RESEARCH ARTICLE

Discovering robust protein biomarkers for disease from relative expression reversals in 2-D DIGE data

Troy J. Anderson[1]*, *Irina Tchernyshyov*[2]*, *Roberto Diez*[3], *Robert N. Cole*[4], *Donald Geman*[5], *Chi V. Dang*[2] and *Raimond L. Winslow*[1]

[1] Center for Cardiovascular Bioinformatics and Modeling and The Institute of Computational Medicine, Johns Hopkins University, Baltimore, MD, USA
[2] Division of Hematology, Department of Medicine and The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, USA
[3] Mass Spectrometry & Proteomic Facility, Johns Hopkins University, Baltimore, MD, USA
[4] Biological Chemistry, Johns Hopkins University, Baltimore, MD, USA
[5] Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA

This study assesses the ability of a novel family of machine learning algorithms to identify changes in relative protein expression levels, measured using 2-D DIGE data, which support accurate class prediction. The analysis was done using a training set of 36 total cellular lysates comprised of six normal and three cancer biological replicates (the remaining are technical replicates) and a validation set of four normal and two cancer samples. Protein samples were separated by 2-D DIGE and expression was quantified using DeCyder-2D Differential Analysis Software. The relative expression reversal (RER) classifier correctly classified 9/9 training biological samples ($p<0.022$) as estimated using a modified version of leave one out cross validation and 6/6 validation samples. The classification rule involved comparison of expression levels for a single pair of protein spots, tropomyosin isoforms and α-enolase, both of which have prior association as potential biomarkers in cancer. The data was also analyzed using algorithms similar to those found in the extended data analysis package of DeCyder software. We propose that by accounting for sources of within- and between-gel variation, RER classifiers applied to 2-D DIGE data provide a useful approach for identifying biomarkers that discriminate among protein samples of interest.

Received: May 18, 2006
Revised: December 11, 2006
Accepted: January 3, 2007

Keywords:
2-DE / Bioinformatics / Classification / High-throughput / Protein profile

Correspondence: Mr. Troy J. Anderson, Biomedical Engineering, Johns Hopkins University, 3400 N. Charles St., Clark 202, Baltimore, MD 21218, USA
E-mail: troy_anderson@jhu.edu
Fax: +1-410-516-5294

Abbreviations: **BVA**, biological variance analysis; **DIA**, differential in-gel analysis; **EBNA2,** Epstein–Barr viral nuclear antigen 2; **EDA**, extended data analysis; **K**, cross-validated parameter for number of pairs; **k-NN**, k-nearest neighbors; **K-TSPs**, 'K' top scoring pairs; **LOOCV**, leave one out cross validation; **PAM,** prediction analysis for microarrays; **RDA**, regularized discriminant analysis; **RER,** relative expression reversal; **RER K-TSP**, 'K' top scoring pair algorithm; **RER TSP**, top scoring pair algorithm; **SVM**, support vector machines; **TSPs,** top scoring pairs.

1 Introduction

Rapid progress in the development of new technologies for measuring protein expression patterns has stimulated great interest in how these patterns may be used to discover biomarkers that support sensitive and specific prediction of disease class and progression. 2-D PAGE is a classic method for the evaluation of global protein expression that has long been plagued with a lack of reproducibility due to gel to gel technical variability. Specifically, Voss and Haberl [1] observed that in the use of 2-D PAGE, "the same amount of

* These authors contributed equally to this work

the same protein can run in different positions with different degrees of spatial resolution, and have different spot intensities on different gels." Unlu *et al.* [2] noted that technical variability may arise from "inhomogeneities in the polyacrylamide gels, electric and pH fields and thermal fluctuations." To combat the effects of technical variability in 2-D PAGE, Unlu *et al.* [2] developed 2-D DIGE. Alban *et al.* [3] then introduced the use of an internal standard and demonstrated clearly that 2-D DIGE, using this internal standard, exhibits increased reproducibility and accuracy compared to 2-D PAGE. By reducing technical variability, 2-D DIGE, with the use of an internal standard, has become a powerful tool for comparative proteomic analysis.

While the 2-D DIGE platform decreases the effects of technical variability, it does not help to reduce monotonic expression variability without the use of a normalization technique. Monotonic expression variability is a type of technical variability that arises when there is uniform scaling of protein spot expression values from one gel to the next or from one fluorescent CyDye to another. A shift in the internal standard (or sample) protein loading can monotonically alter protein spot intensities from gel to gel. Furthermore, small pipetting differences when applying the dyes to the samples can cause one dye to yield a uniformly greater magnitude of spot intensity than the other during analysis, creating a CyDye to CyDye monotonic shift. Differences in the properties of the CyDyes have also been suggested to cause uniform scaling of protein spot expression from CyDye to CyDye [4, 5]. As a consequence, many experimenters use normalization and dye swapping techniques to neutralize this type of variability. It has been demonstrated that these normalization techniques can influence the protein spots exhibiting differential expression. Specifically, different normalization techniques yield different significant proteins [4, 6]. Thus, it would be advantageous to analyze 2-D DIGE data with a method that does not require normalization or dye swapping, yet produces results that remain invariant to monotonic expression variability.

Another important factor currently limiting biomarker discovery, concerns the so-called "small-sample dilemma" that has been documented in gene microarray studies [7]. The small-sample dilemma refers to the difficulties that arise in statistical learning and inference when the number of samples is very small compared to the number of features of interest (matched spots across gels in this case). DIGE experiments, which often include greater sample sizes relative to other commonly used differential-display proteomic techniques, typically fall within the small sample regime [8–11]. The number of sample gels available is typically extremely small when one considers the complexity of the underlying biological systems and processes being studied. A further complication arises from the common practice of running technical replicates. Technical replicates are often referred to as repeats because they are simply the same biological sample run on two or more gels. This increases the number of samples, however technical replicates cannot be regarded as independent of each other because they share identical biological sample [12]. The dependence among technical replicates is a complication that must be considered in the learning process. Consequently, it is difficult to unravel the underlying structure within these data, particularly correlation patterns or even higher-dimensional interactions among the protein expression values. This obstacle has been well-documented in the gene microarray literature [13–16] and remains a key issue in 2-D DIGE studies. Standard methods in statistical learning, designed for more favorable ratios of independent samples to features, often lead to over-fitting and inflated estimates of classifier performance in the small-sample regime.

A final problem limiting biomarker discovery concerns the interpretability and ultimate utility of complex decision rules based on inference procedures. The problem is most obvious when sophisticated techniques from statistical pattern recognition and machine learning, such as neural networks, random forests and support vector machines, are applied to classify biomedical data [7]. These methods may perform well, but the rules they employ to make decisions may be too complex to interpret. Our objectives extend beyond classification; we want to explicitly characterize the important interactions among the biological variables being measured.

We propose that recently developed statistical learning methods applied to 2-D DIGE, with use of an internal standard, address these issues and are therefore well suited for robust protein biomarker discovery. These methods are known as relative expression reversal (RER) classifiers. They provide classification rules that are easy to interpret and that perform well when learning in the small-sample regime [7, 17, 18]. The RER classifiers developed to date include the top-scoring pair (TSP) classifier [7] and the cross-validated parameter for number of pairs (K)-TSP classifier [17]. Tan *et al.* [17] demonstrated that the RER classifier's performance met or exceeded the performance of five other well known, more complex classifiers, when tested on microarray data. The RER classifiers are especially favorable in terms of the number of features used and the interpretability of the classification rule. Furthermore, the RER family of classifiers has been shown to be useful in the integration of data across different studies for the purpose of increasing sample size, due to their invariance to monotonic data transformations such as normalization [18]. These methods should then be useful in analysis of 2-D DIGE data because the dependence of the analysis on normalization and dye swapping is removed. Due to the relative advantages (in accuracy, interpretability and data integration) of RER classifiers over other algorithms previously used to analyze genomic data, we believe RER classifiers may be equally advantageous in the analysis of 2-D DIGE experiments.

In this work, we will apply RER classifier methods to 2-D DIGE data in order to discover protein biomarkers. Algorithms similar to those found in the DeCyder-2D Differential Analysis Software Extended Data Analysis (EDA) package as well as linear support vector machines (SVM) will be applied to the data for performance comparison. We will also introduce the proper way to estimate generalization error in the presence of technical replicates. RER classifier methods will be used to account for monotonic expression variability from gel to gel (due to protein loading changes) and CyDye to CyDye (due to differences in the fluorescent dyes) in 2-D DIGE. We will show that the RER classifier, coupled with 2-D DIGE using an internal standard, is a useful approach to biomarker discovery. When used together, they provide interpretable results that are formed by reducing the effects of technical variability, exposing the biological differences among samples.

2 Materials and methods

2.1 Experimental design

The proteomic samples were total cellular lysates of the P493-6 cell line, a human B-lymphocyte immortalized by an Epstein–Barr virus genome and transfected with an inducible MYC oncogene construct. The P493-6 cells were rendered estrogen (β-estradiol) dependent by engineering the Epstein–Barr virus genome to have a fusion of Epstein–Barr viral nuclear antigen 2 (EBNA2) with a hormone binding domain of the estrogen receptor. This cell line also carries a conditional MYC construct under negative control of the tetracycline responsive operator [19]. The modifications provide the ability to manipulate MYC expression level within these cells in such a way as to model very different cellular states. In the absence of tetracycline and β-estradiol (a state referred to as "high MYC"), ectopic MYC oncogene is expressed at a very high level similar to that in Burkitt's lymphoma. At this level, P493 cells are tumorgenic in immune-compromised mice (data not shown). Addition of tetracycline (referred to as "no MYC") to the media effectively shuts down expression of MYC from this construct and returns cells to a non-proliferating state characteristic of the primary, resting B-lymphocytes. When the P493 cells are grown in the presence of β-estradiol and tetracycline (referred to as "low MYC"), EBNA2 is activated and it initiates reentry into the cell cycle by directly inducing endogenous MYC. Activation of EBNA2 provides a non-neoplastic proliferating cell model that is not tumorgenic. These three states provide two non-neoplastic or normal states and one cancer state, summarized in Table 1. By building a classifier that distinguishes the normal states from the cancer state, the decision rules can be interpreted to give insight into the proteomic pathways that lead to uncontrolled cellular growth.

2.2 The 2-D DIGE protocol

Cells were cultured in RPMI 1640 culture media, supplied with 10% FCS and 1% penicillin/streptomycin solution. Cells were washed in an ice-cold low salt wash buffer (5 mM magnesium acetate, 10 mM Tris-HCl, pH 7.0, 250 mM sucrose). Approximately 10^7 cells were extracted in 100 μL of the lysis buffer (8 M urea, 4% CHAPS, 10 mM Tris-HCl pH 8.5). The crude cell homogenate was sonicated three times for 5 s on ice then incubated for 30 min at room temperature with vortexing. The protein extracts were clarified by centrifugation for 15 min at 16 000 rpm and protein concentration was measured using the BCA Protein Assay (Pierce).

Table 1. Reference table for the created cellular states.

Name:	no MYC	low MYC	high MYC
Model state:	Normal non-proliferating	Normal proliferating	Cancer
Tetracycline present?	yes	yes	no
Beta-estrodiol present?	no	yes	no

Samples were labeled with CyDye Fluor minimal dyes (GE Healthcare) according to manufacturers instructions. Prior to labeling, each sample was diluted to 2 mg/mL with DIGE labeling buffer (8 M urea, 4% CHAPS, 10 mM Tris-HCl pH 8.5). The pH of the samples were monitored and adjusted to pH 8.5 with 1 M Tris-HCl (pH 9.5). An equal amount of each sample included in the experimental sample set was combined to create the internal standard. Samples were labeled with respective CyDyes according to Table 2 for the training sample set and Table 3 for validation sample set. The internal standard in each case was labeled with Cy2.

The first-dimension IEF was performed in an IPGphor IEF unit (GE Healthcare) on 7 cm IPG strips pH 3–10 (GE Healthcare). Nine micrograms total protein *per* gel were further diluted in rehydration buffer (8 M urea, 4% CHAPS, 1% DTT, 1.5 % IPG buffer pH 3–10) in order to bring the volume to a total of 115 μL and loaded onto the strip by active rehydration. IEF was carried out as follows: 50 V rehydration step for 12 h, 500 V to 250 V·h, 1000 V to 500 V·h, 4000 V to total applied 10 000 volt hours. After IEF strips were equilibrated for 15 min (in buffer containing 50 mM Tris-HCl pH 8.8, 6 M urea, 30% v/v glycerol, 2% SDS, 1% DTT), they were alkylated for another 15 min in the same buffer where DTT was substituted with 4% iodoacetamide. Second dimension SDS-PAGE was performed on 4–12% NuPage Bis-Tris gels (Invitrogen) for 1.5 h at 150 V on the XCell mini gel apparatus (Invitrogen).

Gels were scanned on a Typhoon 9400 (GE Healthcare), at the appropriate excitation and emission wavelengths for each CyDye (Cy2, Cy3, and Cy5), generating three images per gel. Gel images were analyzed using the DeCyder-2D Differential Analysis Software 5.0 [20] (GE Healthcare). For spot detection and quantification, the differential in-gel analysis (DIA) module of DeCyder was employed. The biological variance analysis module (BVA) was then used to match the quantified spots of all gels to a chosen master gel. In addition to log standardized abundances, the matched spot raw

Table 2. Training 2-D DIGE sample layout.

Gel	Cy2	Cy3	Cy5
1	Internal Standard	high MYC 1	no MYC 1
2	Internal Standard	high MYC 1	no MYC 1
3	Internal Standard	high MYC 1	no MYC 1
4	Internal Standard	high MYC 1	low MYC 1
5	Internal Standard	high MYC 1	low MYC 1
6	Internal Standard	high MYC 1	low MYC 1
7	Internal Standard	high MYC 2	no MYC 2
8	Internal Standard	high MYC 2	no MYC 2
9	Internal Standard	high MYC 2	no MYC 2
10	Internal Standard	high MYC 2	low MYC 2
11	Internal Standard	high MYC 2	low MYC 2
12	Internal Standard	high MYC 2	low MYC 2
13	Internal Standard	high MYC 3	no MYC 3
14	Internal Standard	high MYC 3	no MYC 3
15	Internal Standard	high MYC 3	no MYC 3
16	Internal Standard	high MYC 3	low MYC 3
17	Internal Standard	high MYC 3	low MYC 3
18	Internal Standard	high MYC 3	low MYC 3

Table 3. Validation 2-D DIGE sample layout.

1	Internal Standard	high MYC 4	low MYC 4
2	Internal Standard	low MYC 5	no MYC 4
3	Internal Standard	no MYC 5	high MYC 5

volume data of each sample was used in this analysis. One gel yields (for all *p*): Volume(Cy2)$_{p,g}$,Volume(Cy3)$_{p,g}$, and Volume (Cy5)$_{p,g}$, where *p* is the matched protein reference number and *g* corresponds to a particular gel. The volume of a spot is calculated as a function of the area and intensity of the spot on a gel.

The raw volume data for each gel was output from DeCyder in a tab delimited file, constructed by the XML toolbox of DeCyder. A ratio was created by comparing the raw volume of each protein spot to that of its intra-gel internal standard (Eqs. 1 and 2):

$$\text{RatioCy3}_{p,g} = \text{Volume(Cy3)}_{p,g} / \text{Volume(Cy2)}_{p,g} \quad (1)$$

$$\text{RatioCy5}_{p,g} = \text{Volume(Cy5)}_{p,g} / \text{Volume(Cy2)}_{p,g} \quad (2)$$

for each protein spot number *p* and gel *g*, these ratios will also be referred to as "raw BVA ratios". The log standardized abundances were also output for each gel and parsed into a separate comma delimited file. By outputting both normalized and non-normalized data from the same gels it allowed us to demonstrate the monotonic invariance property of the RER methods.

2.3 LC/MS/MS

A preparative gel was run and protein spots were visualized by CBB staining (not shown). Spots matched to spot numbers 530 and 786 of the master gel were excised from the gel and tryptic digests of their proteins were separated by RP chromatography using a 2-D nano HPLC (Eksigent) and electrosprayed directly into a LTQ mass spectrometer (Thermo Finnigan). Proteins were identified using the MASCOT Daemon software (Matrix Science) to search fragmentation spectra of tryptic peptides against the nonredundant mammalian database from National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov).

2.4 Data considerations

There are three biological replicates of each state labeled as: high MYC 1, 2, 3; no MYC 1, 2, 3; and low MYC 1, 2, 3 (see Table 2). Therefore, there are 9 total samples that are biological replicates and the rest are technical replicates. As discussed previously, a recent paper by Karp *et al.* [12] demonstrated that it is incorrect to assume independence of technical replicates. To properly account for the dependence among these technical replicates, a new method for estimation of generalization error must be used, described in Section 2.5.

A difficulty with BVA is that it is not always possible to match spots on the master gel with a corresponding spot on non-master gels, a problem we refer to as "missing data". The missing data issue is commonly resolved by inputting zeros for the missing protein spot volumes on non-master gels. This has the biological interpretation of the protein not being expressed. Meleth *et al.* [6] found that inputting zeros can lead to inflated *p*-values in parametric analysis due to its effect on variance. Because the RER classifiers are non-parametric rank-based classifiers (see Section 2.5), the effect of inputting all zeros on variance in our experiment is minimal. Therefore, in this study, zeros were substituted for all missing data.

2.5 RER implementation

A complete description of the original RER method, known as the TSP RER algorithm, is presented in Geman *et al.* [7]. The K-TSP RER algorithm, which is used in this study, is an extension of the TSP algorithm. The code for the program can be downloaded by visiting http://www.ccbm.jhu.edu/aboutus/news-ktsp.php. A description of the K-TSP RER algorithm can be found in Tan *et al.* [17]. Below, we give a simple, intuitive description of the RER classifiers applied to 2-D DIGE. Also, we explain how this classifier accounts for monotonic expression variability, eliminating the dependence of the analysis on normalization and dye swapping.

2.5.1 Training the RER classifier

Let $\mathbf{X}_i = \{X_{i,1}, X_{i,2}, X_{i,3}, \ldots. X_{i,p}\}$ represent the expression profile of sample *i*, $1 \le i \le n$, where *i* is a single fluorescent CyDye of a 2-D DIGE gel and *n* is the total number of Cy3 and Cy5 dyes from all gels in the experiment. $\{X_{i,1}, X_{i,2}, X_{i,3}, \ldots. X_{i,p}\}$ are the raw BVA ratios from equations 1 or 2 (or log standardized abundances) corresponding to each protein spot number *p* of sample *i*. Let Y_i represent the class label of $\mathbf{X}_i$, with $Y_i = 0$ indicating normal and $Y_i = 1$ indicating cancer, and suppose there are *m* total normal samples and *q* total cancer samples ($m + q = n$). Finally, let *M* be the subset of indices of profiles $\mathbf{X}_i$ that are in the normal class and let *Q* be the subset that are in the cancer class.

RER algorithms define a primary score for each possible pair of protein spots based on the ability of their relative expression values within a profile to discriminate between the two classes. Let $P^0_{j,k}$ represent the fraction of normal samples which respond positively to the logical question "Is the raw BVA ratio magnitude of protein spot *j* greater than that of protein spot *k*?" Let $P^1_{j,k}$ represent the fraction of cancer samples which respond positively to the same logical question. More precisely, these fractions, as estimated from the training samples, are given by Eqs. (3) and (4) below:

$$P^0_{j,k} = \frac{\sum_{i \in M} I_{\{X_{i,j} > X_{i,k}\}}}{m} \quad (3)$$

$$P^1_{j,k} = \frac{\sum_{i \in Q} I_{\{X_{i,j} > X_{i,k}\}}}{q} \quad (4)$$

where (Eq. 5)

$$I_{\{X_{i,j} > X_{i,k}\}} = \begin{cases} 1, \ if \ X_{i,j} > X_{i,k} \\ 0, \ if \ X_{i,j} \le X_{i,k} \end{cases} \quad (5)$$

The estimated probabilities $P^0_{j,k}$ and $P^1_{j,k}$ are then used to obtain the primary score $\Delta_{j,k}$ defined by Eq. (6)

$$\Delta_{j,k} = \left| P^0_{j,k} - P^1_{j,k} \right| \quad (6)$$

which can assume values between 0 and 1. Inspection of Eqs. (3)–(6) shows that the largest scores will originate from those pairs of protein spot raw BVA ratios whose expression values within a channel profile invert most often relative to each other from class '0' to class '1', referred to as a RER. In other words, a pair of protein spots has a large score if most of the indicators are true in Eq. (3) and few are true in Eq. (4), or *vice versa*. The primary score, $\Delta_{j,k}$, is calculated for all possible pairs of protein spots *j* and *k* ($j \neq k$), then the pairs are ordered from greatest to least according to this score. Those protein spots that achieve a high primary score are viewed as the most informative for classification. The spot pairs which have tied for the top score are now chosen as the TSPs. Each of the TSPs is then successively used to classify an unknown sample. Consider each assignment of a sample to a class (by a TSP) as a vote for that class. The class which receives the majority of the votes from among the TSPs is then assigned to the unknown sample (see Geman *et al.* [7]).

The K-TSP RER algorithm is a simple extension of the TSP RER algorithm in which K disjoint pairs are allowed to vote. Some of these K pairs may not be top-scoring pairs. The positive integer K is a parameter which itself must be cross-validated in estimating the generalization error of the K-TSP classifier. If the generalization error is estimated by leave one out cross validation (LOOCV), and is to be unbiased, then another ("inner") loop of LOOCV is necessary to discover the best performing number of pairs (the K parameter) [15]. The pairs retained in the K-TSP classifier will be referred to as the K-TSPs. Similar to the TSP method, the K-TSPs are then used in a majority voting scheme to determine the class of unknown samples. See Tan *et al.* [17] for details describing the RER algorithm known as K-TSP.

RER methods use only the relative magnitude of features within the same gel to build a classifier. Thus, monotonic expression variability from CyDye to CyDye and gel to gel (which does not affect the rank order based on magnitude within a sample) does not affect the TSPs or K-TSPs chosen for the classifier. RER invariance to monotonic transformation is discussed further in the study by Xu *et al.* [18].

2.5.2 Estimation of generalization error

In order to estimate the generalization error rate of the K-TSP classifier, a modified version of LOOCV is used. Because technical replicates of the same biological sample cannot be considered independent of each other, simple LOOCV cannot be used. Instead, all technical replicates of a single biological replicate must be left out during each loop of cross validation to insure that the classifier is not learning from data that is dependent upon the data being left out. The following is an example of a single learning loop of the classifier: first, one biological sample and all of its technical replicates (for instance all three low MYC 1 samples) are left out from the *n* total samples in what we will refer to as the "outer loop." We will refer to the number of left out samples as *t*. Following this step, the *n* - *t* remaining samples are used to train a surrogate K-TSP classifier. Here, a loop of LOOCV (referred to as the "inner loop") is used on the *n* - *t* samples in order to find the best performing parameter K. The K-TSPs (K determined from last step) based on the sample size of *n* - *t* determine the surrogate classifier, which is then used to classify the *t* samples left out in the original outer loop. Consequently, no biological information from the *t* samples left out in the outer-loop is used to train the surrogate classifier. Overall, each of the biological samples and its respective technical replicates are left out once and classified by a newly constructed surrogate classifier. It should be noted that each surrogate classifier can potentially have a different value for the parameter K as well as a different set of pairs of protein spots. The final estimated RER generalization error rate is the ratio of the number of biological samples incorrectly classified to the total number of biological samples. Each biological sample will be considered correctly classified only if all corresponding technical replicates of the sample are correctly classified. This method accounts for the presence of technical replicates, properly estimating generalization error. It simulates the situation of attempting to classify new samples of unknown class.

2.6 Implementation of other machine learning algorithms

The EDA package was not used directly to implement its algorithms because, to our knowledge, the DeCyder program has no capability to properly handle technical replicates, as noted by Karp *et al.* [12]. Instead, regularized discriminant analysis (RDA) [21] was implemented using the open source R statistical package (http://www.r-project.org), and the two other algorithms (k-NN and SVM) were implemented using WEKA software [22]. A complete general description of each classifier can be found in the book by Hastie *et al.* [23] Briefly, RDA is a compromise between classical linear discriminant analysis (LDA), which assumes Gaussian data with a common, class-independent covariance matrix, and the more general quadratic discriminant analysis (QDA). SVM separates two classes by generating the hyperplane (in a high-dimensional feature space) which maximizes the distance from the hyperplane to the closest training examples. Finally, k-NN classifies a new sample with the majority label among the k nearest samples in the training set (as measured by Euclidean distance in this case). All three have proven effective in molecular classification, although the decision rules are high-dimensional and difficult to interpret or visualize in practice.

While the RDA and k-NN algorithms are not fully identical to those provided in EDA, they should provide similar classification results, as they use the same basic principles. SVM was chosen as an additional classification method because it had previously been shown to perform as well as the RER methods on microarray data [17]. Estimation of generalization error for these algorithms was performed using the same modified version of LOOCV as for RER to account for technical replicates, with any parameters being determined by an inner loop of cross-validation.

2.7 Permutation analysis

Permutation analysis is used to determine the significance of the RER generalization error rate. In this analysis, a new permuted training set is created from the original training set. The class labels (Y_i) of the original training set are randomly permuted (shuffled), subject to leaving the actual profiles and overall numbers of samples (*m, q*) in each class unchanged. However, because technical replicates of the same biological sample have dependent profiles, the permutation cannot be completely random. In order to account for the presence of technical replicates, only the biological labels are shuffled, then the technical replicates are assigned the same label as their corresponding biological sample. The permuted data set is subsequently run through the classifier to obtain the estimated generalization error (again using the modified version of LOOCV which accounts for the presence of technical replicates.) This process is repeated many times and each repetition, referred to as a "permutation trial", results in an estimated error rate. The distribution of error rates over all permutation trials is used to measure significance. Combining the estimated error rates from all trials yields a histogram of error rates which provides a reference distribution for evaluating the error rate obtained with the original (unpermuted) data. To determine the significance of the original RER generalization error rate, the percentage of permutation trials that achieve an error rate equal to or better than the original is recorded. This percentage is the estimated *p*-value or probability that the original generalization error was achieved under random labeling of the samples. The typical cutoff for significance is a *p*-value of $p \leq 0.05$.

3 Results

DeCyder-2D Differential Analysis (DIA and BVA modules) resulted in the detection, quantization, and matching of 1098 master gel spots across the 18 gels and 54 gel dyes. Log standardized abundance data as well as raw BVA ratio data was output (as described in section 2.1) for each sample. These data were input to k-NN, RDA, SVM and RER classifiers. All the images and data sets used in this study are available for download at: http://www.ccbm.jhu.edu/research/dSets.php.

3.1 Machine learning analysis

The K-TSP RER classifier was trained using the 18 normal samples and 18 cancer samples for both the log standardized data and the raw BVA ratio data. The estimated generalization error rate of the RER classification method was found to be zero on both sets of training data, corresponding to the correct classification of each of the 36 samples. Therefore, the RER classification method is predicted to have 100% sensitivity (3/3 cancer biological replicates correctly classified), and 100% specificity (6/6 normal biological replicates samples correctly classified.)

Next, a single RER decision rule was induced from each of the two training sets. The same decision rule was created from both the log standardized abundance data and the raw BVA ratio data. It is interesting to note that the value of *K* found to perform best based on LOOCV is K = 1. Also, both protein spots in the top pair had no inserted zeros and thus no initial missing data. Since K = 1, the classification rule only involves two protein spots (530 and 786) and is very simple (Eq. 7):

IF BVA ratio: $X_{530} \geq$ BVA ratio: X_{786} **THEN** Cancer **ELSE** Normal. (7)

In words, the rule states that if the expression of spot 530 (divided by its internal standard expression) equals or exceeds that of spot 786 within the same sample then that sample is cancerous. Figure 1 shows a scatter plot of the expression of spot 530 to 786 for all samples. The decision rule completely separates the data into the two classes.

Table 4 summarizes the results of applying each of the four classifiers to both the log standardized and the raw BVA ratio data. K-TSP RER, RDA, and SVM all performed similarly on the log standardized data with k-NN underperforming. RDA was able to accurately classify all samples based on the log standardized expression of only spot number 786. However, the K-TSP RER classifier was the only classifier to maintain its performance level and construct the same decision rule on both the log standardized and raw BVA ratio data. All three other classifier methods dropped in estimated performance on the raw BVA ratio data and created different final decision rules.

3.2 Permutation analysis

There were 45 possible non-redundant permuted training sets that maintained the normal to cancer ratio (18 to 18) when including technical replicates. All 45 permuted training sets were created from the original training set (36 samples) and analyzed with RER. Figure 2 shows the estimated K-TSP RER correct classification rates (1-generalization error rate) associated with the 45 possible permutation trials. None of the permutation trials achieved the zero generalization error rate seen in the unpermuted analysis. The estimated *p*-value is less than 1/45 ($p<0.022$), suggesting that the RER generalization error rate did not occur by random chance.

3.3 Validation samples

Six new samples were created and run on three gels by a different experimentalist than the originals and included a dye swap (the sample layout is shown in Table 3). These gels were then quantified by matching to the original master gel and log standardized data was output from the DeCyder program, consistent with the protocol from the methods section. All six samples were biological replicates: two high MYC samples, two low MYC samples and two no MYC samples (four normal and two cancer samples). These new samples provided a blind validation set to which the final decision rules constructed from the original training analysis for all four classifiers were applied for classification. K-TSP RER used the decision rule shown in Eq. (7) (spots 786 and 530), RDA used only protein spot 786 in its rule, and both SVM and K-NN used all 1098 protein spots in their decision rules. Only the log standardized data was used for classification here as it was shown to provide the best performance across all classifiers in the training data (Table 4). The four classifiers were all able to correctly classify the six new samples. The K-TSP RER classification is displayed in Fig. 3.

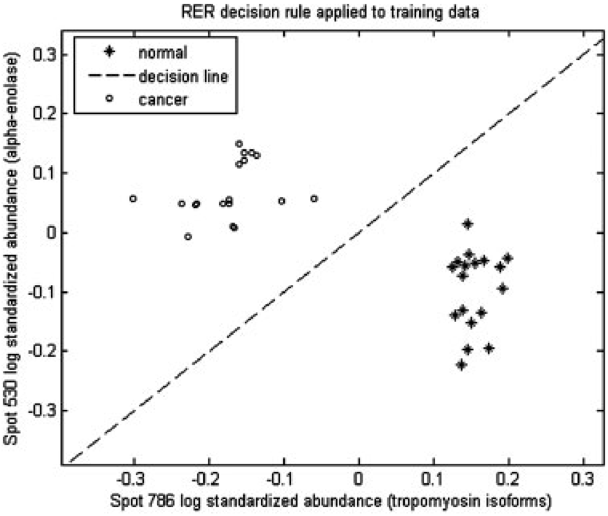

Figure 1. Scatter plot of spot 530 raw BVA ratio *vs.* spot 786 raw BVA ratio in each sample. All 36 training samples are plotted in this space. The dotted line represents where expression for spot 530 equals that of 786. This line is the rule in Eq. (7) that was trained from the original data, above the line samples would be classified as cancer and below as normal. All samples are correctly classified by this rule.

Table 4. Classifier performance comparison using modified LOOCV.

Classifier	*Log Standardized Abundance Data Input*				*Raw BVA Ratio Data Input*			
	Biological Accuracy	Cancer Accuracy	Normal Accuracy	of proteins[a]	Biologcial Accuracy	Cancer Accuracy	Normal Accuracy	of proteins[c]
K-TSP RER	9/9	3/3	6/6	2	9/9	3/3	6/6	2
RDA[b]	9/9	3/3	6/6	1	6/9	2/3	4/6	890
SVM[c]	8/9	3/3	5/6	1098	5/9	1/3	4/6	1098
k-NN[c]	6/9	3/3	3/6	1098	1/9	1/3	0/6	1098

a) The number of proteins that are used in the classification rule for each classifier.
b) Implemented with the RDA package of R, using min-min method described in Guo *et al.* [21]
c) Implemented using the WEKA software package[22].

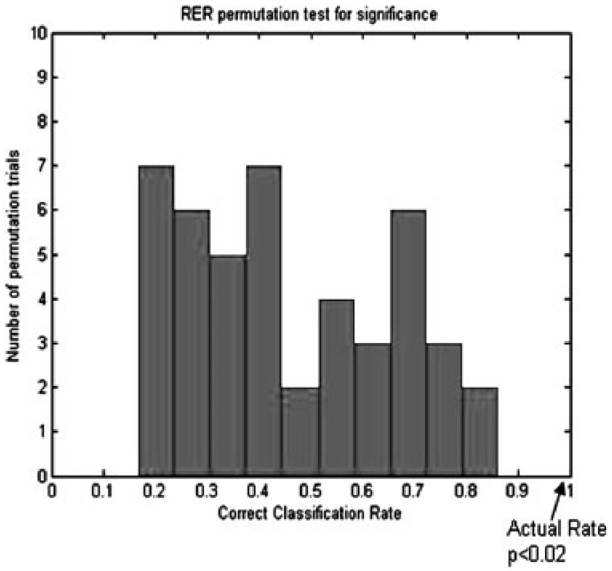

Figure 2. A histogram of the correct classification rate for RER classifiers built in the 45 permutation trials. The arrow points to the original RER correct classification rate on the training data and corresponds to an estimated p-value of <0.022.

The validation samples were also used in order to determine if the technical replicates were necessary for this experiment. Ten new training sets of nine samples (six normal and three cancer) were created from the original training data of 36 samples. For the creation of the new training sets, each sample was selected at random from its group of technical replicates, thus all ten new training sets contained only biological replicates. Each training set was used to train a new K-TSP RER classifier and then tested on the validation samples. All ten new classifiers used spot 786 in their decision rule and were able to correctly classify every validation sample. However, all decision rules were not identical as the other spot in the top scoring pair was not always spot 530.

3.4 Protein spot identification by LC/MS/MS

A preparative gel was run and protein spots were identified as described in Section 2.3. Based on MOWSE search algorithm employed by MASCOT [24], spot number 786 was identified as tropomyosin isoforms 3 and 4 and spot 530 as α-enolase. Protein identifications were based on 12 peptides from tropomyosin 3, 10 peptides from tropomyosin 4, and 13 peptides from α-enolase (Table 5). Only peptides that exceeded the default significant probability-based MOWSE score value ($p<0.05$) were considered. Highly homologous tropomyosin isoforms were distinguished by two peptides specific to each of the isoforms. MS/MS spectra of each peptide were manually examined to ensure that isoform specific amino acid sequences were detected rather than inferred.

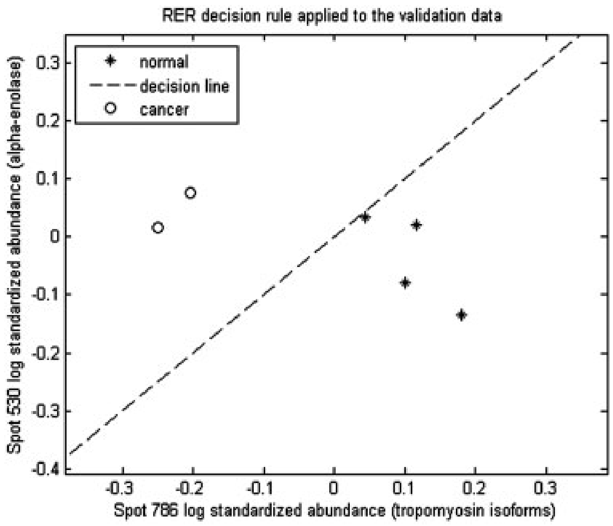

Figure 3. Scatter plot of the validation samples, formatted identically to Fig. 1. It shows that the validation samples were all correctly classified by the RER rule from Eq. (7) that was induced from the training set.

4 Discussion

The EDA package of DeCyder was a step forward in the analysis of 2-D DIGE gels that allowed researchers to employ machine learning methods. This manuscript presented RER as an important alternative to the algorithms found in EDA. Furthermore, we described and implemented the proper method of learning in the presence of technical replicates (currently unavailable in EDA). It is believed EDA does not include capability to handle technical replicates because it is generally accepted that 2-D DIGE reduces technical variability enough to eliminate the running of technical replicates. In the validation section we demonstrated that, when technical replicates were discarded, the K-TSP RER classifier still used spot 786 in its classification rule and correctly classified all validation samples. However, the fact that all decision rules were not identical when technical replicates were randomly discarded indicates that there may be information present in technical replicates. Many researchers still run technical replicates and do not wish to discard that information. Failure to account for technical replicates, when present, could cause over-fitting and inflated performance estimates.

When technical replicates were included, the K-TSP RER algorithm estimated a zero generalization error ($p<0.022$) when applied to the training data and was able to correctly classify all of the validation samples using a simple and interpretable decision rule. However, the RDA algorithm matched the performance of K-TSP RER while using an equally simple decision rule.

Table 4 revealed that K-TSP RER was the only algorithm used in this study that maintained the same performance level and derived the same decision rule when applied to

Table 5. MASCOT results.

Spot number/ Protein name	NCBI accession number	Sequence coverage[a)]	Theoretical mass (kD)[a)]/p*I*[a)]	Peptides (*m/z*, charge)	Matched peptide sequence[b)]/ Ion score[c)]
Spot 530/ Enolase 1 [*Homo sapiens*]	gi|13325287	55%	47481/7.01	450.31, 2+	TIAPALVSK/58
				572.52, 2+	IGAEVYHNLK/63
				640.76, 2+	LMIEMDGTENK/78
				703.85, 2+	GNPTVEVDLFTSK/93
				713.61, 2+	YISPDQLADLYK/73
				760.63, 2+	FGANAILGVSLAVCK/94
				771.17, 2+	VVIGMDVAASEFFR/112
				771.63,2+	LAQANGWGVMVSHR Oxidation(M)/68
				817.63,2+	VNQIGSVTESLQACK/113
				827.19,2+	IDKLMIEMDGTENK Oxidation(M)/57
				902.81, 2+	AAVPSGASTGIYEALELR Oxidation(M)/143
				970.63, 2+	LAMQEFMILPVGAANFR + 2 Oxidation(M)/97
				981.12, 2+	DATNVGDEGGFAPNILENK/91
Spot 786/ Tropomyosin isoform 4 [*Homo sapiens*]	gi|54696136	43%	28619/4.67	508.26, 2+	AEGDVAALNR/ 54[d)]
				546.78, 2+	CGDLEEELK/ 48
				575.54, 2+	MEIQEMQLK/ 53
				585.86, 2+	LVILEGELER/ 74
				622.55, 2+	IQLVEEELDR/ 71
				649.95, 2+	KLVILEGELER/ 65
				700.37, 2+	RIQLVEEELDR/ 71
				808.11, 2+	IQALQQQADEAEDR/ 85
				872.21, 2+	KIQALQQQADEAEDR/ 91[d)]
				595.33, 3+	KLVILEGELERAEER/ 46
Spot 786/ Tropomyosin isoform 3 [*Homo sapiens*]	gi|55665781	48%	29019/4.72	566.51, 2+	MELQEIQLK/ 63
				578.79, 2+	LVIIEGDLER/ 80
				622.54, 2+	IQLVEEELDR/ 71
				643.15, 2+	KLVIIEGDLER/ 67
				658.85, 2+	EQAEAEVASLNR/ 55
				700.40, 2+	RIQLVEEELDR/ 68
				772.47, 2+	AREQAEAEVASLNR/ 97[d)]
				822.14, 2+	IQVLQQQADDAEER/ 101
				864.68, 2+	IQLVEEELDRAQER/ 46
				886.23, 2+	KIQVLQQQADDAEER/ 113[d]
				981.25, 2+	ALKDEEKMELQEIQLK + Oxidation(M)/ 71
				1000.24,2+	IQVLQQQADDAEERAER/ 65

a) Calculated automatically by MASCOT

b) Only peptides with individual scores >45 were included in the matched peptide list.

c) MASCOT seacrch algorithm reports ion scores as -10*log(P), where P is the probablility that the observed match is a random event. Scores >45 indicate identity or extensive homology ($p<0.05$)

d) Peptides are specific to the particular isoform

both types of training data. The fact that the decision rules of the other classifiers changed when they were applied to the raw BVA ratio data indicates that the biomarkers and/or molecular pathways and mechanisms they uncover are dependent on normalization of the data. RER is unaffected by normalization which demonstrates that monotonic shifts in the data, such as the types discussed in the introduction, will not effect the final decision rule.

The K-TSP RER classification algorithm identified the comparison of protein expressions for tropomyosin isoforms 3 and 4 and α-enolase (spots 786 and 530) as the best method for discriminating between normal and cancer states in both types of training data. Visual inspection of the gels was used to verify the existence of protein spots 786 and 530. The RER classifier rule can be evaluated in Fig. 4, both visually and numerically, in order to show that the rule in Eq. (7) holds for

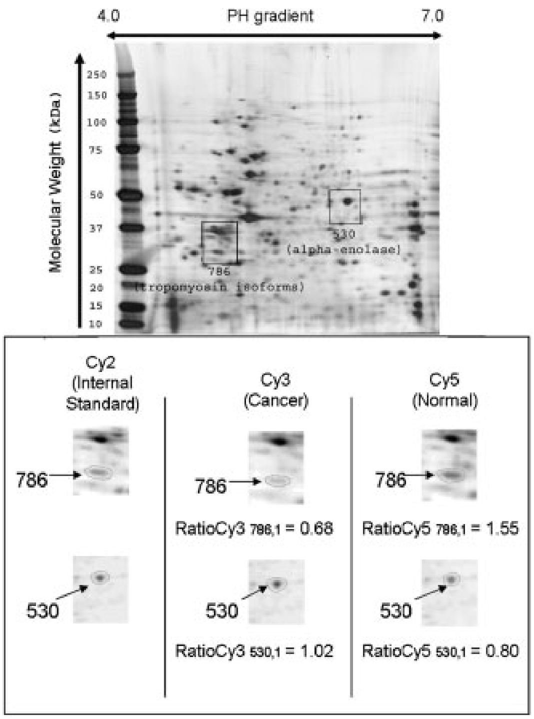

Figure 4. Visualization of protein spots 786 and 530, tropomyosin isoforms and α-enolase, on gel #1. Top is the silver stained image of gel #1. The boxes shown on the top gel are the areas that are enlarged below for visualization of the spots on each of the three channels. For Cy3 and Cy5, the corresponding ratio to the internal standard for the displayed proteins is also shown quantitatively.

gel 1. Two key concepts should be kept in mind here: first, the feature of interest is the expression relative to that of the internal standard; second, the rule is applied within each CyDye not across CyDyes. Visualization of the other 17 gels verified that two spots are present in every gel and appear to have been matched correctly across all gels (not shown). Overall, evidence provided by an estimated zero generalization error, successful permutation analysis, and visual inspection indicated that the spots used in the RER decision rule exhibit RER that can be used to perform accurate and sensitive classification.

The K-TSP RER classifier's performance was clearly better than k-NN and SVM when considering correct classification rate and interpretability. However, RDA classification was able to match K-TSP RER in terms of performance and interpretability (use of a small number of features) when applied to the log standardized training data (Table 4.) The RDA classifier's final decision rule only involved protein spot 786 (tropomyosin isoforms 3 and 4.) The fact that both the K-TSP RER and RDA classifiers independently identified the tropomyosin isoforms as a biomarker to be used for further classification increases our confidence that it is a true biomarker. This exemplifies the reason it is important to have several independently developed algorithms shown to perform well on the data.

The proteins used in the K-TSP RER and RDA decision rules have also been implicated in other cancers and previously proposed as biomarkers. Both of these proteins appear often on the differential expression profiles of various cancers [25, 26]. Tropomyosin participates in the formation of cytoskeleton by binding to and stabilizing actin fibers and it is involved in a variety of functions such as cytokinesis, intracellular transport, cell motility and creating cell structure. Eisenman and colleagues [27] identified isoforms 3 and 4 of tropomyosin in their quantitative analysis of c-myc function as cytoskeletal proteins down regulated in c-myc over expressing cells. They also demonstrated that disruption of actin network proteins results in a reduction of actin fibers and an increase in cell motility leading to higher metastatic potential. α-Enolase is an important glycolytic enzyme which catalyses the conversion of 2-phosphoglycerate to phosphoenol pyruvate. The ENO1 gene was previously identified as one of the several genes involved in glucose metabolism directly regulated by c-Myc [28]. Enhanced expression of glycolytic enzymes in tumors provides significant advantage to the rapidly proliferating cancer cells. Several studies reported a correlation between the tumor progression and α-enolase expression [25, 29] and suggested it as a suitable candidate for a biomarker of cancer. In this study, we have provided evidence that the two previously indicated biomarkers, tropomyosin isoforms and α-enolase, may enhance utility when used together. We also demonstrated that the tropomyosin isoforms were found to be an important biomarker by two different types of machine learning algorithms.

Classification of the validation samples did not help to differentiate between the performances of the RDA and RER classifiers. However, the fact that all validation samples were correctly classified is an impressive result considering that the new gels were run at a much later time point and run by a different experimentalist. It should also be noted that the classifiers were trained on a data set that did not include a dye swap (Table 2) and tested on a 6 sample set that did include a dye swap (Table 3) RER invariance to monotonic shifts predicted that it would perform well when a dye swap was implemented and the results support the hypothesis. Also, even though the training set had many technical replicates, the trained classifiers were able to perform well on the validation data with only biological replicates, which indicates that they were not over fit to the training data.

While the RER method failed to outperform all algorithms, it is important to note that the use of multiple independently developed and validated algorithms increased confidence in the results. The analysis provided in this study indicates that RER methods have all the same advantages in the proteomic arena as they did in the genomic arena [7, 17, 18]. RER is advantageous for three main reasons. Firstly, because it is based on RERs, it accounts for monotonic expression variability eliminating the dependence on the method of normalization and use of a dye swap. Secondly, it achieves high classification rates that appear to generalize well.

Thirdly, the classification rules are easily interpretable, which is essential for biomarker discovery. Thus, RER methods are an important option to the EDA package of algorithms for discovering robust biomarkers in 2-D DIGE data. Lastly, implementation of all these algorithms needs to account technical replicates when present as demonstrated in this paper.

This study focused on the analysis of a proof of principle cohort of model cell states. The same technique can be applied to any set of proteomic samples in order to find the discriminatory protein pairs. While this study focused on biomarker discovery, the potential also exists to use these algorithms for pathway information and discovery of underlying molecular mechanisms. Furthermore, this method can be used to focus research of a cohort to a small amount of robust discriminatory protein pairs, facilitating follow-up studies.

This work was supported by NIH RO1-HL72488, National Heart, Lung, and Blood Institute (contract N01-HV-28180), CA51497 and CA57341. Chi V. Dang is Johns Hopkins Family Professor in Cancer Research. Special thanks to Dr. Aik Choon Tan for his careful review and editing of the manuscript and to Dawn Chen for her MS help.

5 References

[1] Voss, T., Haberl, P., *Electrophoresis* 2000, *21*, 3345–3350.

[2] Unlu, M., Morgan, M. E., Minden, J. S., *Electrophoresis* 1997, *18*, 2071–2077.

[3] Alban, A., David, S. O., Bjorkesten, L., Andersson, C. *et al.*, *Proteomics* 2003, *3*, 36–44.

[4] Karp, N. A., Kreil, D. P., Lilley, K. S., *Proteomics* 2004, *4*, 1421–1432.

[5] Karp, N. A., Lilley, K. S., *Proteomics* 2005, *5*, 3105–3115.

[6] Meleth, S., Deshane, J., Kim, H., *BMC Biotechnol.* 2005, *5*, 7.

[7] Geman, D., dÁvignon, C., Naiman, D. Q., Winslow, R. L., *Stat. Appl. Genet. Mol. Biol.* 2004, *3*, EPub 16646797.

[8] Friedman, D. B., Hill, S., Keller, J. W., Merchant, N. B. *et al.*, *Proteomics* 2004, *4*, 793–811.

[9] Seike, M., Kondo, T., Fujii, K., Yamada, T. *et al.*, *Proteomics* 2004, *4*, 2776–2788.

[10] Somiari, R. I., Sullivan, A., Russell, S., Somiari, S. *et al.*, *Proteomics* 2003, *3*, 1863–1873.

[11] Alfonso, P., Nunez, A., Madoz-Gurpide, J., Lombardia, L. *et al.*, *Proteomics* 2005, *5*, 2602–2611.

[12] Karp, N. A., Spencer, M., Lindsay, H., O'Dell, K., Lilley, K. S., *J. Proteome Res.* 2005, *4*, 1867–1871.

[13] Sebastiani, P., Gussoni, E., Kohane, I. S., Ramoni, M. F., *Stat. Sci.* 2003, *18*, 33–70.

[14] Speed, T. (Ed.), *Statistical Analysis of Gene Expression Microarray Data*, Chapman and Hall/CRC, Boca Raton 2003.

[15] Simon, R., Radmacher, M. D., Dobbin, K., McShane, L. M., *J. Natl. Cancer Inst.* 2003, *95*, 14–18.

[16] West, M., Blanchette, C., Dressman, H., Huang, E. *et al.*, *Proc Natl Acad Sci U S A* 2001, *98*, 11462–11467.

[17] Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L., Geman, D., *Bioinformatics* 2005, *21*, 3896–3904.

[18] Xu, L., Tan, A. C., Naiman, D. Q., Geman, D., Winslow, R. L., *Bioinformatics* 2005, *21*, 3905–3911.

[19] Schuhmacher, M., Staege, M. S., Pajic, A., Polack, A. *et al.*, *Curr. Biol.* 1999, *9*, 1255–1258.

[20] Decyder-2D Differential Analysis Software, Version 5.0 User Manual, GE Healthcare (Amersham Biosciences) 2003.

[21] Guo, Y., Hastie, T., Tibshirani, R., *Biostatistics* 2006.

[22] Witten, I. H., Frank, E., *Data Mining: Practical machine learning tools and techniques with java implementations.*, Morgan Kaufmann Publishers, USA 2000.

[23] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning*, Springer-Verlag, New York 2001.

[24] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[25] Zou, L., Wu, Y., Pei, L., Zhong, D. *et al.*, *Leuk. Res.* 2005, *29*, 1387–1391.

[26] Bae, S. M., Lee, C. H., Cho, Y. L., Nam, K. H. *et al.*, *Gynecol. Oncol.* 2005, *99*, 26–35.

[27] Shiio, Y., Donohoe, S., Yi, E. C., Goodlett, D. R. *et al.*, *EMBO J.* 2002, *21*, 5088–5096.

[28] Kim, J. W., Zeller, K. I., Wang, Y., Jegga, A. G. *et al.*, *Mol. Cell. Biol.* 2004, *24*, 5923–5936.

[29] Takashima, M., Kuramitsu, Y., Yokoyama, Y., Iizuka, N. *et al.*, *Proteomics* 2005, *5*, 1686–1692.