Prediction of biologically significant components from microarray data: Independently Consistent Expression Discriminator (ICED)

Rahul Bijlani [1,†], *Yinhe Cheng* [1,†], *David A. Pearce* [2, 4, 5], *Andrew I. Brooks* [2, 3,*] *and Mitsunori Ogihara* [1, 2]

[1]*Department of Computer Science,* [2]*Center for Functional Genomics,* [3]*Department of Environmental Medicine,* [4]*Center for Aging and Developmental Biology, and* [5]*Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA*

ABSTRACT

Motivation: Class distinction is a supervised learning approach that has been successfully employed in the analysis of high-throughput gene expression data. Identification of a set of genes that predicts differential biological states allows for the development of basic and clinical scientific approaches to the diagnosis of disease. The Independent Consistent Expression Discriminator (ICED) was designed to provide a more biologically relevant search criterion during predictor selection by embracing the inherent variability of gene expression in any biological state. The four components of ICED include (i) normalization of raw data; (ii) assignment of weights to genes from both classes; (iii) counting of votes to determine optimal number of predictor genes for class distinction; (iv) calculation of prediction strengths for classification results. The search criteria employed by ICED is designed to identify not only genes that are consistently expressed at one level in one class and at a consistently different level in another class but identify genes that are variable in one class and consistent in another. The result is a novel approach to accurately select biologically relevant predictors of differential disease states from a small number of microarray samples.

Results: The data described herein utilized ICED to analyze the large AML/ALL training and test data set (Golub *et al.*, 1999, *Science*, **286**, 531–537) in addition to a smaller data set consisting of an animal model of the childhood neurodegenerative disorder, Batten disease, generated for this study. Both of the analyses presented herein have correctly predicted biologically relevant perturbations that can be used for disease classification, irrespective of sample size. Furthermore, the results have provided candidate proteins for future study in understanding the disease process and the identification of potential targets for therapeutic intervention.

Contact: andrew_brooks@urmc.rochester.edu

INTRODUCTION

Microarray technologies including high-density oligonucleotide and cDNA arrays make it possible to monitor the mRNA levels of thousands of genes in a single experiment. Data generated by these types of experiments has been used for disease classification and class prediction (Golub *et al.*, 1999), drug target identification (Kozian and Kirschbaum, 1999), and development and validation of biological pathways (Gray *et al.*, 1998; Marton *et al.*, 1998). However, no universally accepted methodology for the analysis of such large and complex data sets exists. Commonly used techniques include clustering methods (Eisen *et al.*, 1998; Alon *et al.*, 1999; Perou *et al.*, 1999; Ben-Dor *et al.*, 2000), Support vector machines (SVMs; (Furey *et al.*, 2000; Brown *et al.*, 2000)), classification trees (Dubitzky *et al.*), genetic algorithms (Li *et al.*, 2001; Moore and Parker, 2001), neural networks (Hwang *et al.*, 2001), and a weighted correlation method called Neighborhood Analysis (Golub *et al.*, 1999).

We present an algorithm in which samples of microarray data divided into two classes, recognizes genes that are good class discriminators and uses them for the identification of unknown samples. Its novelty lies in its ability to effectively bypass two assumptions which are not addressed in some of the classical methods:

(a) *The distribution of the gene intensities in a sample is normal.* For example, in their analysis of the well-known leukemia data set (Golub *et al.*, 1999) uses the Pearson Correlation Value to calculate importance of a gene in distinction between two sample classes. This

---

value is a fraction, whose denominator is the sum of the standard deviation of the gene's expression in the two classes and whose numerator is the absolute difference of the average of the gene's expression between the two classes. In the case when one distribution is normal and the other is bipolar but they have the same mean, the Pearson Correlation Value is zero while the two distributions look quite different from each other. It is known that the distributions of some replicates are apparently uniform while others are bimodal or trimodal, with further irregularity introduced by the inclusion of absent calls calculated by Affymetrix algorithms (Grant *et al.*, 2001).

(b) *A gene is a good discriminator if it is present at a consistently high level in one class, and absent or present at a consistently low level in the other class.* This appears to be an acceptable but incomplete search strategy when attempting to identify biologically important discriminators.

In the Independently Consistent Expression Discriminator (ICED) the first assumption is weakened by allowing the distributions of relevant gene expression to be multi-polar in one of the two sample classes. The second assumption is addressed by broadening the search criterion—ICED searches for genes that are consistent in one class of data, but not consistent at the same level in the other class. Inherent genetic variation and environmental influence may lead to differential gene expression at baseline in any given population. The perturbation of a pathway(s) leads to changes in gene expression that results in a similar pattern exhibited by all subjects affected by the change in biological state. ICED is designed to identify all differentially expressed genes as a function of potential genetic and environmental variability. Recently, Califano *et al.* has presented an approach that selects genes which are consistent in one condition and variable in the other; the approach goes on to use those genes in a classification scheme (Califano , 2000). This approach has significant differences from ICED, a major one being that in Califano's approach statistical dependence between genes are taken into account indirectly, as opposed to the gene by gene approach employed by ICED. Other differences between the two approaches lie primarily in the scalability and sample sizes needed to achieve a robust classification scheme.

All data points are assigned a weight according to a formula, and a search criterion is used to decide the optimal number of genes with the top weight should be used in the classifier. A voting mechanism based on the genes selected and their weights to assign class membership along with a prediction strength confidence value. To test the efficiency of this method, ICED was run on a 72-sample leukemia data set (Golub *et al.*, 1999), as well as an 8-sample Batten disease study performed at

the University of Rochester Medical Center. To test the robustness of this analysis, a full leave-one-out cross or jack-knife evaluation (Efron, 1982) of the classification performance was performed with 100% accurate results for both data sets. The ICED algorithm has been compared to analysis of similar approaches including SVMs (Furey *et al.*, 2000) and Neighborhood Analysis (Golub *et al.*, 1999). In addition, the leukemia data set was resampled 100 times into a pair of 36-sample groups, where data in the first group was used to classify samples in the second. In this repeated resampling test, ICED made highly accurate predictions and consistently identified similar groups of genes as good discriminators. Interestingly, accurate results were obtained in data sets from both the leukemia and Batten disease studies, in spite of their significantly different sample sizes. Subsequent analysis of the results from both analyses has found the highest weighted genes to have strong biological relevance to the disease states being classified.

SYSTEM AND METHODS

Microarray methods

For Batten disease data set, total RNA was prepared from the cerebellum of cln3 knock-out (Batten mice) and WT littermates ($n = 4$; for each group) and gene expression studies were performed as described in Chattopadhyay *et al.* (2002) using Affymetrix Mu19K Genechips.

The AML/ALL data set used was generated at the Whitehead Institute and the Center for Genome Research at the Massachussets Institute of Technology and is available at their website (http://www-genome.wi.mit.edu/MPR/data_sets.html). The methods used for generating microarray data have been previously described (Golub *et al.*, 1999).
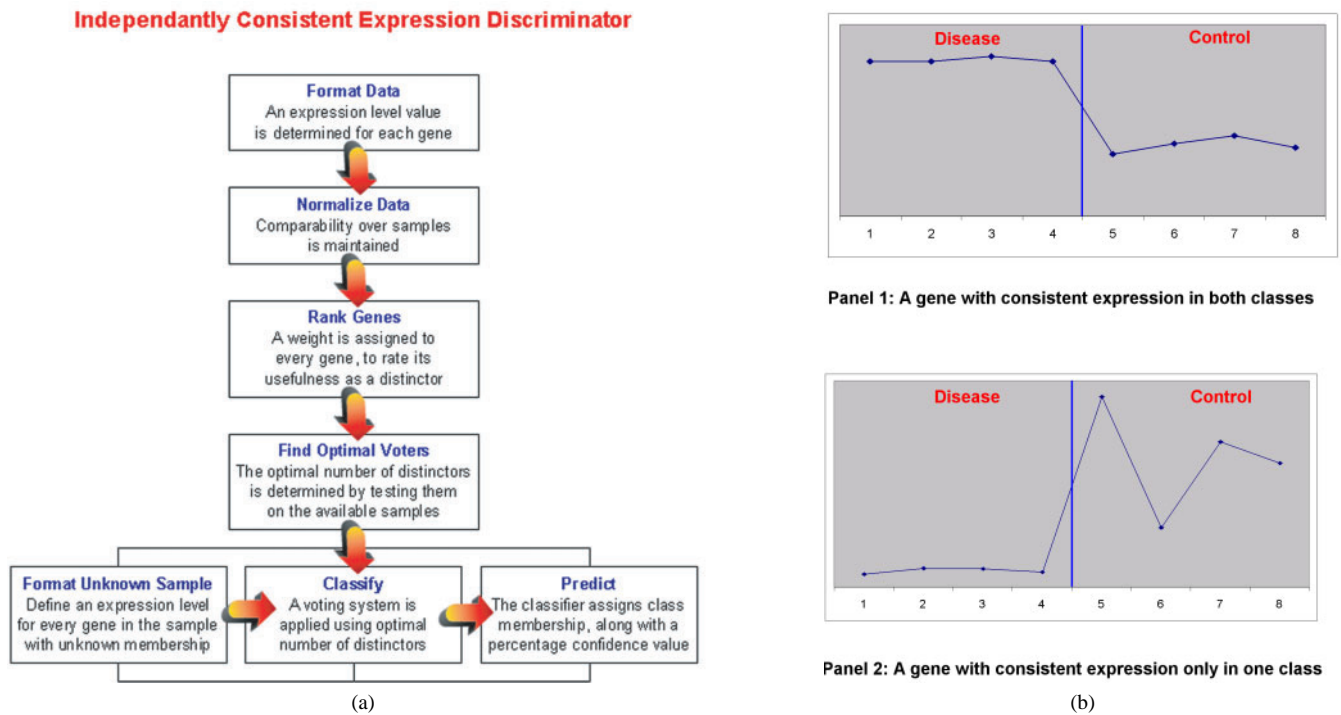
ICED Analysis

There are four steps involved in the execution of the ICED protocol (Figure 1a). They were performed in the following sequence and are described herein:

(1) *Normalization.* For comparability over different experimental standards and samples, normalization of microarray data is critical for analysis. In these experiments, gene expression levels within a sample were scaled to a mean of 0 and standard deviation of 1.

(2) *Gene weight generation.* Two typical measurements used in the computation of the relationship between gene expression patterns and class distinctions are Euclidean Distance and Pearson Correlation Coefficient.

(a) Euclidean Distance was used in (Furey *et al.*, 2000).

(b) A more sophisticated measure is the Pearson Correlation Coefficient, modified as follows (Golub *et al.*, 1999) to emphasize the 'signal-to-noise' ratio in using a gene as

Independantly Consistent Expression Discriminator

(a)

Panel 1: A gene with consistent expression in both classes

Panel 2: A gene with consistent expression only in one class

(b)

Fig. 1. An overview of ICED (Independently Consistent Expression Discriminator). (a) A schematic of the ICED algorithm: The figure provides a brief overview of the various processes that make up the algorithm; (b) The principle behind gene selection with ICED. Rather than restrict itself to genes with expression patterns similar to panel 1, that have widely differing means corresponding to class distinction, and demonstrate low variability within each class, ICED also finds genes with expression patterns like the one panel 2 that are significantly consistent in one class, and less consistent or even variable in the other.

a predictor. Let $[\mu_1(g), \sigma_1(g)]$ and $[\mu_2(g), \sigma_2(g)]$ denote the means and standard deviations of the log of the expression levels of gene g for the samples in class 1 and class 2, respectively.

$$P(g) = [\mu_1(g) - \mu_2(g)]/[\sigma_1(g) + \sigma_2(g)]$$

This value is used to reflect the difference between the classes relative to the standard deviation within the classes. Large values of $P(g)$ are meant to indicate a strong correlation between the gene expression and the class distinction, while the sign of $P(g)$ being positive or negative corresponds to g being more highly expressed in class 1 or class 2.

The reasoning behind this criterion seems to be that variability in the data may occur due to the presence of outliers, and gene expressions differing consistently as per class identity can function as well defined discriminators. Namely, genes that serve as good discriminators between two classes are expressed consistently at a certain level in one class and consistently at a different level in the other classes, and the more consistent the expressions are and the wider the gap between the expression levels is, the better the gene is for discrimination. This measure is used

by the Golub *et al.* team to select a number of predictor genes to achieve around 85% accuracy.

This study is motivated by the question of whether this measure can be improved to achieve much higher accuracy in prediction. In particular, it seeks to investigate whether the requirement that the two genes should have consistent expressions in each of the classes is too strict. There are two reasons for this hypothesis—first, ideal genes may be limited in numbers and hard to find. With a limited number of good discriminators, the quality of prediction may suffer. Second, it is conceivable that a particular disease state has the effect of bringing a gene's otherwise variable expression level up or down to a certain uniform value.

A weight-assignment formula was developed to identify genes that showed consistent expression levels in one class, and did not show consistent expression levels in the same numerical range in the other class. This definition intrinsically includes genes that fit the idealized scenario described earlier, but also identifies genes that may be good discriminators but would be excluded from the previous analysis.

For every gene g, two values, $W_1(g)$ and $W_2(g)$, were computed from the known samples, which are its weights

as a discriminator for class 1 and class 2 respectively:

$$W_1(g) = \frac{\frac{1}{m} \sum\limits_{i=1,m} |g_{2,i} - \mu_{1,n}(g)|}{\sigma_{1,n}(g)};$$

$$W_2(g) = \frac{\frac{1}{n} \sum\limits_{j=1,n} |g_{1,j} - \mu_{2,m}(g)|}{\sigma_{2,m}(g)}$$

where

$g_{1,i}$ is the expression level of the ith sample of gene g in class 1;

$g_{2,j}$ is the expression level of the jth sample of gene g in class 2;

$\mu_{1,n}(g)$ and $\sigma_{1,n}(g)$ are the mean and standard deviation of the n samples of gene g in class 1;

$\mu_{2,m}(g)$ and $\sigma_{2,m}(g)$ are the mean and standard deviation of the m samples of gene g in class 2.

Instead of using the sum of the standard deviations in both classes to calculate the denominators, the focus is on the deviation in one class. Furthermore, instead of using the Euclidean Distance between two genes or the difference in the class means to compute the numerator, the sum of the absolute distance between every sample in one class and the mean of the other class is calculated. A high value for $W_x(g)$ implies that the gene expresses itself consistently in class x, and not consistently at a similar value in the opposite class. Genes recognized usually have either relatively high $W_1(g)$ or $W_2(g)$ values, less often both, validating our design goals. The biggest difference between our weight and the Pearson Correlation Value is that, for W_1, averaging is taken over the absolute value of the distance of $g_{2,j}$ from μ_1. Suppose that the averaging is not over the absolute value, as in the Pearson Correlation Value, but over the simple difference between the two values, that g_1 is subject to a normal distribution, and that g_2 is subject to a symmetric bipolar distribution having the same average as g_1. In our measure the evaluation is some positive value while in the alternative definition the measure is 0. To achieve our goal of identifying genes that are expressed consistently at one level in one class but not consistently expressed at the same level in the other class, it is thus crucial that averaging is over the distance of $g_{2,j}$ from μ_1.

Sorting genes by their $W_1(g)$ and $W_2(g)$ values can be used to rank their discriminating abilities and investigate biological relationships between highly ranked genes and the state represented by their respective classes.

(3) *Voting methodology.* To analyze an unknown sample x, we compute a pair of votes for every gene g in the data set using the following formulas as proposed by Golub *et al.*:

$$V_1(g) = W_2(g) \bullet |g_x - \mu_{2TR,m}(g)|$$
$$V_2(g) = W_1(g) \bullet |g_x - \mu_{1TR,n}(g)|$$

where

g_x is the expression level of gene g in the unknown sample;

$\mu_{1TR,n}(g)$ is the mean of the n training samples of gene g in class 1;

$\mu_{2TR,m}(g)$ is the mean of the m training samples of gene g in class 2.

Finally, a prediction strength $P(x)$ is determined for unknown sample x, using the sum of the votes of p top genes in class 1 and q top genes in class 2, to generate a value within the range $[-1, 1]$:

$$P(x) = \frac{q \bullet \sum\limits_{i=1,p} V_1(g_i) - p \bullet \sum\limits_{i=1,q} V_2(g_i)}{q \bullet \sum\limits_{i=1,p} V_1(g_i) + p \bullet \sum\limits_{i=1,q} V_2(g_i)}$$

A positive value of $P(x)$ denotes that x is a member of class 1, and vice versa for a negative value. This absolute value of $P(x)$ reflects the confidence in the prediction. Large values of $P(x)$ are meant to indicate a high prediction strength.

(4) *Vote-based classifier.* The classifier is designed to answer the question—'How many genes are necessary for an accurate discriminant analysis?', i.e. to determine the value of p and q highest weighted genes in the prediction strength equation.

If P indicates the number of the potential gene predictors in class 1, and Q indicates the number of the potential gene predictors in class 2, the task of the classifier is to find the value of p and q that maximize *fitness*(p, q). p is in the range of $[1, P]$ and q is in the range of $[1, Q]$. *fitness*(p, q) reflects the discriminating ability of the predictor using p highest weighted genes in class 1 and q highest weighted genes in class 2. It can be computed in a number of different ways, as per the individual researcher's requirement. The result derived from using a particular number of predictor genes can be evaluated by a number which is either the average, median or minimum value of the prediction strengths made by the voting mechanism. The classifier's objective would be to find the number of top ranking genes from either class that maximizes this value.

There is a precedent in Golub *et al.* for using the median prediction strength of classifier results to characterize the performance of a classifier. Our system allows the user to select between maximizing the average, minimum or median prediction strength as per her assumptions of the biological model. In our experiments, we first made an estimate of the optimal number of genes required by the classifier after observing the distribution of values in the ranking of weights. After using this estimate to make predictions with ICED, we observed the distribution of the prediction strengths of the classified samples to select a fitness measure.

In these initial estimates for the ALL/AML data set, there was a relative variability in prediction strengths. It did not seem wise to use the median since a certain percentage of the population seemed to lie at lower prediction strengths. Maximizing the minimum recorded prediction strength in the fitness function did not seem practical either, since Golub *et al.* and our earlier experiments had consistently found some outliers in the data. Hence, maximizing the average prediction strength seemed like the best prospect for a fitness function, and ICED's results achieved a high enough degree of accuracy to attest to this choice.

We estimated the optimal number of genes required to analyze the Batten disease data set and discovered that the prediction strengths for all samples in a leave-one-cross were similar, suggesting the use of maximal median prediction strength as a selection criterion. Again, ICED achieved a high degree of accuracy in its predictions.

Ideally, given a training data set, the classifier should test out every possible value of p and q to find the optimal number of genes that need to be considered in each class. This form of two-dimensional analysis is, however, very time-intensive (runtime of the order $O(P*Q*$analysis-time)). One compromise could be achieved by stepping through values of p and q determined by threshold values on weights in the respective classes. i.e.

p = number of genes with $W_1(g_i)$ above the threshold value th_1 for class 1

q = number of genes with $W_2(g_i)$ above the threshold value th_2 for class 2

A user specified increment value would thus limit the running time of such a search.

We also designed a novel algorithm that determines optimal thresholds in significantly less time, with demonstrably high accuracy.

(1) Initially, step through the training set using the same thresholds for both classes, with user specified increments, and test the same data set for accuracy. Let the threshold yielding the best results be t_1.

(2) Fix the threshold for class 1 at t_1 and repeat the search while only implementing the increment steps in threshold values for class 2, testing after each weight list is generated. Let the threshold value for class 2 that generates the best result in combination with t_1 be saved as t_2.

(3) Now fix the threshold for class 2 at t_2 and step through incrementing threshold values for class 1, finally logging the value that yields the best result in the variable t_1.

(4) Repeat 2 and 3 until a user specified number of iterations is completed or the values of t_1 and t_2 remain constant, whichever occurs first.

Experimental evidence suggests that values for t_1 and t_2 stabilize in only 1 or 2 pairs of iterations. So the runtime is reduced to $O(c*(P + Q)*$analysis-time), where c is the average number of iterations. As the genes are sorted in the decreasing order of the weights, we expect that our incremental search is likely to find a local optimum that can achieve accuracy close to the accuracy achieved by the global optimum.

In the scheme described above, a proper increment step can be selected to fit the user's requirement of the runtime and preciseness.

IMPLEMENTATION AND DISCUSSION

Molecular diagnostics is a growing discipline that has the potential to impact both preventative medicine and the treatment of established disease. ICED analysis of the AML/ALL data set resulted in an accurate classification of the samples, as well as the identification of biologically relevant discriminator genes. The 38 sample (27 ALL, 11 AML) training data set used by the Golub *et al.* study was used to develop a weighting/voting system that correctly identified all 34 test samples, 33 of which were with strong predictions (i.e. with a confidence level higher than 0.2). The median confidence level of ICED predictions was 0.911 (Figure 2b). In contrast, Neighborhood Analysis correctly identified 29 samples correctly with strong predictions, and misidentified 2 of the other 5 samples that had weak predictions. The median prediction strength was 0.73. SVMs or Support Vector Machines, are used by Furey *et al.* (2000) to classify examples in the test set, producing results ranging from 30 to 32 accurate predictions from the 34 sample data set. In all their tests, the SVM correctly classifies the 29 predicted by Golub *et al.* and for the five unpredicted samples, each is misclassified in at least one SVM test. Two samples are misclassified in all SVM tests, and no prediction strengths are available for the analyses.

A leave-one-out cross or jackknife testing of the 38 training samples by ICED resulted in 100% accuracy; all 38 samples were identified correctly with strong predictions, and demonstrated a median confidence level of 0.897 (Figure 2a). An SVM based approach also correctly identified all 38 training samples. On the other hand, Neighborhood Analysis identified 36 of the samples accurately with strong predictions, with a median prediction strength of 0.77, and one of the two weak predictions was inaccurate. In addition, Table A (http://fgc.urmc.rochester.edu/resource.html) depicts genes that ICED consistently allotted a high weighting to over 100 resampled analyses of the Golub *et al.* 72 sample data set.

(a)

(b)

(c)
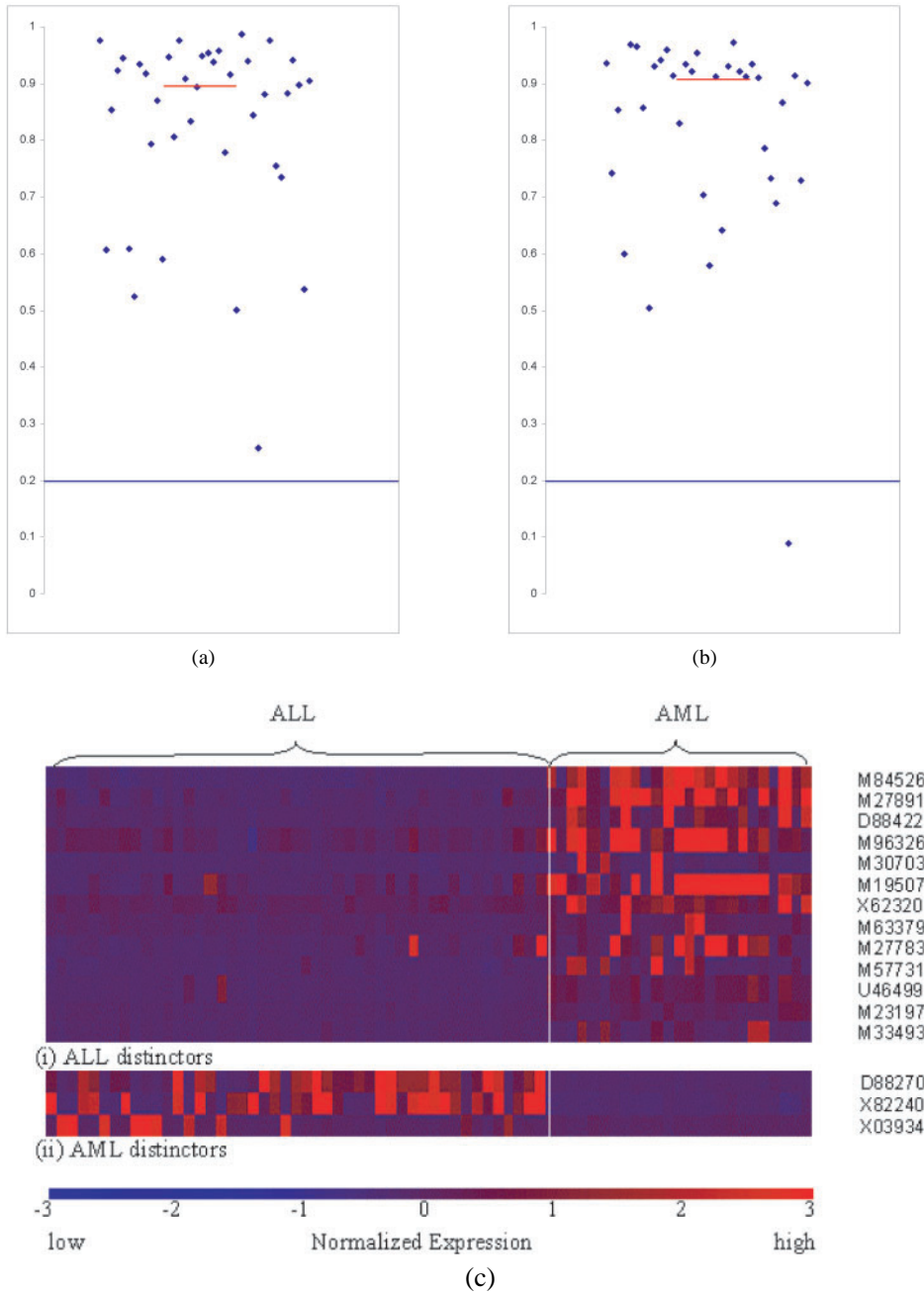
Fig. 2. ICED analyses of 47 ALL and 25 AML samples, each containing 7129 genes, from the AML/ALL data set. (a) Scatter plot of confidence levels in ALL/AML, class predictions for leave-one-out cross or jackknife testing of 38 training samples (27 ALL, 11 AML): A class prediction (ALL/AML), with a percentage confidence level, is made for every sample in the training data set after using the other 37 samples to create a weighting/voting system. All 38 samples were identified correctly with strong predictions, with a median confidence level of 89.7%, as shown by the red line. The blue line at 20% indicates the threshold between strong and weak predictions. This boundary depends on the number of genes in the classifier as proposed by Golub *et al.*; (b) Scatter plot of confidence levels in ALL/AML class predictions for 34 test samples (20 ALL, 14 AML); ICED used the 38 sample training data set to develop a weighting/voting system that correctly identified all 34 samples, 33 of which were with strong predictions, and the median confience level was 91.1%, as shown by the red line. The blue line at 20% indicates the threshold between strong and weak predictions; (c) The optimal genes selected as ALL/AML distinctors based upon their expression levels over all 72 samples: The columns correspond to the samples, and the colors are graded from low (blue) to high (red) gene expression levels, as determined after normalizing the data. *Top Table:* Distinctor genes that are consistent in ALL samples. *Bottom Table:* Distinctor genes that are consistent in AML samples.

(a)

(b)
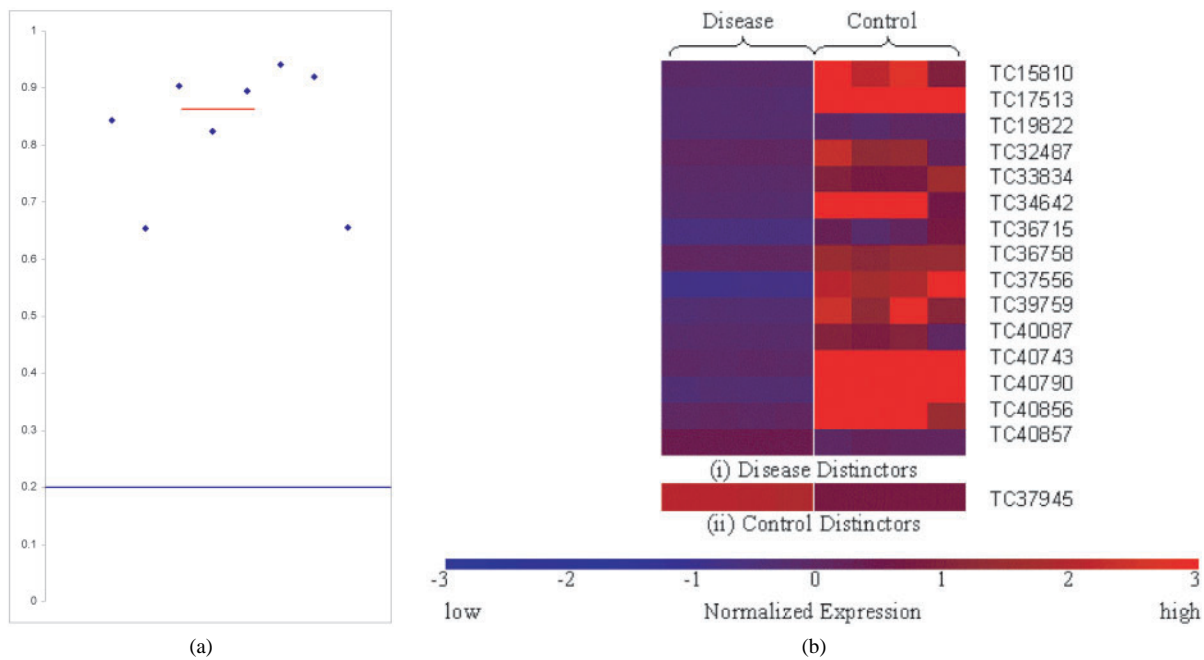
Fig. 3. ICED analyses of 4 diseased and 4 control samples, each containing 21 146 genes from the Batten disease data set. (a) Scatter plot of confidence levels in diseased/control class predictions for leave-one-out cross or jackknife testing: A class prediction (diseased/control), with a percentage confidence level, is made for every sample in the data set after using the other 7 samples to create a weighting/voting system. All 8 samples were identified correctly with strong predictions, with a median confidence level of 87%, as shown by the red line. The blue line at 20% indicates the threshold between strong and weak predictions; (b) The optimal genes selected as Batten's disease distinctors based upon their expression levels over 8 samples: The columns correspond to the samples, and the colors are graded from low (blue) to high (red) gene expression levels, as determined after normalizing the data. *Top Table:* Distinctor genes that are consistent in diseased samples. *Bottom Table:* Distinctor genes that are consistent in control samples.

In order to evaluate the robustness of ICED, for 100 runs the 72 samples were randomly divided into 36 sample pairs of training and test data sets, the first of which was used to train ICED and the second to test it. 98.7% of the 3600 predictions made were accurate, 95.6% of which were strong predictions. Table A ranks genes that were consistently weighted among the top 10, top 20, top 50 and top 100 highest weights in the lists generated by ICED from the training samples in the 100 resampled analyses. Each cell in Table A represents the percentage of occurrences of that particular gene in the respective ranking.

A comparison of the resultant genes selected as predictors using the Nearest Neighborhood analysis and ICED are interesting from a statistical and biological perspective. This point is illustrated in Table B (http://www. urmc.rochester.edu/research/FGC/resource.html) where the optimal number of genes selected by ICED for the AML/ALL data set is directly compared to the output of the NNA. Although some gene function similarities are observed the highest ranked genes (or strongest predictors of class distinction) exhibit differences based on the

search criteria of the two approaches. One example of this difference is the highest ranked gene from the AML/ALL data set, Zyxin. Although there is no clear biological link between leukemia and this gene product, it is considered an excellent class predictor. By contrast, two of the highest ranked ICED genes in the AML/ALL data set are Cystatins, known cystein proteases responsible for protein folding with clear implications in cancer biology (Finney *et al.*, 2001; Yano *et al.*, 2001; Foghsgaard *et al.*, 2001; Kos *et al.*, 2000; Stabuc *et al.*, 2000).

Determining the molecular basis of disease etiology and progression is another application for applying a classification approach to microarray data sets. This approach can be used not only to identify gene based and pharmacological targets for disease treatment, but monitor the progression of established disease and/or measure the efficiency of a therapeutic intervention. Batten disease is inherited in an autosomal recessive manner and is the most common progressive neurodegenerative disease of childhood. The disorder is characterized initially by visual deterioration at age 5–7 which ultimately results in blindness, followed

Table 3. A functional classification of genes ICED optimally selected for class distinction in an animal model of Batten's disease

TIGR	Identity	Functional class
TC34642	TRAM-protein	Trafficking
TC17513	Death Associated Protein (DAPI)	Cell death
TC40790	Myosin Light Chain	Cytoskeleton
TC40743	Transketolase	Pentose-P pathway
TC36758	Ribonuclease T2	RNA degradation
TC36715	CRIP protein	Immune response
TC32487	Neurofilament B	Neuron structure
TC15810	Oxoacyl coA thiolase	Lipid modification
TC37556	p53	Cell death
TC39759	TB2 like protein	Immune response
TC33834	Unknown	
TC19822	Unknown	
TC40087	Unknown	
TC40857	Probable glycosyltransferase	Protein modification
TC38959	Acyl coA desaturase	Lipid modification
TC37945	Cytochrome C1	Oxidative phosphorylation

The optimal number of diseased discriminator genes for the Batten data set with a biological classification relating the importance of function as a result of selection criteria.

by an increased frequency of untreatable seizures, mental retardation, loss of motor skills and premature death. The CLN3 gene responsible for Batten's disease was positionally cloned in 1995 (International Batten disease Consortium, 1995), with most individuals affected harboring a 1.02 kb deletion of the gene. One of the paradoxes of Batten disease is that it is characterized by the accumulation of autofluorescent hydrophobic material in the lysosome of neurons and other cell types with the cerebellum being greatly affected. However, the accumulation of this lysosomal storage material, which no doubt contributes to the neurologic disease, does not apparently lead to disease in these other cell types making this observation a poor choice for clinical diagnosis. A predominant component of the lysosomal storage material has been identified as mitochondrial ATP synthase subunit *c* (Palmer *et al.*, 1992, 1995; Kominami *et al.*, 1992; Ezaki *et al.*, 1996). However, how these cellular alterations relate to the neurodegeneration in NCL's is unknown.

We have compared gene expression in the cerebellum of 10-week old cln3-knockout mouse model for the neurodegenerative disorder, Batten disease (Mitchison *et al.*, 1999), as compared to normal mice, of approximately 19 000 transcripts by high-density oligonucleotide arrays (Chattopadhyay *et al.*, 2002). To minimize technical and surgical variation, cerebella were collected from three male cln3-knockout and three male normal mice, and each type pooled for extraction of RNA. We have recently shown that surgical resection of individual sub-structures, or pieces thereof, contribute significantly to the variability of the assay irrespective of genetic and biological

variability (Brooks *et al.*, unpublished observation). To this end, we have minimized the experimentalist induced variation by pooling the cerebella of three genetically identical animals. Total sample size equaled a biological replicate of four samples for each group. The resultant probes derived from the RNA were hybridized to Affymetrix high-density Mu19K sub arrays A, B and C. Reproducible changes in expression of two-fold or more (determined by averaging the fold change values of all 16 possible pairwise comparisons) were found for 756 genes by performing the comparative analysis using the Affymetrix algorithms. We have classified those genes that have an altered expression pattern into 14 functional categories based on what is known in the public domain about the biology of each gene product. Functional analysis revealed gene expression changes in the cln3-knockout cerebellum as compared to normal for genes involved in neuronal cell structure and development, immune and inflammatory response, and lipid metabolism (Brooks *et al.*, unpublished).

ICED analysis of this data set corroborates the functional analysis described above. The weighted results for Batten disease versus control from the ICED analysis, illustrated in Table 1, provide an interesting correlation with what we know about the pathogenesis of the disease, as well as some interesting new insights. For example, as a neurodegenerative disease, atrophy of the brain occurs in Batten disease, and it is therefore not surprising that two proteins associated to cell death, DAP1 and p53, are heavily weighted as being important predictors of a disease state. Similarly, Batten disease is characterized by the accumulation of ceroid deposits in neurons, and one would predict altered lipid metabolism, which is borne out by the weighting of Oxoacyl CoA thiolase and acyl CoA desaturase. In addition, up-regulation of inflammatory proteins is often associated with neurodegenerative disease. The high weighting of the CRIP protein and a TB2-like protein is therefore intriguing, suggesting perhaps that a novel inflammatory response of an immunological nature may be occurring in this mouse model for Batten disease.

In summary, the ICED algorithm has predicted proteins already known to be associated with the disease processes in addition to providing new insight to disease etiology and progression by the selection of novel gene products. These novel gene products upon subsequent study may prove to be valuable in ultimately understanding the mechanism of disease. We have also demonstrated that this approach can be applied successfully to both large and small data sets as demonstrated by the cancer and neurodegenerative disease experiments described herein. We conclude that ICED is a powerful tool that can be utilized to focus microarray data into identification of key proteins that require further investigation.

ACKNOWLEDGEMENTS

REFERENCES

Alon,U., Barkai,N. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Ben-Dor,A., Bruhn,L. *et al.* (2000) Tissue classification with gene expression profiles. *J. Comput Biol*, **7**, 559–583.

Brown,M.P., Grundy,W.N. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Califano,Stolovitzky, Tu (2000) Analysis of gene expression microarrays for phenotype classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 75–85.

Chattopadhyay,S., Ito,M., Cooper,J.D., Brooks,A.I., Curran,T.M., Powers,J.M. and Pearce,D.A. (2002) An autoantibody inhibitory to glutamic acid decarboxylase in the neurodegeneartive disorder Batten disease. *Hum. Mol. Genet.*, **11**, 1421–1431.

Efron,B. (1982) *The Jackknife, The Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Eisen,M.B., Spellman,P.T. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Ezaki,J., Wolfe,L.S. and Kominami,E. (1996) Specific delay in the degradation of mitochondrial ATP Synthase subunit c in late infantile neuronal ceroid lipofuscinosis is derived from cellular proteolytic dysfunction rather than structural alteration of subunit c. *J. Neurochem*, **67**, 1677–1687.

Finney,H., Williams,A.H. *et al.* (2001) Serum cystatin C in patients with myeloma. *Clin. Chim. Acta*, **309**, 1–6.

Foghsgaard,L., Wissing,D. *et al.* (2001) Cathepsin B acts as a dominant execution protease in tumor cell apoptosis induced by tumor necrosis factor. *J. Cell Biol.*, **153**, 999–1010.

Furey,T.S., Cristianini,N. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Golub,T.R., Slonim,D.K. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Grant,G. *et al.* (2001) Using non-parametric methods in the context of multiple testing to identify differentially expressed genes. CAMDA00 proceedings (publication due to appear in 2001).

Gray,N.S., Wodicka,L., Thunnissen,A.M., Norman,T.C., Kwon,S., Espinoza,F.H., Morgan,D.O., Barnes,G., LeClerc,S., Meijer,L. *et al.* (1998) Exploiting Chemical Libraries, Structure, and Genomics in the Search forKinase Inhibitors. *Science*, **281**, 533–538.

Hwang,K.-B., Cho,D.Y., Park,S.-W., Kim,S.D. and Zhang,B.T. (2001) Applying machine learning techniques to analysis of gene expression data: cancer diagnosis. In Liu,S.M. and Johnson,K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer Academic Publishers.

Jason,H.M. and Joel,S.P. (2002) Evolutionary computation in microarray data analysis. In Liu,S.M. and Johnson,K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer Academic Publishers.

Kominami,E., Ezaki,J., Muno,D., Ishidoh,K., Ueno,T. and Wolfe,L.S. (1992) Specific storage of subunit c of mitochondrial ATP synthase in lysosomes of neuronal cedroid lipofuscinosis (Batten's disease). *J. Biochem*, **111**, 278–282.

Kos,J., Krasovec,M. *et al.* (2000) Cysteine proteinase inhibitors stefin A, stefin B, and cystatin C in sera from patients with colorectal cancer: relation to prognosis. *Clin. Cancer Res.*, **6**, 505–511.

Kozian,D.H. and Kirschbaum,B.J. (1999) Comparative gene-expression analysis. *Trends Biotechnol*, **17**, 73–78.

Li,L.-P., Pedersen,G., Darden,T.A. and Weinberg,C.R. (2001) Computational analysis of leukemia microarray expression data using the GA/KNN. In Liu,S.M. and Johnson,K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer Academic Publishers.

Marton,M.J., DeRisi,J.L. *et al.* (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.*, **4**, 1293–301.

Mitchison,H.M. *et al.* (1999) Targeted disruption of the Cln3 gene provides a mouse model for Batten's disease. *Neurobiol. Dis.*, **6**, 321–334.

Moore,J.H. and Parker,J.S. (2001) Evolutionary computation in microarray data analysis. In Liu,S.M. and Johnson,K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer Academic Publishers, ISBN 0-7923-7564-5.

Palmer,D.N., Fearnley,I.M., Walker,J.E., Hall,N.A., Lake,B.D., Wolfe,L.S., Haltia,M., Martinus,R.D. and Jolly,R.D. (1992) Mitochondrial ATP synthase subunit c storage in the ceroid-lipofuscinoses (Batten Disease). *Am. J. Med. Genet.*, **42**, 561–567.

Palmer,D.N., Bayliss,S.L. and Westlake,V.J. (1995) Batten disease and the mitochondrial ATP synthase subunit c turnover pathway: raising antibodies to subunit c. *Am. J. Med. Genet.*, **57**, 260–265.

Perou,C.M., Jeffrey,S.S. *et al.* (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.

Stabuc,B., Vrhovec,L. *et al.* (2000) Improved prediction of decreased creatinine clearance by serum cystatin C: use in cancer patients before and during chemotherapy. *Clin. Chem.*, **46**, 193–197.

Yano,M., Hirai,K. *et al.* (2001) Expression of cathepsin B and cystatin C in human breast cancer. *Surg. Today*, **31**, 385–389.