltransferase 1, catalyses the conversion of diacylglycerol to triacylglycerol	triacylglycerol metabolism

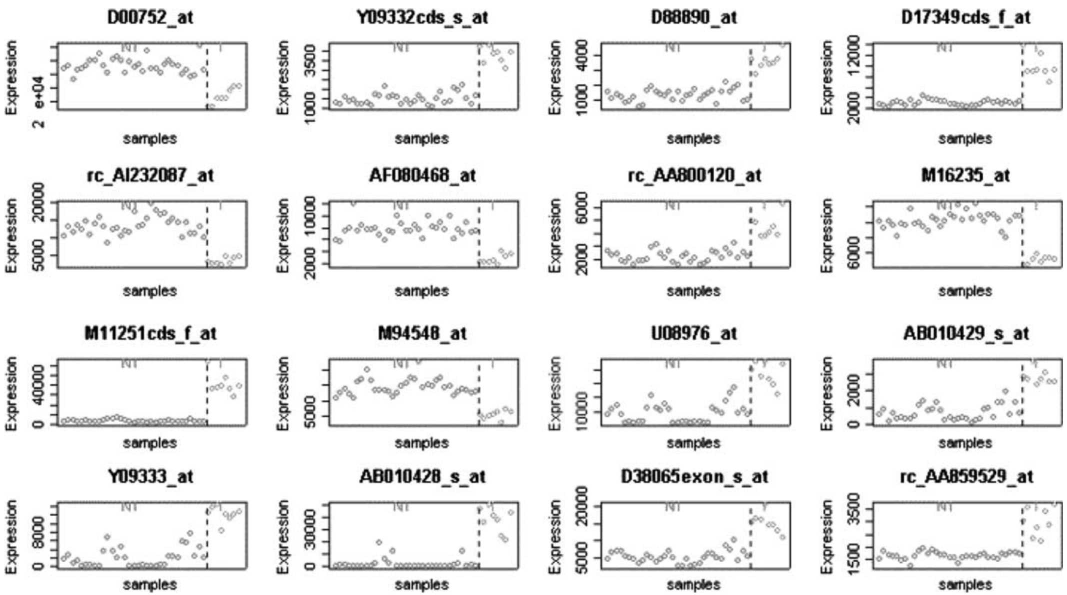

Figure 3. Sample scatter plots of gene expression of 16 candidate biomarkers identified by the data-driven analysis procedure. Affymetrix probe set names are shown on the top of each plot. NT, non-toxic cluster; T, toxic cluster.

serum creatinine level obtained for the corresponding animal and time point using a conventional clinical chemistry assay. Because serum creatinine is a well-established marker of renal blood flow, it is routinely used to correct for differences in urine output and hence the effect of dilution. This correction was necessary owing to the increased urine output that occurred as the ZDF rats became increasingly diabetic.

Results from PCA analysis appear in Figure 4A, which shows three clusters clearly separated from one another. Cluster 1 includes all lean Zucker rat samples and 5-week ZDF rat samples. Cluster 2 includes samples of ZDF rats at 8 weeks. Cluster 3 includes 12 samples from 12-week-old ZDF rats plus duplicates from rat #11 at 8 weeks. The cluster of rat #11 is consistent with accelerated disease progression of this animal, which had the highest serum glucose level of the six ZDF rats at 8 weeks (340 mg dl^{-1}). The corresponding mean and %CV for the six rats was 210.3 mg dl^{-1} and 38%. Carefully examining the plot shows that the direction of the largest variation, i.e. PC1, corresponds to disease progression (Figure 4D). Direction of the second largest component of variation (PC2) corresponds to normal development (Figure 4C). Interestingly, the two directions were roughly orthogonal and, thus, disease progression appeared independent of normal development.

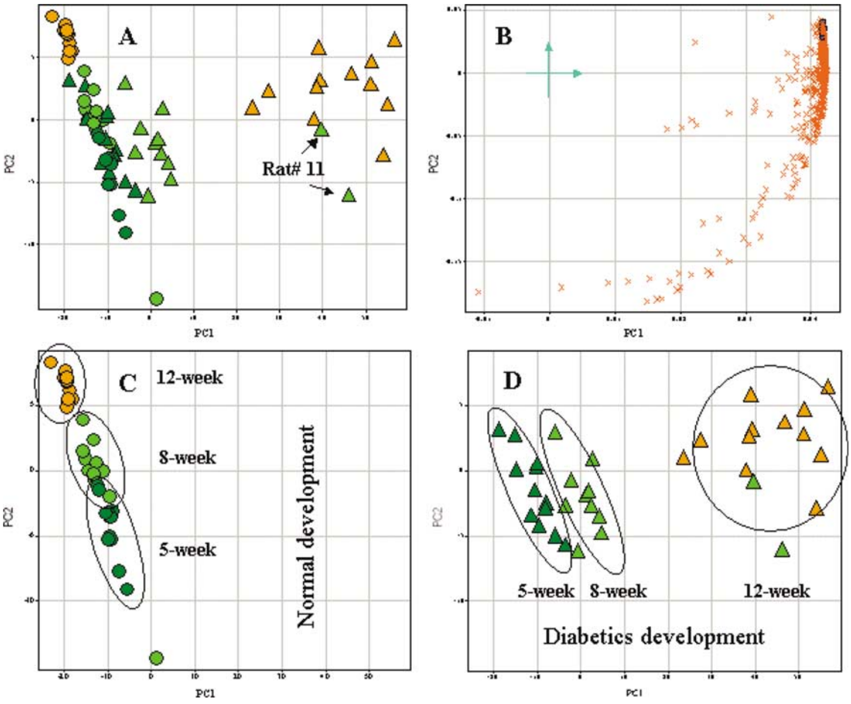

Figure 4A. Two-dimensional projection of 72 urine samples collected at different times (5, 8, 12 weeks) of lean and Zucker diabetic fatty rats. PCA was performed over 584 fragment ions. (A) Projection of 72 urine samples. Arrows indicate outlier samples. (B) Loading plots over PC1 and PC2. Each symbol stands for a fragment ion and its position on the plot represents its contribution to PC1 and PC2 respectively. Blue symbols are 26 fragment ions selected by a shrunken centroid algorithm as candidate biomarkers for diabetic disease progression. They are located consistently at the leftmost on PC1. (C, D) Reproduced from (A) for better visualization of the two major sources of variations revealed by PCA projections, one for disease progression and the other for normal development.

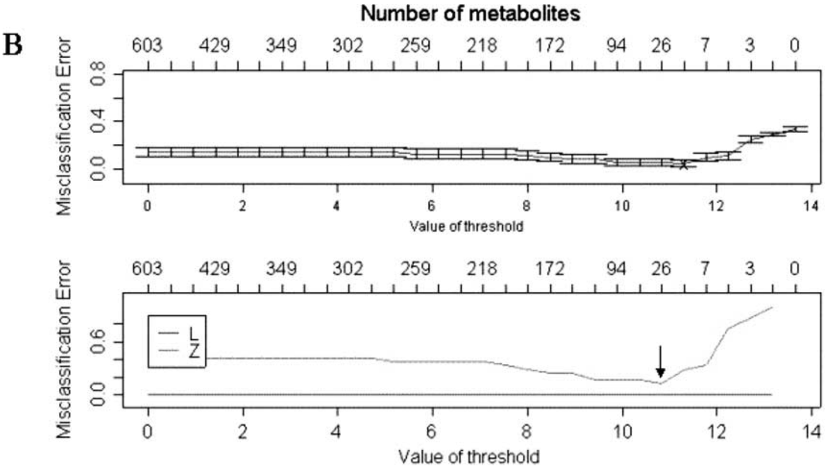

Figure 4B. Misclassification errors at different levels of shrinkage generated from a ten-fold cross-validation. The first graph is total misclassification errors; the second graph is classification errors for each group. An arrow indicates the threshold at which the minimum misclassification errors were achieved. L, lean cluster; Z, ZDF diabetics cluster.

Although cluster 2 (8-week ZDF rats) did not separated from cluster 1 (lean and 5-week ZDF rats) as well as it did from cluster 3, it did deviate from cluster 1 in the direction of disease progression (Figure 6 and 4D), suggesting there were correlated changes of molecular features relevant to disease progression. While distinct data structures existed within cluster 1 due to normal development (Figure 4C), it was chosen to group them as a single normal group because the objective was to identify biomarkers for diabetic progression not ones for normal development. Given the information revealed by PCA, two approaches could be taken to identify candidate biomarkers consistently regulated in the disease progression. In the first approach, a set of discriminant molecules between cluster 1 and 2 and another set between cluster 1 and 3 could be identified by separately applying the shrunken centroid algorithm. The common set of the two would constitute the best candidate biomarkers. In the second approach, clusters 2 and 3 were pooled to form the ZDF diabetic group, while all lean rats were grouped with 5-week ZDF rats to form the lean group. By merging clusters 2 and 3 into a single cluster, the shrunken centroid algorithm was forced to identify the molecular features with consistent changes from early time (8 weeks) to fully developed disease state (12 weeks). The direction of disease progression thus constituted the best candidate biomarkers in a single algorithm run.

Thus, the shrunken centroid algorithm was applied to the dataset with sample grouping information obtained from the previous analysis. Two prominent fragment ions from the mass spectrum of trimethylsilyl-derivatized glucose at *m/z* 291 and 305 were included in the analysis serving as positive controls. Figure 4E shows the sample classification error rates at various thresholds obtained from cross-validation. It was determined that at $\Delta = 10.8$, the model achieved the smallest classification error rate, i.e. four of 72 samples. At $\Delta = 10.8$, there were 26 urine metabolite peaks including the two glucose peaks whose $d'_{ik} > 0$. The top ranked six features were plotted across all samples (Figure 5). Urine metabolites identified show robust changes between the two types of rats lean versus ZDF diabetics. In particular, they also show a strong

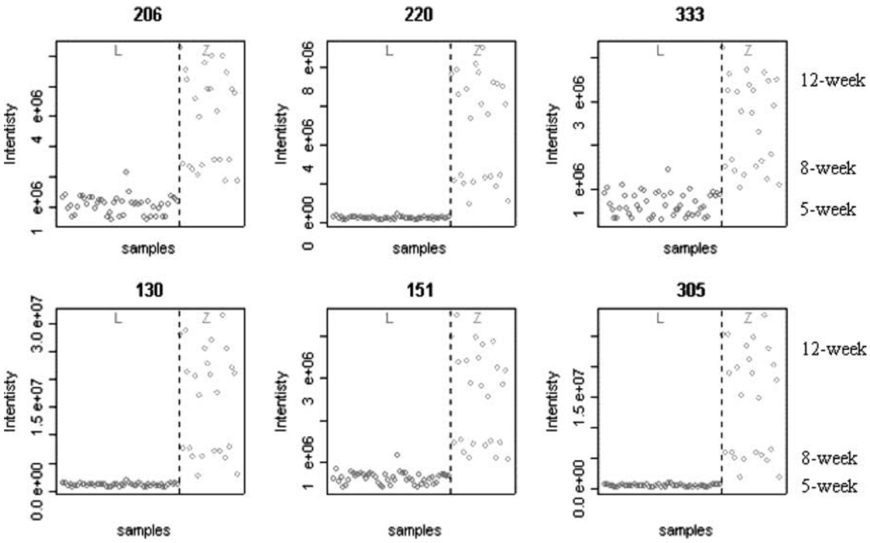

Figure 5. Sample scatter plots of six of 26 fragment ions of urine metabolites identified by the data-driven analysis procedure as putative biomarkers for diabetics development in Zucker diabetic fatty rats. Peak identification numbers are shown on the top of each plot. L, lean cluster; Z, Zucker diabetic fatty rat cluster. Peak 305 represents glucose ions and serves as a positive control.

correlation with progression of the disease. Thus, these features that represent *m/z* were derived from metabolites corresponding to the best candidate biomarkers. The authors are currently using this information as part of an ongoing investigation to derive chemical structures for these putative markers and validate their association with the disease progression.

Discussion

High-throughput technologies can interrogate thousands of molecules simultaneously and provide new avenues for identifying biomarkers that help select appropriate animal models and lead compounds in pre-clinical studies, and to confirm the efficacy and mechanism of action, minimize safety risk, guide protocol selection, and stratify patient population in clinical studies. Supervised analysis methods have been applied to aid in identifying predictive molecular profiles for disease diagnosis or prognosis (Golub et al. 1999, Bittner et al. 2000). These methods require predetermined sample grouping information to guide feature selection and model building. In practice, such information may not be readily available or experimental effects might not always be achieved as designed. The present paper adopted a data-driven analysis approach in which an unsupervised analysis was first applied to confirm the intended experimental effects or to discover relevant data structures, knowledge of which was then exploited in a supervised analysis to identify a small number of molecules as candidate biomarkers for further experimental characterization.

Many unsupervised algorithms, particularly clustering analysis, have been widely applied in molecular-profiling data analysis to discover new disease subtypes (Golub et al. 1999) or cluster molecules into functionally distinct groups (Eisen et al. 1998). Several popular clustering algorithms were benchmarked with the well-known Golub's dataset. HCA performed poorly with this data set. This is largely because HCA searches for local optimum based on a distance metric that may not be global optimum, examples of which can are shown in Figure 1B. Two samples of cluster 2 are much closer in distance to cluster 1, and 3 than to the cluster 2 thus would be clustered together with either cluster 1 or 3 by HCA. However, globally it best clusters with cluster 2, its own sample group. In the other two datasets where there were well separated clusters (Figures 2A and 4A), all clustering algorithms generated similar clustering results. Although divisive methods achieved better accuracies than HCA, it suffers two shortcomings. It needs a predetermined number of clusters, which is not readily available in many situations, and clustering results are not stable and depend on cluster centroids selected at the beginning of the algorithm. With a random selection scheme, as many programs implement, it would be difficult to obtain an optimum clustering result in a single algorithm run. The bagging procedure was initially proposed to stabilize 'unstable' classifiers such as neural networks and decision trees (Breiman 1996). It has been recently reported that the bagging procedure could significantly improve clustering accuracies of PAM (Dudoit & Fridlyand 2003). The present paper found otherwise when one of the two proposed bagging procedures was tested on the leukaemia dataset. The apparently inconsistent result may result from the way the dissimilarity matrix was implemented in the proposed algorithm. The dissimilarity of any two samples is inversely proportionate with the proportion of times they are clustered together in bootstrap samples by multiple PAM runs. It can be expected that samples close one another in the original sample space tend to cluster. Thus, the bagging PAM algorithm proposed by Dudoit & Fridlyand (2003) may stabilize clustering results of PAM, but might not be able to increase its accuracy. Figure 6 shows an example in which a sample from cluster 1 and a sample from cluster 2 tend to cluster together due to their proximity in Euclidean space. The two samples will join one of the two larger clusters, either 1 or 3, in a PAM run, which results in one error. Multiple PAM runs over bootstrap samples may stabilize clustering of the two samples with one of the two larger clusters 1 or 3.

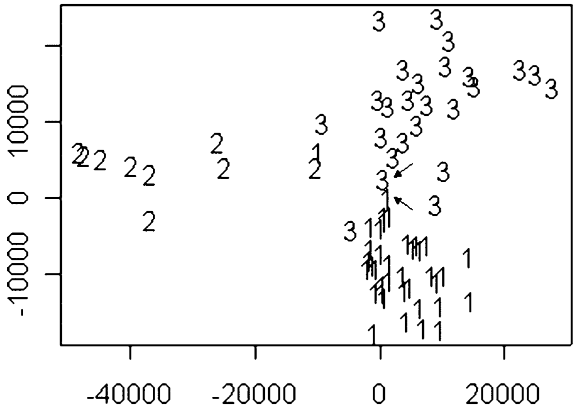

Figure 6. Two samples arrowed can be correctly clustered visually with the PCA plot. However, due to their proximity in distance, they will be clustered together for most times in multiple PAM runs, resulting in one error. The bagging clustering may stabilize clustering of the two samples with one of the two nearby larger clusters 1 or 3, but it might not break the two due to their proximity in Euclidean space.

Our understanding of complicated biological processes is greatly hampered by the inability to visualize readily data structures inherent in a high-dimensional dataset generated by measuring these biological processes. PCA is an unsupervised analysis technique that can project a complicated dataset in a high-dimensional space into its subspaces defined by a few prominent principal components while retaining the majority of variation. Thus, PCA-assisted projections of samples into subspaces allow one to explore visually and identify data structures that can be used to identify outlier samples or confirm intended experimental effects. The rationale for the validity of this method is because a biological system tends to respond to physiological changes caused by disease development, or to external treatments, in a coordinated fashion such as co-regulated gene expression observed in transcriptional profiling experiments (Whitfield et al. 2000). As demonstrated in the three examples, it was shown that PCA-assisted projections of samples into subspaces could effectively disclose inherent data structures that confirm intended experimental effects when compared with expected group membership, forming an accurate piece of knowledge required for a supervised learning algorithm to identify potential molecular biomarkers.

Knowledge of sample clusters obtained in the previous analysis step was employed to identify putative biomarkers by applying a supervised analysis. The supervised analysis could be as simple as statistical testing or any other machine learning algorithms combined with a feature selection procedure (Golub et al. 1999, Bittner et al. 2000, Furey et al. 2000). Machine learning algorithms are preferred over statistical testing because the multivariate classifier constructed from selected features or candidate biomarkers are evaluated either by a cross-validation procedure or by independent testing with samples not used in the modelling process. Thus, the biomarkers identified by these methods are more relevant. Several supervised analyses that have been recently developed combine the feature selection directly into the model training step including biomarker identification by feature wrappers as done by Xiong et al. (2001), and the nearest shrunken centroid method developed by Tibshirani et al. (2002). The present authors chose to apply the nearest shrunken centroid method in the second step of the analysis workflow due to the following considerations. First, the principle on which the algorithm was built is consistent with the desired biomarkers in drug development, i.e. the small set of molecules with large yet stable changes. Second, the algorithm evaluates each molecular feature independently and ignores any correlation among the molecules. This apparently wasteful practice actually fits the typical experimental setting in an early stage of many biomarker discovery projects using high-throughput technologies, because with a large number of molecular features surveyed in a small number of samples, it is difficult to estimate reliably any complicated correlations among the large number of molecules. By using this algorithm, it was chosen to ignore any correlation information that could not be reliably obtained in this stage of discovery. Due to a small sampling size compared with a large number of molecular features surveyed, over-fitting is a great concern in the supervised analysis. Thus, instead of creating a complex but unstable predictive model, our analysis that is objective at this stage of biomarker discovery is to use the shrunken centroid algorithm to rank molecules according to their discriminant power and to select a set of top-ranked molecules, ideally evaluated with functional relevancy and a practical, easy assay, for further experimental characterization.

Acknowledgement

The authors thank Dr Seppo J. Karrila for critical comments.

References

Affymetrix. 2002. GeneChip Expression Analysis Technical Manual. Affymetrix, Santa Clara, CA, USA.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, et al. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406:536–540.

Bozinov D, Rahnenfuhrer J. 2002. Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering. Bioinformatics 18:747–756.

Breiman L. 1996. Bagging predictors. Machine Learning 24:123–140.

Butte A. 2002. The use and analysis of microarray data. Nature Reviews 1:951–960.

Duda RO, Hart PE, Stork DG. 2001. Pattern classification. New York: Wiley.

Dudoit S, Fridlyand J. 2003. Bagging to improve the accuracy of a clustering procedure. Bioinformatics 19:1090–1099.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, USA 95: 14863–14868.

Etgen GJ, Oldham BA. 2000. Profiling of Zucker diabetic fatty rats in their progression to the overt diabetic state. Metabolism 49:684–688.

Frank R, Hargreaves R. 2003. Clinical biomarkers in drug discovery and development. Nature Reviews 2:566–580.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16:906–914.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537.

Hamadeh HK, Bushel PR, Jayadev S, DiSorbo O, Bennett L, Li LP, Tennant R, Stoll R, Barrett JC, Paules RS, Blanchard K, Afshari CA. 2002. Prediction of compound signature using high density gene expression profiling. Toxicology Science 67:232–240.

He QY, Chiu JF. 2003. Proteomics in biomarker discovery and drug development. Journal of Cell Biochemistry 89:868–886.

Ilyin SE, Belkowski SM, Plata-Salaman CR. 2004. Biomarker discovery and validation: technologies and integrative approaches. Trends in Biotechnology 22:411–416.

Jolliffe IT. 2002. Choosing a subset of principal components or variables. In: Jolliffe IT, editor. Principal Components Analysis. 2nd ed. New York: Springer-Verlag. pp. 111–147.

Lee Y, Lee CK. 2003. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics 19:1132–1139.

Lindon JC, Holmes E, Antti H, Bollard ME, Keun H, Beckonert O, Ebbels TM, Reily MD, Robertson D, Stevens GJ, et al. 2003. Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. Toxicology and Applied Pharmacology 187:137–146.

Maindonald J, Braun J. 2003. Data analysis and graphics using R. Milan: McGraw-Hill.

Pang JX, Ginanni N, Dongre AR, Hefta SA, Opitek GJ. 2002. Biomarker discovery in urine by proteomics. Journal of Proteome Research 1:161–169.

Robertson DG, Reily MD, Albassam M, Dethloff LA. 2001. Metabonomic assessment of vasculitis in rats. Cardiovascular Toxicology 1:7–19.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, et al. 2000. Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics 24:227–235.

Simon R, Radmacher MD, Dobbin K, McShane LM. 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute 95:14–18.

Soukas A, Cohen P, Socci ND, Friedman JM. 2000. Leptin-specific patterns of gene expression in white adipose tissue. Genes and Development 14:963–980.

Tibshirani R, Hastie T, Narasimhan B, Chu G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences, USA 99: 6567–6572.

Timbrell JA. 1998. Biomarkers in toxicology. Toxicology 129:1–12.

Valentini G. 2002. Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles. Artificial and Intelligent Medicine 26:281–304.

Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, et al. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Molecular Biology of the Cell 13:1977–2000.

Xiong M, Fang X, Zhao J. 2001. Biomarker identification by feature wrappers. Genome Research 11:1878–1887.

Yamada J, Suga K, Furihata T, Kitahara M, Watanabe T, Hosokawa M, Satoh T, Suga T. 1998. cDNA cloning and genomic organization of peroxisome proliferator-inducible long-chain acyl-CoA hydrolase from rat liver cytosol. Biochemical and Biophysical Research Communications 248:608–612.

Zhang H, Yu CY. 2002. Tree-based analysis of microarray data for classifying breast cancer. Front Biosci 7:c63–c67.