Conf-9107133--1

ANL/CP--73117

DE91 014015

DISCLAIMER

QUANTITATIVE MUTATION DETECTION USING TWO-DIMENSIONAL ELECTROPHORESIS*

J Taylor and CS Giometti
Biological and Medical Research Division
Argonne National Laboratory
Argonne, IL 60439-4833, USA

---

*This work is supported by the U.S. Department of Energy, Office of Health and Environmental Research under Contract No. W-31-109-ENG-38.

QUANTITATIVE MUTATION DETECTION USING TWO-DIMENSIONAL ELECTROPHORESIS

J Taylor and CS Giometti

Biological and Medical Research Division Argonne National Laboratory, Argonne, IL 60439-4833, USA

1 INTRODUCTION

Two-dimensional electrophoresis (2DE) of proteins has been used in several studies to detect genetic mutations [1-4]. Mutations are generally observable in 2DE patterns as spots which appear out of context, either in position or amount (quantitative variants). Most studies to this point have screened only for positional variants, as these are detectable by eye and do not require computer analysis. Screening for quantitative variants in addition to the positional ones widens the class of detectable events. We have been using 2DE to detect heritable mutations expressed as quantitative alterations in liver protein expression in inbred mice with the goal of extending these studies to analysis of mutation expression in human populations [1]. The purpose of this paper is to describe the computerized screening techniques which have been developed and to summarize their performance.

2 METHODS

The data used in these studies came from the Argonne National Laboratory 2DE database of mouse liver proteins. All data used in this paper came from data sets N89 and M20. N89 consists of 749 distinct patterns from offspring of BALB/cJANL females bred with C57BL/6JANL males. Of these offspring, 582 were sired by males exposed to 60 cGy of fission-spectrum neutrons and 167 were sired by unexposed males. Data set M20 [1], acquired several years prior to N89, consisted of 797 patterns from offspring of the same hybrid cross sampled at age 7 weeks. Of these, 105 were sired by untreated males, 323 by males treated with 150 mg/kg ethylnitrosourea (ENU), and 369 by males treated with 3 Gy of ^{60}Co gamma radiation. Male and female offspring were analyzed for both N89 and M20. Biopsies were performed at age 12 weeks for the N89 set and 7 weeks for M20.

Digital images were produced by scanning each gel in a tray of water using a flat-bed scanner. These images were converted to 8-bit optical density values during the scanning and the data were stored on disk. Images were processed to remove background and modeled with sets of two-dimensional Gaussian distributions with zero correlation between x and y, each distribution representing one spot. The integral of each distribution is called its volume and is a measure of the total integrated density for that spot. The ensemble of Gaussian distributions from a single image is herein called a pattern. Each pattern was matched to a single master pattern in order to establish the correspondence of each spot (represented by a Gaussian distribution) to a spot on other patterns. The master pattern is essentially a template of spot positions originally generated from one of the early patterns and to which any additional spots can be added as encountered. Matching was done in groups of 20 to 40

patterns and used both interactive and automatic procedures. The mean number of matched spots was 496± 44 for N89 and 400± 54 for M20.

Several characteristics of the data must be considered in the design of quantitative mutation screening systems:

- Quantitative variants are evidence for mutations in regulatory systems or for events where one of two homozygous alleles is no longer capable of expression. In the former case one might observe many proteins altered in amount. In the latter case, the expected effect is a decrease in the amount of one or more proteins by 50% (assuming measurement linearity and no compensation). Positional variants are often also detectable by means of a quantitative decrease at the electrophoretic position of the original allele.
- Because mutation events are relatively rare, one must analyze many patterns and utilize as many spots as possible. Experiments with more than 500 gels are typical. Fewer than 10 mutations were confirmed out of more than 1500 samples for the current data.
- Mutation experiments are generally carried out over an extended period of time since laboratories are currently unequipped to run and analyze the necessary number of samples in a short time. The experimenter must control long-term drifts that change the electrophoresis patterns and make comparison difficult. Because of this long time period it is advantageous to separate the screening process into a training phase and a test phase, with data being processed soon after the samples are taken.
- Even the best 2DE patterns contain artifacts which are often confused with mutation events. Only a very small fraction of the outliers are later confirmed to be actual mutations, and most of them can be eliminated by inspection using an interactive analysis system. A key feature of the analysis program is its ability to display many patterns simultaneously on a display screen, thus providing a context by which an event can be viewed.
- Most (if not all) of the sample types are composed of subpatterns which vary in proportion. Components may come from multiple cell types, organelles, inadvertent contamination, etc. At least three cell types are known to be present in the liver samples discussed here.

There are many possible ways to implement mutation screening systems to detect quantitative differences in individual spot amounts. All of them involve the examination of the observed values as compared with a prediction. Values that differ significantly from the prediction are flagged for review. Two methods which we have implemented are summarized below:

- Predictions based on a mean of the values for the control set. An average pattern is formed using the means of the spots in the controls. Control data are scaled using the sums of amounts over a predetermined set of spots, and the standard deviations are computed. A scale factor is similarly computed for each object pattern being screened. The predicted values to which the observed values are compared are taken to be values from the control pattern divided by the scale factor. A deviation value is computed by divid-

ing the difference of the observed and predicted values by the standard deviation. Large absolute values for this deviation are considered as possible events. This method is essentially a univariate procedure on scaled data. Our first screening studies used this technique.

- Predictions based on a Principal Components Analysis (PCA). Here the control data are analyzed by PCA, producing a set of eigenvectors. The first few eigenvectors (corresponding to the largest eigenvalues) are kept and used as a basis for a prediction system. Predictions for an individual spot in a pattern are predicted using the reduced set of eigenvectors and the vector of observed amounts for that pattern. Thus ,information for many spots in a pattern is used to predict the amount for any single spot. Standard deviations of observed amounts from the predictions are computed for the control data in the training phase. A deviation value is calculated for every spot by dividing the difference of the observed and predicted values by the standard deviation. Spots in individual object patterns are screened for large absolute values of this deviation. This screening method accounts for variation in protein loading as well as shifts in proportions between different components and has the advantage that it is based on a standard statistical package (SAS) [5].

3 RESULTS AND DISCUSSION

Performance of the two detection methods was assessed using the N89 data set. Data were randomly split into a training set (500 patterns) and a test set (249 patterns). The training set was used both for calculating the statistics required for the univariate method and as the object of a PCA analysis. Data included 276 spots from each pattern. For the PCA analysis we considered using up to the 40 largest eigenvalues and their associated eigenvectors. Seventy percent of the variation was explained by the first 23 eigenvectors and 77% by the first 40.

We used simulation to help decide how many eigenvectors to retain. Simulation was required because thousands of events are necessary for performance assessment, and the number of actual mutations expected in a data set the size of N89 was expected to be low (only two were confirmed). As we were particularly interested in detecting mutations that result in a 50% decrease in protein abundance, we simulated these events by taking 61 patterns from control animals in the test set, halving the value of a single spot in each pattern, and computing the deviation. For comparison with the univariate method, we computed its deviation by subtracting the scaled value for the spot from the average (by sex) over the training set and dividing by the standard deviation. This process was repeated for every spot under consideration, yielding approximately 16,000 simulation trials.

Figure 1 plots the detection efficiency vs. the false positive rate for the univariate and PCA methods. Examination of these and similar curves is essential for optimizing the detection efficiency. Because the entire set of principal component axes will exactly reproduce the data, care must be taken not to choose too many, as that would decrease the efficiency. Also, we wish to avoid excessive overtraining, that is, a vastly different performance be-

tween the training and test sets. With these considerations in mind, we chose to use 30 eigenvectors. Examination of curves similar to those in Figure 1 showed little advantage in using more than thirty. Some overtraining was evident by the false positive rate being significantly smaller for the training set than for the test set when more then 30 principal component axes were used. As measured with the simulation runs, the overall efficiencies in detecting 50% reductions was 51% for the univariate method and 76% for the PCA method. Other experimental strategies (analyzing duplicate samples, for example) would allow the use of a higher efficiency range.

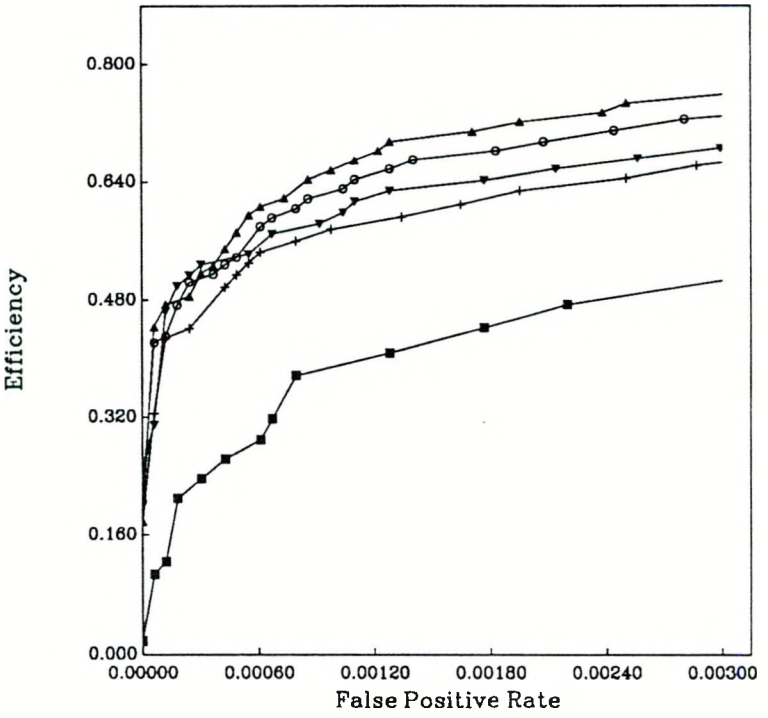

Figure 1. Detection characteristics of the various detection methods for spots which are decreased by 50%. The curves from bottom to top are the univariate method and the PCA method using 5, 10, 20, and 30 eigenvectors, respectively.

To compare the performances, we combined the N89 and M20 data sets, thereby adding considerable diversity to the data. Five hundred of the 1546 patterns (including all of the controls and none of the known mutants) were selected for training. Detection thresholds for the two methods were set by choosing an overall false positive rate of 0.3%. The thresholds were -1.9 for the univariate case and -3.5 for the PCA method. Table I summarizes the results on the set of confirmed mutations. With these diverse data, the univariate method detected two of eight events and the PCA detected six.

Mutation Event	Sex	ID of Mutated Protein	New Spot	Spot #s Affected	Deviations Univariate	PCA
ENU1	F		yes	121	-1.24	-3.90*
ENU2	M	OAT	yes	99	-1.51	-1.65
ENU4	F	CEH	yes	39	-1.42	-2.89
ENU6	M	APO A1	yes	150	-3.31*	-5.82*
ENU8	M	HSP70 fam.	yes	5	-2.07*	-7.08*
NEUT1	M		no	188	-1.69	-13.7*
	F		no	188	-1.52	-14.6*
NEUT2	F		no	94	-1.81	-3.21
				95	-0.07	-4.39*
				113	-0.75	-2.72
				188	-0.08	-2.46
				337	-0.37	-3.36

Table I. Comparison of the two detection methods on the confirmed mutation events with the training performed on pooled data from data sets N89 and M20. The * signifies the quantitative detection of the event. A "yes" in the "New Spot" column signifies that the pattern contained a qualitative mutation event.

It was concluded from these comparisons that the PCA method was superior to the univariate method, especially for studies in which the data are diverse. We have therefore adopted the PCA technique as our primary screening method.

4 ACKNOWLEDGMENTS

We thank Sandra Tollaksen, Sharron Nance, and Anne Gemmell for their excellent work in preparing the samples and running the gels. This work was supported by the United States Department of Energy, Office of Health and Environmental Research, under Contract W-31-109-ENG-38.

5 REFERENCES

[1] C.S. Giometti, M.A. Gemmell, S.L. Nance, S.L. Tollaksen, and J. Taylor. J. Biol. Chem. 1987, 262, 12764–12767.

[2] S.M. Hanash, M. Boehnke, E.H.Y. Chu, J.V. Neel, and R.D. Kuick. Proc. Natl. Acad. Sci. USA. 1988, 85, 165-169.

[3] J. Klose. Arch. Toxicol. 1977, 38, 53-60.

[4] R.R. Marshall, A.S. Raj, F.J. Grant, J.A. Heddle, Can. J. Genet. Cytol. 1983, 25, 457-466.

[5] SAS Institute. SAS/STAT User's Guide, SAS Institute Inc., Cary, NC, 1988.