BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 4 2007, pages 487–492
doi:10.1093/bioinformatics/btl626

Data and text mining

Oligonucleotide microarray identification of *Bacillus anthracis* strains using support vector machines

M. Doran[1,*], D. S. Raicu[1], J. D. Furst[1], R. Settimi[1], M. Schipma[2] and D. P. Chandler[2,†]

[1]Intelligent Multimedia Processing Laboratory, School of Computer Science, Telecommunications and Information Systems, DePaul University, Chicago, USA and [2]Biodetection Technologies, Argonne National Laboratory, Chicago, USA

Received on October 1, 2006; revised on November 30, 2006; accepted on December 4, 2006
Advance Access publication January 3, 2007
Associate Editor: Martin Bishop

ABSTRACT

The capability of a custom microarray to discriminate between closely related DNA samples is demonstrated using a set of *Bacillus anthracis* strains. The microarray was developed as a universal fingerprint device consisting of 390 genome-independent 9mer probes. The genomes of *B.anthracis* strains are monomorphic and therefore, typically difficult to distinguish using conventional molecular biology tools or microarray data clustering techniques. Using support vector machines (SVMs) as a supervised learning technique, we show that a low-density fingerprint microarray contains enough information to discriminate between *B.anthracis* strains with 90% sensitivity using a reference library constructed from six replicate arrays and three replicates for new isolates.

Contact: doran_michael@msn.com

1 INTRODUCTION

Microarray technology allows for parallel testing for the presence of many DNA sequences with a single test. A custom microarray using a set of 390 random 9mer probes was designed to allow for fine grained classification of microbial isolates without targeting a specific genome. Prior work with planar arrays has shown that this basic microarray methodology is able to easily distinguish between species or genera of isolates (Beattie, 2000; Belosludtsev *et al*., 2004) but has more limited ability when applied to closely related strains or species (Willse *et al*., 2005). Technical challenges for high-resolution DNA fingerprinting include noisy data, cross-hybridization to mismatched probes, and low-signal to noise ratios for informative signatures. Statistical challenges include methods for normalizing data across arrays, time and users; defining minimum replication requirements for constructing reference libraries or analyzing new isolates; and quantitatively comparing profiles to an established library (Willse *et al*., 2005; Chandler *et al*., 2006). *Bacillus anthracis* strains are of particular interest in this context because they are one of the most genetically homogenous bacterial species and represent a significant public health and bioterrorist threat. The ability to quantitatively distinguish between unique strains will enhance our ability to track the dissemination and movement (intentional or natural) of microorganisms through the populace or non-human vectors.

*To whom correspondence should be addressed.
†Present address: Akonni Biosystems, Inc., Frederick, MD, USA

In this paper, we propose using a probe-based supervised learning approach, the support vector machine (SVM), for classification of microarray-based DNA fingerprints, utilizing closely related *B.anthracis* strains as the model system and test set for SVM development. Our proposed approach consists of the following stages (Fig. 1): first, the data are extracted from the microarray images (Fig. 2), transformed through a logarithmic ratio between foreground and background and then normalized using quantile normalization in order to account for any systematic differences in the intensity readings caused by experimental factors (washing process and exposure time, etc). In the second stage, a classification model is built using SVMs and k-fold cross-validation. Since the proposed classification model can also be used for classifying new samples, as a third stage, a confidence interval for the classification sensitivity is calculated in order to estimate the expected sensitivity for new data.

An ANOVA based approach was successfully used to classify single replicates from a biologically diverse set 62.8% of the time without averaging replicates (Willse *et al*., 2005). Using the same dataset and a similar 3-fold cross-validation evaluation process, SVMs perform slightly better by correctly classifying 75% of the cases. While the prior work demonstrated that the microarray could be used to discriminate between a heterogeneous set of samples, the work presented in this paper explores the microarray's potential to discriminate between highly homogeneous samples by using the *B.anthracis* test dataset. Because the data are homogeneous, the classification problem presented in this paper is much harder, so the end classification results between the two experiments might not be not directly comparable. An empirical comparison of various machine learning algorithms showed that SVMs classify *B.anthracis* more accurately than a collection of other techniques. Based on these findings SVMs are an effective classification tool to study further in the context of this data mining application.

The experiment uses samples of the following strains: *B.anthracis* K4516, *B.anthracis* A0392, K2165, *B.anthracis* K8215, *B.anthracis* A0001, K0300, *B.anthracis* A0172, K3897, *B.anthracis* K7222, *B.anthracis* K4596, *B.anthracis* K4834, *B.anthracis* K2762, *B.anthracis* K0123, *B.anthracis* K0610, *B.anthracis* K1340, *B.anthracis* K2478, *B.anthracis* K9002, *B.anthracis* K7441, *B.anthracis* K1694, *B.anthracis* K5135, *B.anthracis* A0362, K8091, *B.anthracis* K1285, *B.anthracis* K1256, *B.anthracis* K2802, *B.anthracis* K7948, *B.anthracis* K7038, *B.anthracis* K6835 and *B.anthracis* K0404.

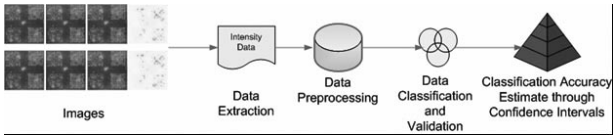

Fig. 1. Knowledge discovery process.

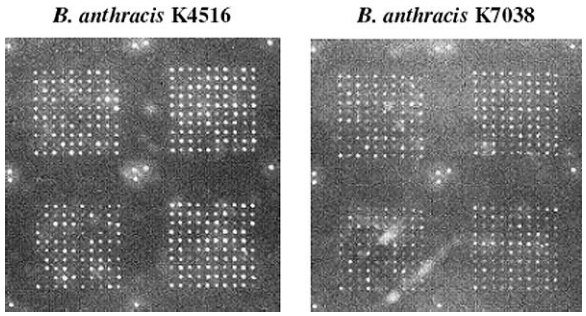

Fig. 2. Microarray images.

2 MICROARRAY AND FINGERPRINTING METHODS

The rationale, justification, design and use of the nonamer microarray is described in detail elsewhere (Chandler *et al*., 2006). In this study, we converted from planar microarrays to a gel element microarray format to address the technological issues of low-signal to noise and cross-hybridization (to be described in detail elsewhere). Briefly, nonamer probes from (Chandler *et al*., 2006) were re-synthesized with a custom allylamide linker, high-performance liquid chromatography (HPLC)-purified and incorporated into gel element arrays at 0.25 mM concentration via co-polymerization in a method (Rubina *et al*., 2004), with slight modification. Co-polymerization solutions (including the oligonucleotide probes) are printed with a QArray2 robotic arrayer (Genetix, New Milton, UK) using 150 μm blunt pins. After printing, slides were re-equilibrated by incubating overnight in an airtight container with 2–4 ml of a mixture that includes all the components of the co-polymerization mixture (minus oligonucleotides). After equilibration, arrays were photopolymerized for 30 min in a nitrogen atmosphere under a ultraviolet (UV) lamp. Finally, the slides were washed in 0.01× SSPE washing buffer (Ambion, Austin, TX) for 1 h, thoroughly rinsed with MilliQ water, and air dried. Co-polymerized arrays are stable for up to one year at room temperature. The SD of signal intensity for an array of gel elements printed with a given pin on a given slide ranged from 5 to 11%, while the average intensities for sub-arrays printed with different pins on a given slide ranged from 6 to 9%.

Genomic DNA from *B.anthracis* strains was purified and amplified by PCR (Chandler *et al*., 2006), except that amplification primers were unlabeled. Each DNA sample was amplified in nine independent reactions, and applied individually to nine independent oligonucleotide arrays. Amplification was confirmed by analyzing 20% of the reaction mixture on 2% agarose gels (Invitrogen, Carlsbad, CA). The remaining amplification products were purified in 96-well plates using a ChargeSwitch PCR Clean-Up Kit and 96-well Magnetic Separator (Invitrogen, Carlsbad, CA). Purified products were eluted in 0.01 M sodium carbonate buffer (pH 8.5) and subsequently fragmented and labeled a single-tube, radical-coordinating chemistry and lissamine rhodamine (Kelly *et al*., 2002). Resulting fragments averaged 25–150 nt in length, as determined by PAGE. Fragmented, labeled and purified amplification products were diluted in hybridization buffer to achieve a final concentration of 4× SSC, 5× Denhardt's solution, heat denatured at 95°C for 5 min, and hybridized to gel element arrays under a non-adhesive perfusion chamber overnight at 4°C. After hybridization, arrays were washed five times in ice-cold 4× SSC, dipped briefly in deionized water, air dried and imaged on a custom epifluorescent CCD imager as described previously (Chandler *et al*., 2006). Data extraction utilized the freely available AMIA software (White *et al*., 2005); normalization techniques and SVM analysis is described (as Results) below.

3 SVM OVERVIEW

A SVM is a supervised learning technique that has been successfully applied to a variety of domains, such as handwriting recognition, face detection and identification, and object recognition (Byun and Lee, 2003). More recently, the SVM approach has also been applied to pattern recognition problems in the field of computational biology; such problems include protein remote homology detection, microarray gene expression analysis, recognition of translation start sites, functional classification of promoter regions, prediction of protein–protein interactions and peptide identification from mass spectrometry data (Noble, 2004).

The motivation behind using the SVM for computational biology consists of the approach's ability to handle high-dimensional, noisy data and non-vector data inputs, (such as variable length sequences or graphs). Another advantage of SVMs is their ability to make otherwise inseparable datasets separable by mapping the data into a higher dimensional space using a kernel function. Commonly used kernel functions include polynomial kernels and radial basis functions (Vapnik, 1998).

The SVM approach, as any other supervised classification approach, uses a training dataset to build a classification model and a testing set to validate the model; each instance in the two datasets consists of a class label and a set of features whose cardinality denotes the dimensionality of the feature space. The goal of SVM is to first map the feature space into a higher dimensional feature space such that the data will be linearly separable in the new space (Cristianini and Taylor, 2000), and then find a linear separating hyperplane (Fig. 3) that will maximize the margin (sum of the distances from the hyperplane itself to the closest instances from two different classes) and minimize the empirical risk (sum of the training error).

When finding the hyperplane that separates two classes, the SVM uses a parameter c (modeling the complexity of the classifier) that will make the classifier more flexible in its ability to disregard outliers, (which might introduce too much variance in the model). The ability to disregard outliers is an important advantage of SVMs since it reduces the burden on the experimenter to develop a separate process for removing outliers as new data or classes of data are accumulated.

The classification model produced by the SVM approach consists of a set of weights which, besides being used for the classification of new data, can also be used to identify the most discriminative

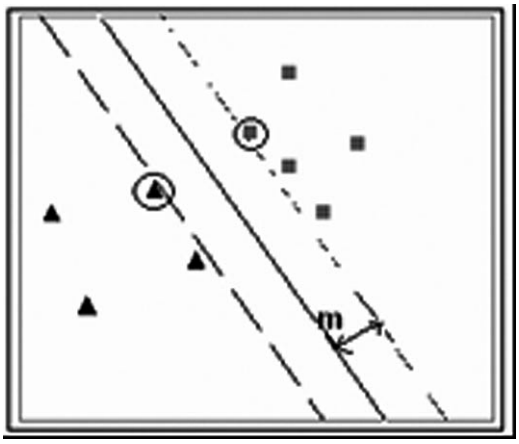

Fig. 3. Training data and the SVM classifier: the triangles and squares represent the two classes for the training data, the circled points show the support vectors for the corresponding classes, the solid line shows the optimal hyperplane, and *m* represents the geometrical margin.

features for separating the data into two classes. Through a recursive feature elimination (RFE) process, the probes are ranked according to these weights and the probes with the lowest weights are recursively eliminated from the classification process while keeping the accuracy of the classification at a specified level. RFE is applied with a linear kernel. Reducing the number of features is a key issue for datasets with small number of cases/records but high-number of features, a situation typified by microarray data (Brown *et al*., 2000).

While the foregoing explanation of a SVM approach is for binary classification, the model can be generalized to multiple classes in two different ways (Weston and Watkins, 1998). The first approach for multiple classes builds a one compared to all model: for each class, the optimal hyperplane separating the positive class (target class) and the negative class (all the other classes) is sought. The second approach builds a pair-wise model: a hyperplane for each pair of classes is found. Using the pair-wise approach an unknown (unlabeled) instance can be classified in a number of ways, such as by comparing it to all of the hyperplanes and observing which class it resembles most frequently.

4 SVM FOR DNA FINGERPRINTS CLASSIFCATION

In this study, we applied the SVM approach for the identification of *B.anthracis* strains using oligonucleotide microarray hybridization data. The microarray datasets used to create and validate the classification model are in the form of a matrix with each row denoting a bacterial strain and one microarray experiment, each column denoting a probe on the microarray, and each element of the matrix representing a probe signal intensity. The classification is performed along the row dimension (bacterial strain) after the raw data has been preprocessed due to the variances produced as side effects of experimental factors unrelated to hybridization levels.

4.1 Data preprocessing

The preprocessing stage consists of three steps: (1) filtering to remove extraneous signal un-related to the immobilized probe, (2) transformation to obtain the hybridization level for each probe and (3) normalization to adjust for experimental factors that may have made some replicates more or less intense than others.

4.1.1 Filtering and transformation Each microarray and data vector contains nine fluorescently-labeled control probes (or beacons) that provide information about the orientation of the chip, manufacturing quality and signal intensity levels and values that are independent of the isolate tested on the chip, hybridization or wash procedures. These probes are excluded from the dataset since they do not include any information about the DNA sample being tested. The information represented by the control probes could be used for intensity normalization, but they are not necessary for the normalization approach used here.

After the filtered probe intensities are acquired from the microarray images and the control probes are eliminated, the probe intensities are transformed using a logarithm function computed as the log of the ratio of the foreground to background; this transformation is a standard practice in the industry (Affymetrix, 2002; Eisen *et al*., 1998) and was also used in a prior applications of the fingerprinting chip methodology (Willse *et al*., 2005). Using the ratio of the foreground to background intensity adjusts for anomalies in the chip image close to the spot and reduces the variance in the data at the high end of the intensity range. In the microarray images, the high-intensity probes show up as bright dots and the lowest intensity probes appear to be missing points in the grid. A few probes have a bright cloud around them showing how the background can be influenced by the probe response.

4.1.2 Normalization While the logarithmic transformation reduces the variance, it does not always make the distributions close enough to the Gaussian normal distribution for the application of statistical inference later in the process (Parrish and Rudolph, 2004). There are many systematic errors that can skew intensity readings in microarray data, including (but not limited to) the exact amount of target DNA applied to the chip, the washing process, the exposure time and even the chip production process (Lee *et al*., 2000). While protocols are established to minimize these errors, some form of preprocessing is usually needed to standardize the data.

Different normalization techniques have been used to normalize microarray data. A common industry practice is to adjust the means of each microarray to be equal and then use a linear transform to equalize the high and low-end ranges (Affymetrix, 2002). If the ranges are similar, another approach that can be applied is the low-end mode normalization (Willse *et al*., 2005) which may, however, present some automation challenges. Scaling the variance's heterogeneity uses a linear scaling to force interquartile distances to be identical between arrays and offers another normalization technique that is more resistant to outliers (Parrish and Delongchamp, 2004). M-A plots (Bolstad *et al*., 2003), and loess smoothing (Cleveland and Devlin, 1988) are two other techniques used for normalizing microarray data. From (Willse *et al*., 2005), we know that all of the isolates used in this gel element array and SVM study are so closely related (as few as three probes may be significantly different) that assuming an identical distribution is reasonable. Therefore, we apply a quantile normalization to produce similar probe intensity distributions based on the properties of the corresponding microarray data (these custom chips use random 9mers).

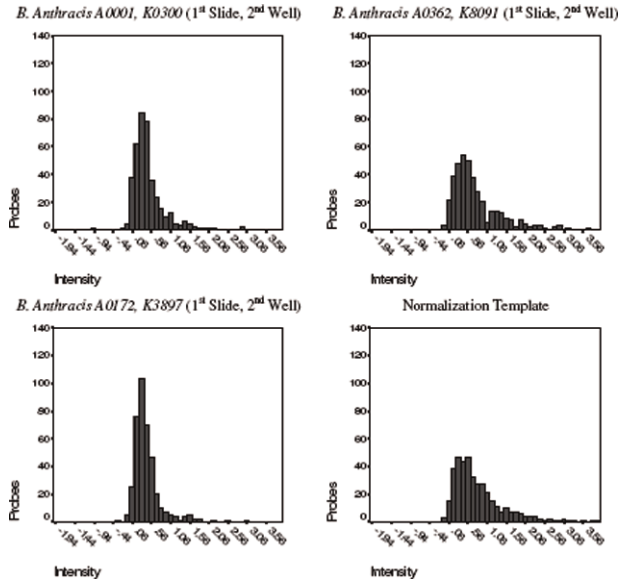

Fig. 4. Intensity histograms and normalization template.

Quantile normalization generates a mean distribution template that is used to normalize every replicate. A baseline template is generated as a vector

$$b = \left(\frac{1}{n} \sum_{i=1}^{n} q_{i1}, \frac{1}{n} \sum_{i=1}^{n} q_{i2}, \ldots, \frac{1}{n} \sum_{i=1}^{n} q_{iM} \right), \tag{1}$$

where n is the number of microarrays, M is the number of probes and q_{ij} is the jth quantile of the ith microarray. With the baseline template generated, the values of each microarray q are transformed such that

$$q_{ij} = b_j. \tag{2}$$

To make the algorithm slightly more resistant to outliers, the b_j values are computed as the median instead of the mean of all sample quantiles j. This form of normalization eliminates any problems involved in selecting a base line image (Bolstad *et al*., 2003) and, when using the median instead of the mean, is not greatly influenced by outliers. Examples of the changes in signal intensity histograms after the quantile normalization are shown in Figure 4. It is worth mentioning here that, for datasets with different distributions among the bacterial strains, other normalization techniques should be applied.

4.2 SVM classification model creation and validation

Prior work applying SVMs to microarray data has shown that they work well with sparse datasets, large numbers of attributes and outliers (Brown *et al*., 2000). These strengths address key issues with the data set of *B.anthracis* strains (390 probes, 9 replicates per isolate and 13 strains/isolates) that has been chosen to demonstrate the capability of the SVM approach to quantitatively discriminate between closely related DNA samples and microarray signatures.

Even though there are hundreds of probes, only a few are discriminating probes and thus, the data are sparse by almost any standard. Furthermore, the large number of independent probes (390) compared to the number of replicates (9) introduces the 'curse of dimensionality' problem and produces Type I and II errors during classification (Table 1). Therefore, any array-based, DNA fingerprint classifier must be able to work with large number of probes or it must be combined with an attribute selection process.

Table 1. Null hypothesis = there is no difference among the samples, alternative hypothesis = there is a significant difference among the samples

	Computed accept	Computed reject
Accept null hypothesis	True accept	Type I error
Reject null hypothesis	Type II error	True reject

Because the SVM approach, along with the RFE wrapper approach, can be used to classify high-dimensional data and also rank the probes accordingly to the margins calculated in building the SVM model (Witten and Frank, 2005), SVMs provide an opportunity to satisfy biological intuition and data reduction requirements. Thus, the probes with the highest margins are considered to be the most significant for discrimination among bacterial strains. Moreover, the two or three most discriminant probes can be used as the axes of a 2D or 3D space since the hyperplanes found by the SVM model will optimally divide the set of points with respect to the probes that are differentially hybridized or detected. Extending the utility of confidence intervals and margins was beyond the scope of this study and will be reported elsewhere.

In order to build and validate the classification model, a holdout approach is usually used to sample the data: 66% of the original data are randomly selected from the original data to form the training set on which the model is constructed and the rest forms the testing set on which the model is validated. However, since the data used here contain only few replicates/samples per strain, we applied a stratified cross-validation technique to build and validate the classification model. The data are partitioned into k sets of equal size that are referred to as folds. Each set contains as equal a number of each isolate as possible. The fold size k is chosen to be a factor of the number of replicates so that an equal number of samples are available for each fold. Each one of the k folds becomes a testing set while the other $k - 1$ folds form the training set. Therefore, k classification models are produced on the training sets and tested on the corresponding testing sets. The most general measure of model performance is classifier error rate (E), which is calculated as:

$$\text{Classification Error Rate} = \frac{\text{\# of misclassifications}}{\text{\# of test set instances}}. \tag{3}$$

The error rates for each of the k repetitions are averaged to estimate the model's expected error rate (Han and Kamber, 2001). The error rate E is also equivalent to 1—Sensitivity, where the sensitivity is calculated as:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{4}$$

True positives are identified as the number of cases (replicates) that are classified correctly; false positives are defined as the total number of cases incorrectly classified. Figure 5 illustrates the cross-validation approach.

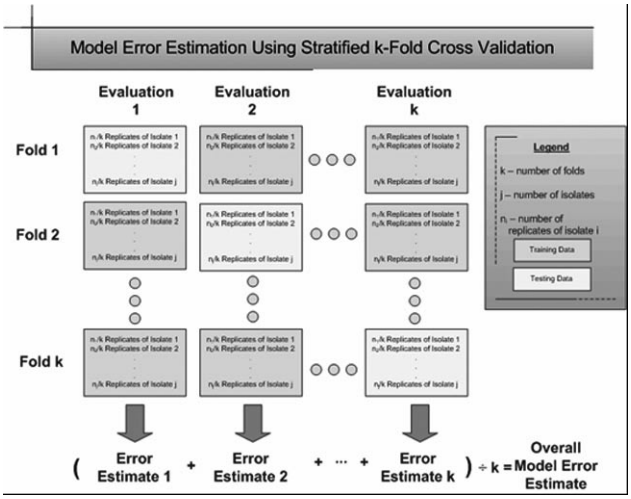

Fig. 5. The cross-validation approach.

In general, when testing the accuracy of any classifier, the classifier is applied individually; i.e. on any case from the testing set. However, since we do not have enough cases per isolate and in order to minimize the impact of the possible outliers, a few replicates are averaged together and then classified as a single combined testing case. The cases from the training set are not averaged since we want the SVM classifier to learn the amount of expected variance among the replicates within the same isolate. The possible outliers from the training set are expected to not influence the classification model since the constant c calculated when building the SVM classifier will allow disregarding the outliers from the training set (lower values of c are more likely to discard data points considered to be outliers).

4.3 Confidence intervals for classification sensitivity

The purpose of a classifier error rate computation is to give an indication about the likely future performance of the model. The error rates of each individual isolate can be combined into an overall average E because they are assumed to be estimates of the same number (the isolates are all very closely related so they are expected to have very similar classification error rates). Each test can be viewed as a Bernoulli trial since each sample is independent. The steps involved to determine a confidence interval for a computed error rate E for a set sample S of size n are as follows (Roiger and Geatz, 2003):

(1) compute the sample variance

$$\text{Variance} = E(1 - E). \tag{5}$$

(2) compute the standard error

$$SE = \sqrt{\frac{\text{Variance}}{n}}. \tag{6}$$

(3) calculate an approximate upper bound for the 95% confidence interval as

$$\text{Sensitivity} + 2 * SE. \tag{7}$$

(4) calculate an approximate lower bound for the 95% confidence interval as

$$\text{Sensitivity} - 2 * SE. \tag{8}$$

The above formula for the approximate confidence interval also shows that if we are increasing the number of test set instances (in our case, the number of replicates per isolate), we are decreasing the range of the confidence interval.

5 RESULTS

SVM models were constructed using the Sequential Minimal Optimization algorithm (Platt, 1998) implemented in the Weka machine learning package (Witten and Frank, 2005). A 3-fold stratified cross-validation was used to evaluate the models. The three-replicates of each isolate are included in each fold so that there are enough cases to average together when performing the testing. This is a logical way to break up the data since 3 is the only factor of 9 that allows for a uniform split of the data while including multiple samples in each fold to allow for averaging of replicates. The entire cross-validation was repeated four times to reduce the effect of particular random samples on the results. One of the 26 isolates was excluded from this analysis because of an unusable replicate. With only eight replicates available the number of test or training cases available in each fold would be lower than the other isolates.

Normalizing the data increased classification results by almost 9%. With only nine replicates per isolate, this means that on average <1 replicate for each isolate was misclassified. Averaging test samples together to decrease intensity variation and using quantile normalization improved results from 73.3 to over 89%.

Typically, an advantage of SVMs is their ability to make otherwise inseparable datasets separable by transforming the original feature space (each feature corresponds to a probe) into a higher dimensional space using a kernel function. We tried several polynomial kernel functions $K(x,y) = (x^T y + 1)^d$, for $d = 1, \ldots, 7$, and different values for the parameter c ($c = 1, \ldots, 6$) and we noticed that the optimal results were achieved using a linear kernel [$d = 1$, commonly defined simply as $K(x,y) = (x^T y)$] with a c parameter >1. Obtaining the best results using a linear kernel can be explained by the fact that each probe is probably completely independent and so, transforming the data does not offer any clear advantages. Modification of the c parameter had a small effect on results for the linear kernel but in general had no noticeable impact. The approximate 95% confidence interval for the linear model's sensitivity was between 85.6 and 93.8%.

The accuracy of the SVM classification model was also compared to the number (and pattern) of REP-PCR bands observed on the agarose gels. In this study, the number of visible PCR amplicons ranged from 1 to 7 per isolate. Experimentally, the number of visible bands can be attributed either to differences between input concentrations of genomic DNA entering the PCR (range 10–100 ng per isolate) and PCR output (experimental variance); normally, the number and pattern of bands is taken to reflect true differences in genome structure between isolates (biological variation). Each aliquot of *B.anthracis* DNA was received with a reported DNA concentration. DNA concentrations were also verified and measured

separately by gel electrophoresis prior to PCR amplification. Reported concentrations, experimentally verified concentrations and the ratio between the two were then compared to the sensitivity of the SVM model for each isolate, where the sensitivity is defined as the percentage of test cases of the isolate that are correctly identified. The correlation between the reported DNA concentration and the SVM sensitivity was 0.009; the correlation between the experimentally estimated DNA concentration and the SVM sensitivity was 0.049; and the correlation between the ratio of the two DNA concentrations and the SVM sensitivity was 0.034. Based on these very low correlations, it is unlikely that the quantity of genomic DNA entering the PCR (between 10 and 100 ng) had a significant affect on the discriminating capability of the microarray and SVM method reported here.

On the other hand, 58% of the incorrectly identified replicates were classified as an isolate that generated a similar number of PCR bands (i.e. gel-based DNA fingerprint), and 76% of the remaining, incorrectly identified cases are confused with an isolate that generated only 1 additional visible PCR band. The test cases that were difficult to identify via the microarray and SVM method were also difficult or impossible to differentiate based on the agarose gel data (data not shown). These results indicate that discriminatory microarray signatures arise from true biological differences in the isolates, and are not an artifact of experimental variation during the PCR and microarray procedure, and that statistical confusion between microarray fingerprints correlates with bacterial relatedness as typically estimated by band-sharing in a gel-based fingerprint.

6 CONCLUSION

Based on these results we conclude that the low-density, genome-independent fingerprint microarray contains enough information to distinguish between very closely related isolates of *B.anthracis* when combined with a SVM classifier method. With ~95% confidence between 85.6 and 93.8% of test samples for *B.anthracis* isolates were correctly identified using a library constructed with six replicates from each isolate and averaging three test replicates together to reduce the variance. The sensitivity and accuracy of the SVM classification can be improved by increasing the number of replicates used to generate the initial library and reference set (beyond six replicates), or increasing the number of replications (beyond three replicates) performed on new, uncharacterized isolates prior to classification and library comparisons.

7 FUTURE WORK

One of the objects of exploring new microarray data analysis techniques is to reduce the number of microarray replicates that are required to quantitatively discriminate between closely related strains at a known level of statistical confidence. The results presented here, using a gel element array format for microarray-based DNA fingerprinting, combined with previous data using planar microarrays, provides the opportunity to quantitatively compare microarray substrates, apply different data analysis techniques to create a classification model with better sensitivity, or develop new microarray image processing techniques for better microarray image quality. The margins used in each SVM model can also be ranked to facilitate the identification of discriminating probes and potentially identify signatures that are generic to a class of organisms (e.g. *B.anthracis* rather than *Bacillus cereus* or *Bacillus thuringiensis*) rather than classifying isolates at the sub-species level.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2002) *Statistical Algorithms Description Document*.

Beattie,K.L. (2000) Arbitrary sequence oligonucleotide fingerprinting. U.S. patent 6,156,502.

Belosludtsev,Y.Y. *et al.* (2004) Organism identification using a genome sequence-independent universal microarray probe set. *BioTechniques*, **37**, 654–660.

Bennet and Campbell (2000) Support vector machines: hype or hallelujah? *ACM SIGKDD Explorations Newslett.*, **2**.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.

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

Byun,H. and Lee,S.-W. (2003) A survey on pattern recognition applications of support vector machines. *Intern. J. Pattern Recognit. Artif. Intell.*, **17**, 459–486.

Chandler,D.P. *et al.* (2006) Diagnostic oligonucleotide microarray fingerprinting of *Bacillus* isolates. *J. Clin. Microbiol.*, **44**, 244–250.

Cleveland,W.S. and Devlin,S.J. (1988) Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, **83**, 596–610.

Cristianini,N. and Shawe-Taylor,J. (2000) *'An Introduction to Support Vector Machines and Other Kernel-based Learning Methods'*. Cambridge University Press.

Dudoit,S. *et al.* (2002) Multiple Hypothesis Testing in Microarray Experiments. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 110.

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

Han,J. and Kamber,M. (2001) *Data Mining Concepts and Techniques*. Academic Press.

Kelly,J.J. *et al.* (2002) Radical generating coordination complexes as a tool for rapid and effective fluorescent labeling and fragmentation of DNA or RNA for microarray hybridization. *Anal. Biochem.*, **311**, 103–118.

Lee,M.L. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.

Noble,W.S. (2004) Support vector machine applications in computational biology. In Schoelkopf,B., Tsuda,K. and Vert,J.-P. (eds), *Kernel Methods in Computational Biology*. MIT Press, pp. 71–92.

Parrish,R.S. and Delongchamp,R.R. (2004) 'Normalization of microarray data' in DNA microarrays and related genomic techniques: design, analysis and interpretation of experiments. In Allison,D.B., Page,G.P., Beasley,T.M. and Edwards,J.W. (eds), Chapman & Hall CRC, NY, pp. 9–28.

Platt,J. (1998) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Microsoft Research Technical Report MSR-TR-98-14*.

Roiger,R.J. and Geatz,M.W. (2003) *'Data Mining: A tutorial-based primer'*. Addison-Wesley, pp. 222–224.

Rubina,A.Y. *et al.* (2004) Hydrogel drop microchips with immobilized DNA: properties and methods for large scale production. *Anal. Biochem.*, **325**, 92–106.

Vapnik,V. (1998) *Statistical Learning Theory*. John Wiley & Sons, Inc., NY.

Weston,J. and Watkins,C. (1998) Multi-class support vector machines. *Technical Report CSD-TR-98-04*. Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, England.

White,A.M. *et al.* (2005) Automated microarray image analysis toolbox for MATLAB. *Bioinformatics*, **21**, 3578–3579.

Willse,A. *et al.* (2005) 'Comparing bacterial DNA microarray fingerprints'. *Stat. Appl. Genet. Mol. Biol.*, **4**, 19.

Witten,I.H. and Frank,E. (2005) *'Data Mining: Practical Machine Learning Tools and Techniques'*. 2nd edn. Morgan Kaufmann, San Francisco.