Automated Validation of Polymerase Chain Reactions Using Amplicon Melting Curves

Tobias P. Mann
Department of Genome Sciences
University of Washington
Seattle, WA, USA

Richard Humbert
Department of Molecular Biology
Regulome
Seattle, WA, USA

John A. Stamatoyannopolous
Department of Molecular Biology
Regulome
Seattle, WA, USA

William Stafford Noble
Department of Genome Sciences
University of Washington
Seattle, WA, USA

Abstract

PCR, the polymerase chain reaction, is a fundamental tool of molecular biology. Quantitative PCR is the gold-standard methodology for determination of DNA copy numbers, quantitating transcription, and numerous other applications. A major barrier to large-scale application of PCR for quantitative genomic analyses is the current requirement for manual validation of individual PCR reactions to ensure generation of a single product. This typically requires visual inspection either of gel electrophoreses or temperature dissociation ("melting") curves of individual PCR reactions—a time-consuming and costly process.

Here we describe a robust computational solution to this fundamental problem. Using a training set of 10,080 reactions comprising multiple quantitative PCR reactions from each of 1,728 unique human genomic amplicons, we developed a support vector machine classifier capable of discriminating single-product PCR reactions with better than 99% accuracy. This approach has broad utility, and eliminates a major bottleneck to widespread application of PCR for high-throughput genomic applications.

1. Introduction

PCR (18) is perhaps the most widely applied technique in molecular biology and functional genomics, with applications ranging from gene discovery to microarray probe synthesis (8; 16). A major difficulty in applying PCR to large scale genomic analyses is the current lack of tools for automated validation of PCR reactions. Here we describe such an automated method based on classifier analysis of standard, post-amplification PCR product temperature dissociation curves (so-called "melting curves").

All quantitative applications of PCR require generation of a single product during the PCR reaction. This is also true of manufacturing applications such as production of cDNA microarrays. One approach to validating individual PCR reactions relies upon measurement of the melting curve of the amplicon following completion of PCR thermal cycling. Currently, such melting curves are examined by human operators, who must decide if the reaction was acceptable. When the number of reactions is large, this manual screening process becomes tedious and time consuming. Furthermore, manual screening may be error prone due to the superficial homogeneity of melting curves and the potential for inter-observer variability.

We aim to automate the screening of PCR reactions based on analysis of melting curves. We use a manually labeled data set and features derived from amplicon melting curves to train a support vector machine (SVM) classifier to discriminate between acceptable and aberrant PCRs. The features include a low-dimensional representation of the melting curve and information about the representation error. We then employ the classifier to identify the aberrant PCR reactions automatically.

Our approach has three significant advantages over manual screening. First, the classifier is vastly faster and more scalable than human labeling. Second, the classifier produces quantitative probability estimates of melting curve abberancy. These probabilities allow the stringency of filtering to be set at the outset. Controlling the stringency of filtering is important for interpretation of downstream analysis, where it may be very important to have a minimum threshold of certainty that analyzed data represents acceptable reactions. Finally, classifiers can be retrained according to new labeling requirements, and thus can rapidly support different analysis scenarios.

Figure 1 illustrates our method. Our machine learning

0-7695-2344-7/05 $20.00 © 2005 **IEEE**

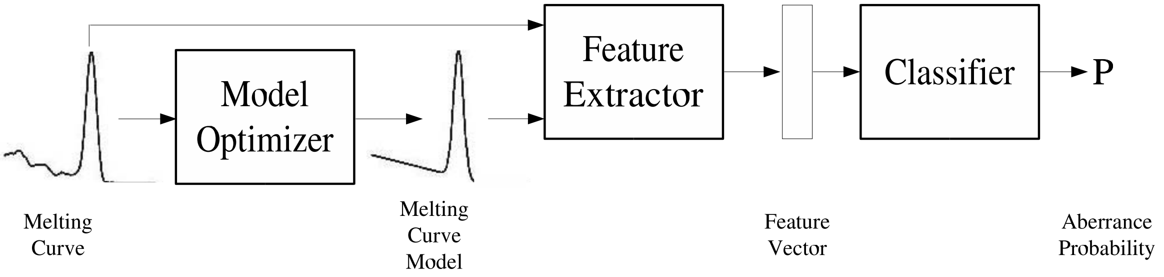

Figure 1. Diagram of automated screening method. The melting curve is given as input to the model optimizer, which fits a model of the melting curve and outputs the model parameters. The melting curve model parameters and the melting curve are then given as input to the feature extractor, which produces a vector of features. Finally, the features are given as input to the classifier, which produces as output the probability that the melting curve is aberrant.

approach relies on two components. First, we use features derived from the melting curves as input to the classifier. These features are computed by modelling each melting curve using a family of functions that can closely approximate melting curves from acceptable PCRs, and then computing statistics about the data as well as the error of the approximation. We developed a set of 36 features describing the model, and the quality of the model's approximation to the data, that are presented to the classifier. Second, we use a support vector machine classifier (20; 2) for discrimination. The SVM implicitly defines a hyperplane in a vector space that is used to classify the data points. This hyperplane is used to compute a numerical score for each melting curve based on a nonlinear mapping of the 36 features. This numerical score is then processed to yield an estimate of the probability of melting curve aberrance.

Using this machine learning approach, we demonstrate outstanding accuracy in distinguishing acceptable from aberrant melting curves, achieving a receiver operating characteristic (ROC) score of 0.998. We show that high accuracy can be achieved with a relatively small training set, and furthermore that SVM outputs can be calibrated for *a priori* specification of the decision threshold in terms of probabilities. This work is significant because it removes a bottleneck in large scale analysis of genomic DNA by PCR.

Below we describe (i) the melting curve collection process; (ii) a melting curve model together with features used to classify melting curves; (iii) experimental validation of our approach; and (iv) the overall results.

2. Data Generation

We used real-time quantitative PCR (qPCR) (7) data to develop our classifier. Quantitative real-time PCR is an extension of PCR in which a signal from a fluorescent reporter is monitored throughout the reaction. This fluorescence signal is used to track the progress of the reaction after each cycle of amplification. Fluorescence data can in turn be analyzed to determine accurately the initial template concentration. In the dataset we analyze, the dye SYBR Green I (6) is used as the fluorescent reporter. SYBR Green I's fluorescence increases substantially upon binding to double-stranded DNA. Reactions were cycled and fluorescence data acquired using a standard qPCR instrument (ABI 7900, Applied Biosystems, Foster City, CA).

Use of SYBR Green I has the advantage that melting curve analysis can be used to validate the RTPCR immediately after thermal cycling without further sample processing. A melting curve is measured after the last extension phase of the reaction by first bringing the reaction mixture to a relatively low temperature, and then slowly heating the reaction mixture to a temperature at which the DNA strands are expected to be denatured. The fluorescence signal decreases during this process, and usually most of the decrease occurs over a narrow temperature range. This sudden decrease is caused by rapid denaturation of the PCR amplicon in a cooperative process (15). Typically, PCR amplicons will denature with one relatively sharp transition, although some sequences have been shown to denature with more complex profiles (10). Melting curves showing one transition are used for downstream analysis, and melting curves with unusual profiles are reserved for further examination.

Figure 2A shows a typical melting curve. The melting curve is plotted as the negative of the first derivative of the SYBR Green I fluorescence signal as a function of temperature. As the temperature increases, SYBR Green I fluorescence decreases slightly due to a temperature dependent effect that is independent of DNA dissociation. When regions of the amplicon melt out from helix to coil, peaks are observed in the melting curve due to significant decrease in SYBR Green I fluorescence. When the amplicon is entirely dissociated at high temperatures, there is a small decrease in SYBR Green I fluorescence as the temperature is further increased.

Melting curves were generated using a standard option on the ABI 7900 instrument. For each qPCR reaction, we have measurements of the fluorescence signal at a discrete set of temperatures. Each melting curve consists of approximately 95 such measurements taken at temperatures ranging from 67° C to 93° C.

3. Algorithms

Although the classifier can be trained on the raw melting curves, discrimination based on the raw data will not generalize well for two reasons. First, using raw data introduces sensitivity to the sampling of the temperature axis. In order to use a classifier on datasets with different temperature sampling, an interpolation scheme would be necessary. Second, new data gathered on a different temperature range would introduce difficulties in applying the classifier, because data would need to be truncated or extrapolated. For these reasons, we opted for an approach in which features were extracted from the raw data.

Given that acceptable amplicon melting curves typically show a single transition, we choose to model melting curves as the superposition of a Gaussian function and a linear function. An example of a melting curve model fit to data is shown in figure 2A. Although the melting curve transitions have a slight asymmetry, a Gaussian function is able to fit the transition well. We chose to use a linear function that terminates at a chosen temperature, because after the amplicon melts out, the fluorescence decrease is negligible. The melting curve model we use is motivated by examination of the melting curve data, rather than the physical processes underlying the interaction of the fluorescent dye and the dissociating DNA duplexes. Because we do not model the underlying processes, our melting curve will have systematic biases in the error between the model and the data. However, because our goal was to reliably identify aberrant melting curves, our focus was on developing a model that could be used to extract informative features.

The melting curve approximation $M(t; m, b, t_0, a, \mu, \sigma)$ is a function of the temperature t and parameters of the linear and Gaussian components of the model,

$$\begin{aligned} M(t; m, b, t_0, a, \mu, \sigma) &= L(t; m, b, t_0) + G(t; a, \mu, \sigma) \\ &= (mt + b)\frac{e^{-(t-t_0)}}{1 + e^{-(t-t_0)}} \\ &\quad + ae^{-\frac{(t-\mu)^2}{2\sigma^2}}. \end{aligned} \tag{1}$$

The linear component $L(t; m, b, t_0)$ is determined by three parameters, which are the slope m, line intercept b, and linear component stopping temperature t_0. The linear component is the product of a line segment and a sigmoidal activation function. Using the product of these two functions allows the line segment to terminate continuously. This family of line segments was chosen to preserve the continuity of the derivative of the model, so that a nonlinear optimization procedure could be used. The Gaussian function $G(t; a, \mu, \sigma)$ is determined by three parameters consisting of the mean μ, standard deviation σ, and amplitude a.

Thus, the melting curve model has six parameters. These parameters are chosen for each melting curve by minimizing the mean squared error between the melting curve model and the measured melting curve using a Levenburg-Marquardt (11) optimization procedure.

Based on the model and the data, a total of 36 features are computed for each melting curve and used as inputs to the SVM. The features are as follows.

1. The six parameters of the melting curve model.

2. A measure of error relative to the Euclidean norm of the data,

$$relative\ error = \frac{\|data - model\|}{\|data\|}.$$

3. A twelve bin histogram recording the distribution of melting curve values.

4. A twelve bin histogram recording the distribution of error values with respect to the model.

5. The absolute error between the second most intense peak found in the melting curve and the optimized model.

6. The ratio of the amplitude of the second highest peak found and the amplitude of the most prominent peak.

7. The absolute error of the linear portion of the curve.

8. The absolute error after the Gaussian peak.

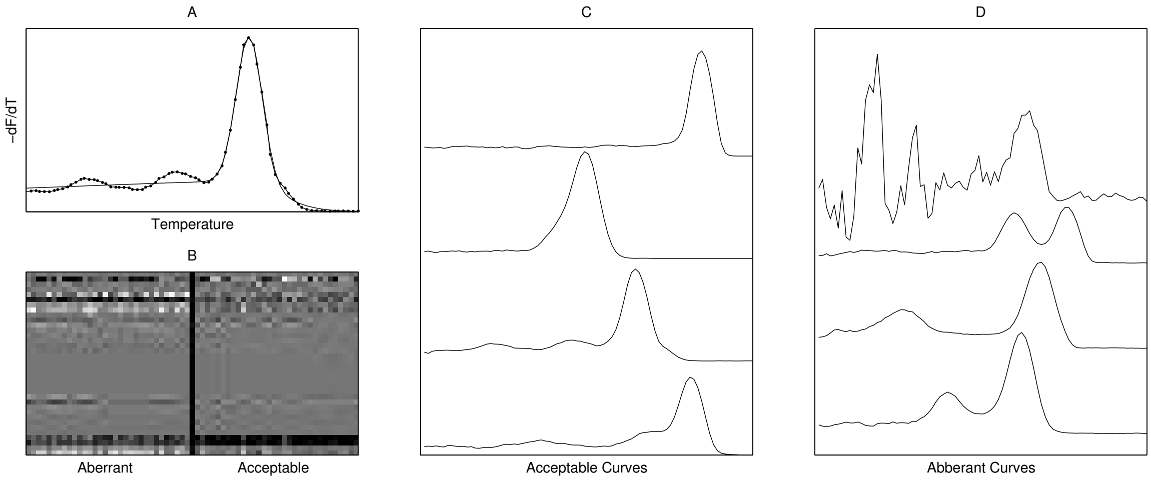

Figure 2. Examples of data and feature vectors. (A) Melting curve example (plotted in dots) and model fit (solid line). The melting curve is the negative derivative (-dF/dT) of the SYBR Green I fluorescence signal as a function of temperature, and characterizes thermally induced dissociation of the amplicon. (B) Heat map of feature vectors derived from aberrant and acceptable melting curves. Low numerical values are plotted as dark pixels, and high numerical values are plotted as light pixels. Each column corresponds to a melting curve, and the 36 rows contain the feature data for each melting curve. The task of the classifier is to distinguish acceptable from aberrant curves on the basis of these features. (C) Acceptable melting curves. These curves have a dominant peak and a relatively linear component before the main peak on the temperature axis. The different peak locations correspond to different amplicon melting temperatures. (D) Aberrant melting curves. Curves are classified as aberrant if they have a large secondary peak (bottom three examples) or some other unusual morphology (top example).

9. The sum of the error in the segment with the maximum error sum, where t_l and t_h are temperatures at which the fluorescence of the mixture was measured:

$$\max_{t_l < t_h} \sum_{t=t_l}^{t_h} (data[t] - M[t; m, b, t_0, a, \mu, \sigma]).$$

The heterogeneous nature of these features reflects the morphological variety of the melting curves. Because there are a variety of ways in which an acceptable melting curve can depart from the superposition of a line segment and a Gaussian peak, a single error statistic does not provide sufficient information for the classifier to reliably identify aberrant curves. For instance, many acceptable melting curves have at least one secondary peak (see figure 2A), but other curves were classified as aberrant if the secondary peak was too high in comparison with the dominant peak. For the subtask of distinguishing aberrant curves from acceptable curves on the basis of the second peak, features five and six are informative. Other melting curves have heavy shoulders that are not distinct peaks (see bottom curve in figure 2C). These heavy shoulders lead to a high error but are nevertheless acceptable. Feature nine is informative in separating these curves from the curves in which high error is due to a prominent secondary peak. The other features were developed to assist the classifier in correctly identifying other marginal cases. None of the data used to develop features was used to evaluate the final classifier performance.

Once features are computed for each curve in the dataset, we train an SVM classifier to predict the label assigned to each melting curve on the basis of the computed features. SVMs provide state-of-the-art classification performance, and have been used extensively in a wide variety of machine learning tasks, such as handwriting recognition, face detection, and text categorization (2). They have also been applied to a number of problems in computational biology, such as peptide identification in mass spectrometry data, remote homology detection, and microarray gene expression analysis (13).

An SVM uses a hyperplane to assign examples to one of two classes, and thus is similar to a perceptron. The perceptron training algorithm chooses a hyperplane in the vector space defined by the input features that separates positive from negative examples in a training set. In contrast, the SVM incorporates three improvements over the classical perceptron algorithm. First, motivated by statistical learning theory (20), the SVM training algorithm chooses the hyperplane with the maximum distance from the positive and negative examples near the decision boundary. Second, the hyperplane is expressed as a weighted combination of the training data, and thus the complexity of the decision function is decoupled from the complexity of the feature space. Finally, the SVM incorporates a generalized measure of similarity called a kernel function, which allows input features to be nonlinearly mapped into a higher-dimensional vector space. An appropriate choice of kernel function can improve the separation of positive and negative examples as compared to the native input feature space.

4. Methods

We used a sample of 10,080 manually labeled melting curves in our study, representing 1,728 primer pairs that were each run at least three times. The average amplicon size was 241 bp, typical for quantitative PCR applications. These melting curves were collected in studies of DNAse I hypersensitive regions in the human genome (3; 17; 5). The amplicons from which the melting curves were derived span a total of half a megabase of human genomic sequence located on chromosome 11. Amplicons encompassed a wide variety of genomic contexts, including CpG islands, repeat elements, genes, and intergenic regions.

These curves were initially labeled by laboratory technicians, who examined one of the replicates for an amplicon and assigned the category of the examined replicate to all melting curves of the amplicon. These melting curves were labelled as acceptable if they had one dominant narrow peak. See Figure 2C and 2D for examples of acceptable and aberrant melting curves. In this dataset, approximately five percent of the reactions were aberrant. Each individual curve was then checked prior to any SVM training for correct categorization by TPM. Upon examination, 47 curves were relabeled as acceptable and 152 curves were relabeled as aberrant.

We used the publicly available python package PyML (pyml.sourceforge.net) to choose the maximum-margin hyperplane separating the positive examples from the negative examples and compute cross-validated ROC 1% scores. Parameters were selected via cross validation on a training set as described below.

We used ROC analysis (12) to evaluate the performance of our classifiers. An ROC curve is generated by plotting the true positive fraction as a function of the false positive fraction as the decision threshold is swept through the range of the classifier outputs. The ROC curve provides information about the ranking of examples induced by the SVM. For the application described in this paper, we are particularly interested in the fraction of aberrant examples that appear early in the ranking of examples from aberrant to acceptable. We thus used the ROC 1% score, which is the normalized area under the ROC curve including only the first 1% of false positives. The ROC 1% score for a perfect classifier would be 1, and for this data set, the expected ROC 1% score for a totally random classifier would be $5 \cdot 10^{-3}$.

ROC analysis is useful because it provides information about the total ranking produced by a classifier for a variety of thresholds. However, in order to use a classifier as a component in a larger analysis task, it is necessary to choose a specific threshold for the classifier outputs, so that data is separated into acceptable and aberrant groups. In order to separate the choice of threshold from the output distribution of the specific SVM under consideration, the SVM outputs are further processed to yield probabilities of aberrancy. Although SVMs produce a numerical score which has no inherent probabilistic interpretation, empirical probabilities can be produced with the use of an appropriate sigmoid function. The parameters of the sigmoid function are estimated from hold-out training data (14). This sigmoid function takes as input the classifier output, and produces an output that can be interpreted as a probability of aberrancy.

5. Results

Our experiments show that our approach can identify aberrant melting curves with high accuracy. We report an average ROC score of 0.997 and a ROC 1% score of 0.92 based on three repititions of a five fold cross validation procedure. A model selection study shows that our method does not strongly depend on the exact kernel function or kernel parameters. By training a classifier on different sized subsets of the data, we also show that we can achieve good accuracy with 1000 training examples and optimal performance with about 3000 training examples. We also demonstrate that we can accurately assess the probability of melting curve abberance given an SVM output. An analysis of the genomic features associated with aberrant melting curves suggests that a prominant cause of aberrance is overlap of the amplicon with a SINE repeat.

5.1. Model Selection

In order to avoid over-fitting the data, we partitioned the data randomly into two halves. Each half had the same proportion of positive and negative examples as the entire dataset. We used one half to develop features and choose SVM parameters. We then used a cross validation approach to evaluate the performance of the method on the other half of the data.

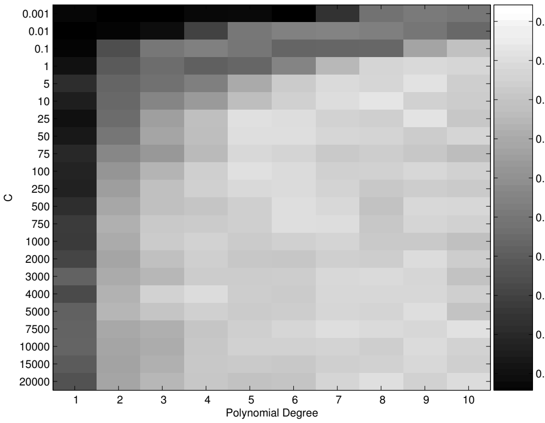

Figure 3. Heat map showing the results of a grid search using a polynomial kernel. Each pixel displays the mean of three different ROC 1% scores produced by five-fold cross validation for particular values of the polynomial degree and C parameter. The maximum ROC 1% score of 0.93 was achieved with a polynomial kernel of degree 7 and a C value of 10.

We used cross-validated ROC 1% scores to evaluate features and SVM parameters. First, we further partitioned the model selection data into five random sets, each with the same proportion of positive and negative examples as the entire dataset. Next, we used an SVM to classify the examples in each set after training the SVM on the other four data subsets. We repeated this entire process three times, and used the average ROC 1% score as a figure of merit.

One of the most important SVM parameters is the kernel function used to compare data examples. There are several kinds of similarity measures appropriate for real valued vector data. The most commonly used are polynomial kernel functions and radial basis functions (2). The polynomial kernel function has one significant parameter, the polynomial degree. The radial basis function also has one significant parameter, which controls the width of the radial basis function. In addition, the SVM learning algorithm has a parameter C that controls the cost of misclassifying training examples. In order to select a kernel, we performed a grid search for both the radial basis function and polynomial kernels. Our grid search procedure evaluated the 5-fold cross validation ROC 1% score as the parameters C and degree were varied for the polynomial kernel, and as the parameters C and radial basis function width were varied for the radial basis function kernel.

We found that the performance of radial basis function and polynomial kernels was similar, but that polynomial kernels had slightly better performance. Figure 3 shows that similar performance was obtained over a range of parameters for different polynomial degrees and values of C. Although the classifier performance was numerically better for some parameter values than others, the overall performance wasn't strongly dependent on the parameter values. For the final results, we used a polynomial kernel of degree 7 and set C to 10. For these parameters, we achieve a ROC 1% score of 0.93 on the training data and 0.92 on the testing data. In order to compare the SVM performance against another classification method, we employed the same model selection process on the training data using a K-nearest neighbor classifier (4), for identical values of polynomial degree and radial basis width while varying the number of neighbors used by the K-nearest neighbor classifier from one to ten. The K-nearest neighbor classifier did worse than the SVM, with a ROC 1% score of 0.89 on both the training and testing data.

5.2. Learning Curve

An important issue for practical application of this method is the determination of the number of PCR reactions necessary to train a successful classifier. To estimate the number of data points necessary to train a successful classifier, we performed cross-validation analysis on different sized subsets of the hold-out data. For training set sizes ranging from several hundred to several thousand examples, we chose 100 non-disjoint subsets of the appropriate size and performed 5-fold cross validation on each subset, recording the ROC 1% score as a measure of accuracy.

Figure 4 shows the learning curve, demonstrating that cross-validated ROC 1% scores of 0.9 are achieved with 3000 training examples. The slight upward trend of the curve as the training set size is increased suggests that further accuracy might be obtained by increasing the training set size, but that large amounts of additional data are likely to be required to achieve significant improvement.

5.3. Probabilities

Classification of data into aberrant and acceptable groups requires application of a threshold to the SVM outputs. The distribution of SVM outputs will vary from classifier to classifier, and depends on the training data set. In order to decouple the thresholding value from the exact classifier output distribution, we calibrate the SVM outputs to empirical probabilities. After this calibration, thresholds can be expressed in terms of probabilities and decided *a priori* based on the user's requirements.

We used the method described by (14) to calibrate the SVM outputs to empirical probabilities. Briefly, we use

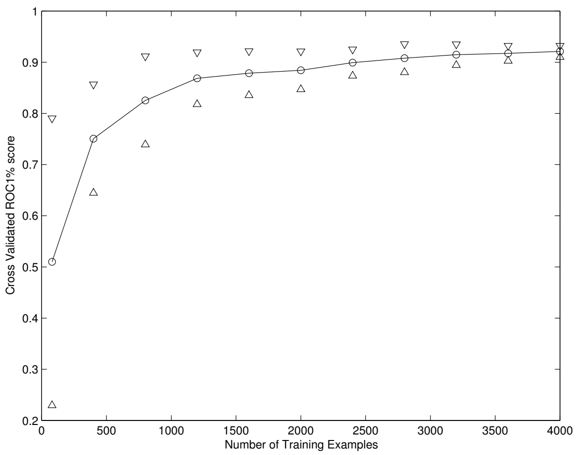

Figure 4. Learning curve showing cross-validated ROC performance as training data size is varied. The solid line with circles plots the mean of the ROC 1% scores, and the triangles show the mean ± one standard deviation. For each training set size, the cross-validated ROC 1% score averaged over 100 trials is plotted.This plot shows that learning saturates at about 3000 points.

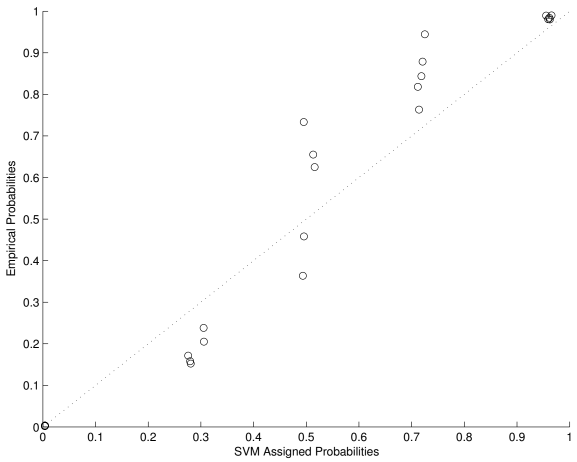

Figure 5. Empirical probabilities and threshold accuracy. (A) Comparison of empirical probabilities and sigmoided SVM outputs. The five sets of circles represent five individual five-fold cross validation runs. Each circle represents the mean value of the SVM decision function for a set of points plotted against the fraction of examples in that set that were labeled aberrant. This plot shows the agreement between the calibrated output of the SVM and the empirical probability of aberrancy.

a sigmoidal function $\sigma(x) = 1/(1 + e^{Ax+B})$ to map the SVM outputs to the interval (0,1). The parameters A and B are chosen by Levenberg-Marquardt optimization to maximize a log-likelihood function of the data. To assess the agreement between our calibrated SVM outputs and empirical probabilities, we performed five-fold cross validation, where two-thirds of the training data was used to train the SVM and one third was used to estimate values of A and B. We then sorted the data based on the assigned probabilities, and partitioned the data into disjoint and ordered subsets. For each subset, the mean assigned probability was plotted against the fraction of positives in that subset. Figure 5 shows the results of this comparison of the probabilities from five different five-fold cross-validation runs. Our experiments show that the estimates of aberrance probability yielded by the sigmoid function track well with the empirical aberrance probability. An accuracy of 99% was obtained for all cross-validation runs by setting the assigned probability threshold to 0.2.

5.4. Amplicon Analysis

In order to better understand the causes of aberrant melting curves, we analyzed the distribution of various genomic features within amplicons that were associated with aberrant and acceptable melting curves, respectively. This analysis showed that aberrant reactions are more likely to occur when the amplicons overlap with SINE (short interspersed element) repeats.

Our data set is derived from amplicons that spanned approximately 0.5 megabases of human genomic DNA on chromosome 11. Among the 1,728 distinct amplicons, 71 uniformly aberrant amplicons have all replicates flagged as aberrant and 1528 uniformaly acceptable amplicons have all replicates flagged as acceptable. The remaining amplicons have an intermediate number of acceptable replicates. We used the $1528+71$ consistently flagged examples for further analysis.

The GC content and the dinucleotide frequencies are essentially the same in the uniformly acceptable and uniformly aberrant amplicons. The abberant amplicons have an average GC content of 45 percent, and the acceptable amplicons have an average GC content of 44 percent. These GC contents are similar to the region on chromosome 11 from which the amplicons were derived, which has an overall GC content of 44 percent.

Next, we used the UCSC Human Genome Browser (9) to retrieve annotations for the genomic regions from which the amplicons were derived. The region of chromosome 11 covered by the amplicons contains 14 genes, six CpG islands, and 982 repeats flagged by the RepeatMasker program (19). The 14 genes contain 96 exons and 82 introns. The percentages of acceptable and aberrant amplicons falling within introns, exons, and CpG islands were similar. LINE repeats also had similar distributions. However, aberrant amplicons were enriched for overlap with SINE repeats. Of the uniformly acceptable amplicons, 74 percent were entirely disjoint from SINE repeats and approximately 20 percent partially overlapped SINE repeats. In contrast, only 50 percent of the uniformly aberrant amplicons were disjoint from SINE repeats, and about 46 percent partially overlapped SINE repeats. This situation was similar for simple repeats (such as $(TG)_n$ or $(CA)_n$). 95 percent of uniformly acceptable amplicons were disjoint from these simple repeats, whereas only 85 percent of the uniformly aberrant amplicons were disjoint from the simple repeats.

The association of aberrant amplicons with repeat elements is not surprising for two reasons. First, because one of the primers must bind to part of a SINE repeat, it will have many other strong binding sites in the genome, given the large number of SINE repeat copies (1). These other binding sites could lead to multiple amplicons being produced where multiple Alus are proximal. The second is that Alu sequences, which are a prominent subset of the SINE repeats, have high GC content at the 5-prime end. Amplicons that straddle the 5-prime end of an Alu could be expected to show a multi-modal melting curve if the flanking sequence was AT-rich due to the difference in GC content between the Alu sequence and adjoining region.

6. Discussion

We developed a method that can accurately distinguish aberrant PCR melting curves from acceptable melting curves. A critical advantage of this method, in addition to its speed, is that the outputs can be calibrated so the decision threshold can be specified without examination of the classifier's output distribution. Due to the sparseness enforced by the SVM optimization, our trained SVMs use relatively few of the training examples: usually around 100 of the training data are used to specify the separating hyperplane. Thus, the procedure produces low complexity models that can be expected to generalize well. Our classifier reliably separates aberrant melting curves from melting curves with one dominant peak.

Some PCR melting curves are expected to generate multiple peaks. For instance, multiplex PCR, in which multiple amplicons are amplified, would generate at least one transition for each amplicon. Another example is PCR genotyping, where a heterozygote at the amplified locus would generate two peaks, one for each amplified allele. The model could be extended for these types of RTPCR data by fitting multiple peaks; this would require a slightly more complex nonlinear optimization procedure and would introduce more features for analysis by the classifier.

Rejection of a melting curve as aberrant does not imply that the PCR amplicon was not successfully amplified. An amplicon that consists of an AT rich segment followed by a GC rich segment could be expected to generate a multi-modal melting curve, such as encountered for amplicons straddling Alu repeat elements. Alternatively, amplification of an undesired amplicon due to nonspecific primer hybridization or primer-dimer formation could also generate multi-modal signals in the melting curve even when the target amplicon is successfully amplified. However, when using RTPCR data to estimate initial template concentrations, the presence of primer-dimers or undesired amplicons will lead to inaccuracy. Thus, this approach is appropriate for flagging questionable data that is likely to cause errors in downstream analysis.

Our method trains a classifier to reproduce human judgments (which is the present gold standard). As such, the classifier may be subject to the same biases as the technicians who initially labelled the data. However, we have shown that if humans can reliably distinguish acceptable from aberrant curves, then so can our classifier. Furthermore, if humans make systematic errors leading to incorrect classification of melting curves based on a particular morphology, then our classifier can easily be adjusted by retraining.

In summary, this article presents a method for automated detection of aberrant PCR data based on melting curves. Our method is fast, accurate, and can be easily adjusted to account for new acceptability criteria, and is at as reliable as the judgments of the laboratory technicians who label the training data. The requirement for manual screening of sets of replicates represents a significant bottleneck in high throughput experiments. Our method relieves this bottleneck in developing and implementing high-throughput genomic assays using PCR.

Acknowledgments: The authors thank Paul Tittel for assistance with data transfer, Michael Dorschner for qPCR data generation, and Jun Neri, Kristen Lee, Molly Weaver, and Jeff Goldy, who initially labelled the melting curves.

References

[1] M. A. Batzer and P. L. Deininger. Alu repeats and human genomic diversity. *Nat. Rev. Genetics*, 3:370–380, 2002.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods*. Campbridge University Press, Cambridge, UK, 2000.

[3] M. O. Dorschner, M. Hawrylycz, R. Humbert, J. C. Wallace, A. Shafer, J. Kawamoto, J. Mack, R. Hall, J. Goldy,

P. J. Sabo, A. Kohli, Q. Li, M. McArthur, and J. Stamatoyannopoulos. Statistical mechanical simulation of polymeric dna melting with meltsim. *Nat. Methods*, 3:219–225, 2004.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[5] ENCODE Project Consortium. The encode (encyclopedia of dna elements) project. *Science*, 306:636–640, 2004.

[6] R. Haugland. *Handbook of Fluorescent Probes and Research Chemicals*. Molecular Probes Inc., Eugene, OR, 2001.

[7] R. Higuchi, G. Dollinger, P. Walsh, and R. Griffith. Simultaneous amplification and detection of specific DNA-sequences. *Biotechnology*, 10(4):413–417, 1992.

[8] M. A. Innis, D. H. Gelfand, and J. J. Sninsky. *PCR Applications: Protocols for Functional Genomics*. Academic Press, 1999.

[9] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome Res.*, 12:996–1006, 2002.

[10] W. Li, B. Xi, W. Yang, M. Hawkins, and U. K. Schubart. Complex DNA melting profiles of small PCR products revealed using SYBR Green I. *BioTechniques*, 35(4):702–706, 2003.

[11] D. Marquardt. An algorithm for least-squares estimation for non-linear parameters. *J. Soc. Ind. Appl. Math.*, 11:431–441, 1963.

[12] C. E. Metz. Basic principles of ROC analysis. *Semin. Nucl. Med.*, 8:283–298, 1978.

[13] W. Noble. Support vector machine applications in computational biology. In B. Schoelkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, USA, 2004.

[14] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelyhood methods. In P. Bartlett, B. Schoelkopf, D. Schuurmans, and A. Smola, editors, *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, USA, 1999.

[15] D. Poland. Recursion relation generation of probability profiles for specific-sequence macromolecules with long-range correlations. *Biopolymers*, 13(9):1859–1871, 1974.

[16] G. Raddatz, M. Dehio, T. F. Meyer, and C. Dehio. Primearray: genome-scale primer design for DNA-microarray construction. *Bioinformatics*, 17(1):98–99, 2001.

[17] P. J. Sabo, R. Humbert, M. Hawrylycz, J. C. Wallace, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl. Acad. Sci. U.S.A.*, 101(13):4537–4542, 2004.

[18] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H.A.Erlich. Primer-directed enzymatic amplification of DNA with a thermostable polymerase. *Science*, 239(4839):487–491, 1988.

[19] A. Smit, R. Hubley, and P. Green. Repeatmasker open-3.0, http://www.repeatmasker.org, 1996–2004.

[20] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.