APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2005, p. 8663–8676 Vol. 71, No. 12
0099-2240/05/$08.00+0 doi:10.1128/AEM.71.12.8663–8676.2005

Evaluation of Gel-Pad Oligonucleotide Microarray Technology by Using Artificial Neural Networks†

Alex Pozhitkov,[1] Boris Chernov,[2] Gennadiy Yershov,[2] and Peter A. Noble[1]*

University of Washington, Seattle, Washington 98195,[1] and Biodetection Technologies Section, Argonne National Laboratory, Argonne, Illinois 60439[2]

Received 21 April 2005/Accepted 12 July 2005

Past studies have suggested that thermal dissociation analysis of nucleic acids hybridized to DNA microarrays would improve discrimination among duplex types by scanning through a broad range of stringency conditions. To more fully constrain the utility of this approach using a previously described gel-pad microarray format, artificial neural networks (NNs) were trained to recognize noisy or low-quality data, as might derive from nonspecific fluorescence, poor hybridization, or compromised data collection. The NNs were trained to classify dissociation profiles (melts) into groups based on selected characteristics (e.g., initial signal intensity, area under the curve) using a data set of 21,044 profiles derived from 186 probes hybridized to a study set of RNA extracted from 32 microbes common to the human oral cavity. Three melt profile groups were identified: one group consisted mostly of ideal melt profiles; another group consisted mostly of poor melt profiles; and, the remainder were difficult to classify. Screening of melting profiles of perfect-match hybrids revealed inconsistencies in the form of melting profiles even for identical probes on the same microarray hybridized to same target rRNA. Approximately 18% of perfect-match duplex types were correctly classified as poor. Experimental variability and deviation from ideal melt behavior were shown to be attributable primarily to a method of local background subtraction that was very sensitive to displacement of the grid frames used for image capture (both determined by the image analysis system) and duplexes with low binding constants. Additional results showed that long RNA fragments limit the discriminating power among duplex types.

The human oral cavity is inhabited by possibly more than 500 microbial species, present either free in saliva or organized in complex multispecies biofilms attached to the surfaces of teeth and oral tissues (21, 34). The general composition of the oral microbiota has been studied using a variety of approaches, early culture-based studies more recently being complemented and extended by molecular characterization (16, 22, 25, 43, 50). However, established approaches are not suited to intensive and extensive monitoring. Molecular techniques such as clone libraries, quantitative PCR, and fluorescent in situ hybridization analyses, although informative, are labor intensive and impractical for routine monitoring. Thus, we anticipate that the development of tools that provide high-fidelity data in a high-throughput format will have significant utility, both as a diagnostic aid for oral diseases such as carries and periodontal disease and for identifying relationships to other disease states that may not be monocausal. DNA microarrays offer one type of highly multiplexed technology, using hybridization to multiple diagnostic sequences to identify different microbial populations or genes of functional significance.

Among the large variety of recent microarray technology platforms (see the review in reference 6), two formats have been most often used for microbial species identification: planar microarrays (51) and gel-pad microarrays (4, 12, 17, 20). Gel-pad microarrays, composed of oligonucleotides in a polyacrylamide gel-pad matrix attached to a glass surface, offer a number of advantages over planar microarrays because (i) they have a greater dynamic range due to the immobilization of a greater number of oligonucleotides (from 3 to 300 fmol) on the surface covered by each gel-pad (17), (ii) when used in combination with a temperature-controlled reaction chamber they can be employed to monitor arrays of probes that have different kinetics of association and dissociation (10), and (iii) when used under conditions that approximate equilibrium, thermodynamic analyses of probe-target duplexes in gel-pads have been reported to correlate to data obtained in solution (13), demonstrating a link to well-established principles of nucleic-acid chemistry (9, 29).

In conventional applications, fluorescent dye-labeled target nucleic acids hybridize to a short complementary oligonucleotide probe immobilized in a gel-pad, which yield stable duplexes under appropriate hybridization conditions. As the temperature of a microarray reaction chamber is increased, bound nucleic acid dissociates from probes, and there is an overall decrease in retained fluorescent dye (as inferred from signal intensity). Studies using synthesized targets (i.e., oligonucleotides) and fragmented nucleic acid extracted from microbes in environmental samples have demonstrated that it is possible to distinguish between target and nontarget sequences that differ by a single internal mismatched base-pair (12, 47, 48). This level of discrimination is needed to resolve variants of highly conserved genes (e.g., those encoding the rRNAs). Numerous studies have been conducted using gel-pad microarrays (8, 11, 12, 20, 23, 24, 26, 45, 47, 48, 49) because melting profiles of probe-target duplexes are thought to offer better discrimination between target and nontarget sequences than planar microarrays, which typically depend on signal intensity (SI) values

* Corresponding author. Mailing address: 201 More Hall, Civil and Environmental Engineering, University of Washington, Seattle, WA 98195. Phone: (206) 685-7583. Fax: (206) 685-7583. E-mail: panoble@washington.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

obtained following hybridization and wash conditions adjusted to an appropriate (average) stringency. (The term "melting profile," used throughout this article, refers to nonequilibrium dissociation curves.) Based on these studies, we proposed that nonequilibrium dissociation analyses, in concert with advanced statistical approaches, could be used to develop a diagnostic tool for identifying microorganisms in complex communities such as that found in the human oral cavity.

The focus of this study was to evaluate the performance of gel-pad microarrays under established conditions, which includes the same nucleic acid sample preparation protocols and concentrations, and image analysis software. The evaluation employed a set of DNA probes complementary to the rRNAs of major groups and species of microbes known to inhabit the human oral cavity. Preliminary screening of numerous melting profiles revealed inconsistencies in signal intensity values and discrepancies in the form of melting profiles, even for identical probes on the same microarray hybridized to the same target. However, our working hypothesis is that the general form of a melting profile in gel-pad microarrays (within the dynamic range of the detection system) should be sigmoid-shaped and not change from one duplex to the next, and that observed differences are due to experimental variations. Modeling the observed melting profiles using a conventional approach (i.e., fitting to a theoretical curve) (29) was not possible since melting profiles in gel-pad microarrays cannot be considered as reflecting an equilibrium process (see Discussion). Additionally, we tried to infer a function that best describes melting profiles. This function was able to explain only a small fraction of melting profiles.

Since no underlying model for interpreting gel-pad melting profiles exists, we sought to determine the sources of the inconsistencies by training artificial neural networks (NNs) to recognize patterns in the form of melting profiles. Artificial NNs can be used to recognize patterns in data such as the variability and shape of melting profiles and to classify melting profiles (e.g., ideal or poor) based on their melt characteristics (e.g., initial signal intensity, area under the curve). NNs are implemented as computer programs and consist of networks of neurons that receive information from inputs or other neurons, make independent computations, and pass their outputs to other neurons in the network (1, 3, 42). Once an NN is properly trained, the optimized weighting factors can be used to generate a model that provides information on the relationships among (input) variables such as melt characteristics and different types of melting profiles (outputs) such as those typical of perfectly matched duplexes versus those of duplexes containing multiple mismatches.

The objectives of this study were: (i) to develop and evaluate a software tool for calculating the variability and shape of melting profiles and (ii) to determine the main sources of inconsistencies that affect interpretation of melting profiles.

We report the development of a melting profile performance (MPP) calculator that correctly determined that ~18% of perfectly matched duplexes yielded highly variable (i.e., poor) melting profiles. Further experiments revealed that the main sources of inconsistencies contributing to the poor melting profiles were: the placement of the grid frames and the background subtraction method used by the image analysis system. Moreover, duplexes with low binding constants had highly variable melting profiles because SI values approached the detection limit of the system.

MATERIALS AND METHODS

RNA preparation. RNA was either isolated from the log phase grown cultures or in vitro transcribed from a cloned rRNA gene (Table 1). Isolation was performed using the FastRNA BLUE kit (Q-BIOgene, Irvine, CA), following the manufacturers instructions by bead beating (two times for 30s each), phenol chloroform extraction, and isopropanol precipitation as previously described (44). To produce in vitro transcribed RNA, a T7 RNA polymerase kit (Invitrogen, USA) was used. The RNA was subjected to fragmentation and labeling with lissamine-rhodamine B ethylenediamine dye according to the previously published protocol, with slight modifications (4, 12). Briefly, 10 μg of total RNA was preheated at 95°C for 4 min in a 1.5 ml reaction tube. Freshly prepared labeling cocktail (150 μl) containing 5 mM 1,10-phenanthroline, 0.5 mM $CuSO_4$, 1 mM lissamine-rhodamine B ethylenediamine (Molecular Probes, Inc., Eugene, Oreg.), and 20 mM sodium phosphate (pH 7.0) was heated at 95°C for 30s. Hydrogen peroxide (2 mM) and freshly prepared 20 mM $NaCNBH_3$ were added to the cocktail, and the mixture was added to the reaction tube. After incubation of the mixture for 10 min at 95°C, the reaction was stopped by adding 9 μl of 500 mM EDTA (pH 8.0). Fragmented nucleic acids were precipitated by adding 15 μl of 5 M ammonium acetate and 450 μl of 100% (vol/vol) ethanol followed by a 10 min incubation at −80°C. Nucleic acids were recovered by centrifugation at 13,200 rpm for 10 min, excess fluorescent label was removed by washing twice with 500 μl of 100% (vol/vol) ethanol, dried, and resuspended in 20 μl of diethyl pyrocarbonate-treated water.

For the experiments involving the affects of image processing and diffusion on melting profiles, and the affects of different binding constants on SI values, target oligonucleotides (Supplementary Table S1) were 5′-labeled with Oregon Green (QIAGEN, Valencia, CA) while native rRNA target sequence from *Bacteroides forsythus* was fragmented and then randomly labeled with Cy3. Briefly, 4 μl aqueous solution of native RNA (2 μg/ml, isolated the same way as above) was hydrolyzed in 2 μl of 0.1 M NaOH for 5 min at 35°C, neutralized by adding 2 μl of 0.1 M HCl to the reaction mixture, and labeled using the Micromax ASAP RNA labeling kit (Perkin Elmer, Boston, MA). The length of RNA fragments was determined by using BioAnalyzer (Agilent Technologies, USA).

Oligonucleotide array fabrication. Oligonucleotides were designed by using the probe design function of ARB software (http://www.arb-home.de) (27). The specificity of the probe for the target was checked with the probe check function in the ARB software, the BLAST search (2) at the National Center for Biotechnology Information, and the Probe Match program in Ribosomal Database Project II (28). Self-complementarities were also examined by using Ribosomal Database Project II. Oligonucleotide probes, ranging in length from 13 to 25 nucleotides (Table S1 in the supplemental material), were synthesized with an amino linker at the 3′-end and fabricated at Argonne National Laboratory (52). The microarray matrix containing polyacrylamide gel pads (100 by 100 by 20 μm) spaced 200 μm apart from each other and fixed to a glass slide, was manufactured by photopolymerization procedure (52). A total of 3 nl of 1 mM amino-oligonucleotide solution was applied to each gel element containing aldehyde groups (45) which were designed and implemented by a robot arrayer (52). A total of 186 oligonucleotide probes were immobilized with the aldehyde group of the activated gel pad on the microarrays as described previously (40).

Hybridization and washing protocols. Hybridizations were carried out at room temperature (20°C) for 12 h in 40 μl of hybridization buffer containing 5 to 10 μg of each target RNA, 0.9 M NaCl, 20 mM Tris-HCl (pH 8.0), and 40% formamide. Following overnight hybridization, the microarray was washed three times at room temperature with a washing buffer consisting of 20 mM Tris-HCl (pH 8.0), 5 mM EDTA, 4 mM NaCl and 1% wt/vol Tween 20. After the final wash, 200 μl of washing buffer was added to the imaging chamber (Grace BioLabs, Bend, OR) for image and melting profile capture.

Image and melting profile capture. To generate melting profiles, the microarray was fixed on a thermotable mounted on the stage of a custom-designed epifluorescence microscope (State Optical Institute, St. Petersburg, Russia) and connected with a thermoelectric temperature controller (LFI-3735; Wavelength Electronics, Inc. Bozeman, MT) and a water bath (Cole Parmer Instruments Co., Chicago, IL). The microscope was equipped with fluorescence filters (Omega Optical, Brattleboro, VT) and a cooled charge-coupled device camera (Princeton Instruments, Trenton, NJ) and manipulated with a program that allows image acquisition, processing, and analysis (13). Melting profiles for all probe-target duplexes were monitored and recorded at 2°C intervals between 20 and 70°C by increasing the temperature at a rate of 1°C per min. The melting profile experiments were performed in triplicate and repeated on different days. We visually

TABLE 1. Organisms tested[a]

Data set no.	Microbial species	GI no.[b]	Source	No. of replicated microarray experiments
1	*Abiotrophia defectiva*	1834295	ATCC 49176[c]	4
	Acinetobacter baumannii	829087	ATCC 19606[c]	6
	Actinobacillus actinomycetemcomitans	173681	Strain JP-2[d]	3
	Actinomyces odontolyticus	853707	ATCC 17929	3
	Bacteroides forsythus	10946530	ATCC 43037	2
	Brevundimonas diminuta	2580430	ATCC 11568[c]	4
	Butyrivibrio fibrisolvens	15011532	ATCC 19171	3
	Enterococcus faecalis	5578753	ATCC 19433[c]	6
	Escherichia coli	174375	TOP10[e]	2
	Fusobacterium nucleatum	4490387	ATCC 25586[d]	3
	Peptococcus niger	45659	ATCC 27731[c]	3
	Peptostreptococcus anaerobius	175621	ATCC 27337[c]	2
	Porphyromonas endodonthalis	294287	ATCC 35406	4
	Prevotella denticola	294420	ATCC 35308	4
	Propionibacterium freudenreichii	45491	ATCC 6207[c]	4
	Ralstonia eutropha	23821283	ATCC 17697[c]	5
	Staphylococcus aureus	576603	ATCC 25923[c]	2
	Streptococcus bovis	176044	ATCC 9809[c]	5
	Streptococcus gordonii	2183315	DL1[d]	4
	Streptococcus mutans	5578899	ATCC 25175[c]	5
	Streptococcus salivarius	176047	ATCC 25975[c]	5
	Treponema denticola	3712666	ATCC 35405	2
2	*Candida albicans*	2507	Ted White[f]	3
	Candida parapsilosis	17266284	Ted White[f]	3
	Capnocytophaga gingivalis	289582	ATCC 33624	2
	Desulfovibrio vulgaris	18034282	Judy Wall[g]	3
	Gemella haemolysans	174677	ATCC 10379	2
	Haemophilus paraphrophilus	174772	Strain C128[d]	3
	Porphyromonas catoniae	929751	ATCC 51270	2
	Porphyromonas gingivalis	294288	ATCC 33277[d]	1
	Rothia dentocariosa	175870	ATCC 17931[c]	2
	Treponema sp.	2586374	Richard L. Lamont[d]	2

[a] rRNA was extracted from the following microorganisms and used to hybridized to the DNA microarrays. Shown are the corresponding GI numbers and the number of replicated microarray experiments for each microorganism.
[b] genInfo identifier; a unique integer assigned by National Center for Biotechnology Information which identifies a particular sequence.
[c] Strains were kindly provided by Paul W. Lepp, Stanford University, California.
[d] Strains were kindly provided by Richard L. Lamont, University of Florida, Gainesville, Florida.
[e] Invitrogen Inc.
[f] Ted White, Seattle Biomedical Research Institute, Seattle, Washington.
[g] Judy Wall, University of Missouri, Columbia, Missouri.

inspected each microarray for gross artifacts prior to recording melting profiles. In total, 13 microarrays were used in this study ($n = 104$ experiments) and each microarray was reused multiple times. Analysis of variance of initial signal intensities and MPP scores of universal probes revealed no significant differences by microarray lot or the number of times the microarray was reused.

To assess the precision and accuracy of the image acquisition system, an alternative approach was employed by creating a custom-designed module in V^{++} (Roper Scientific, Germany). The module controlled the camera, the microscope, and the Peltier element, and recorded the image at each temperature. An additional custom-designed program was created in C^{++} to convert the stack of images to SI values for each gel-pad on the microarray.

In silico prediction calculator. The in silico Prediction calculator (http://noble.ce.washington.edu/programpage.jsp) evaluates probe and target sequence using lexicographical matching to yield information on the type of probe-target duplex (i.e., perfectly matched, P; terminal mismatch, T; internal mismatch, I; more than two mismatches, N). The position and type of the mismatch (e.g., A, T, G, and C) was determined for the duplex structure yielding the best match (highest hybridization score). The hybridization score of a probe to a target sequence was determined by counting the number of correctly base-paired nucleotides for a probe at every possible position in a target sequence, starting at the 5′-end of a target sequence, and moving the probe along the target sequence, one nucleotide position at a time until the end of the sequence was reached.

T_d calculator. The T_d calculator was designed to automatically calculate the experimentally determined T_d and slope for each probe-target duplex by using data obtained from the image acquisition, processing, and analysis software. A Web based interface for this software is available at http://noble.ce.washington.edu/programpage.jsp. The interface contains a Readme documentation describing how to use the software. The documentation contains links to demonstration files (that can be submitted to the calculator) and specifies required formatting for input files.

Since calculating the T_d using a simple curve fitting yielded inconsistent results (presumably due to factors affecting the melting profiles, *see Discussion*), a new version of the T_d calculator (47) was designed to determine the temperature corresponding to the maximum slope of the signal intensities, which was presumably related to the transition from duplex to random coil. Multiple regressions lines, consisting of various number of points, were used to determine the maximum slope. To maximize the number of points used for regression analyses, normalized SIs, i.e., $X_{norm} = (X_{obs} - \min)/(\max - \min)$ in the range of 0.75 to 0.55 arbitrary units were used as the initial start points for regression lines, and a lag of 0 to 5 points was used to vary the endpoints, and thus the length of the regression lines. The minimum number of points used to calculate a regression line was 5. The program considered all possible slopes and calculated the mean slope and T_d for all slopes meeting the following two criteria: (i) the calculated T_d was within the mean temperature (±1.5°C) of normalized SI values between 0.35 to 0.65, and (ii) the calculated T_d fell within a temperature range based on the normalized SI values of 0.55 and 0.45. The default values for the T_d calculator were experimentally determined.

For each melting profile, the number of regression lines meeting these criteria, the initial SI, the T_d, and the slope of SI values and temperature (dSI/dT), the

area under the curve before min/max normalization, and the normalized area under the curve (i.e., after min/max normalization and zeroing) was recorded.

Data sets used for statistical analyses. Melting profile characteristics and corresponding in silico predictions of each gel-pad experiment was merged to a single data set. Data set 1 consisted of melting profiles obtained from 22 known target sequences (Table 1), 186 probes targeting 16S and 23S rRNA (Table S1 in the supplemental material), and 81 hybridizations and melting runs, and included 1,017 perfectly matched (P) probe target duplexes, 110 probe target duplexes containing a terminal mismatch, 2,269 duplexes containing one or two internal mismatches (I), and 12,188 duplexes containing more than two mismatches (N). Data set 2 consisted of melting profiles obtained from 10 known target sequences (Table 1), 186 probe targeting rRNA genes (Supplementary Table S1), and 23 hybridizations and melting runs, and included 276 perfectly matched (P) probe target duplexes, 21 probe target duplexes containing a terminal mismatch, 604 duplexes containing one or two internal mismatches (I), and 4,559 duplexes containing more than two mismatches (N).

NN software. An artificial NN package was developed for this project (Neuroet) (38, 46) and is available at the web site http://noble.ce.washington.edu/Neuroet. Unless otherwise specified, the following settings were used for training NNs: input and output scaling was set to standard linear (0, 1); the logistic transfer function was used for hidden neurons and pure linear transfer function was used for output neurons; 80% of the data was used for training, 10% was used for testing, and 10% was used for validating the NN; and, conjugate gradient error minimization was used as the training method.

The architectures of all NNs were optimized prior to conducting analyses by adjusting the number of hidden neurons (1 to 8) and identifying the architecture that provided the best predictive model. Comparison of different predictive models was conducted by computing their median Akaike's Information Criterion corrected (AICc) value (35), determining the probability that one model was better than another, and calculating the corresponding evidence ratio. The AICc value was used (rather than R-squared) because it balances the complexity of an NN model (i.e., the number of data records and the number of variables [i.e., input variables and number of hidden neurons]) with how well the NN predicts outputs. The model yielding the lowest AIC_c score contained the optimal number of hidden neurons. AIC_c was calculated using the following equation:

$$AIC_c = N\ln\left(\frac{SS}{N}\right) + 2K + \frac{2K(K+1)}{N-K-1} \quad (1)$$

where N is the number of data records, K is the number of input variables used plus 1, and SS is the sum of squares of the residuals (predicted scores versus actual scores) (19).

The probability (P_{model}) that one NN model was likely to be more correct than another was determined using the following equation:

$$P_{model} = \frac{e^{-0.5\Delta AIC_c}}{1 + e^{-0.5\Delta AIC_c}} \quad (2)$$

The evidence ratio (E) was used to assess how more likely one model was to be correct than another model. E was determined using the following equation:

$$E = \frac{P_{model1}}{P_{model2}} = \frac{1}{e^{-0.5\Delta AIC_c}} \quad (3)$$

If, for example, the AICc scores of two models differed by 5.0, then E equals 12.2, meaning that the model with the lower AICc score was about 12 times more likely to be correct than the other model. However, if AICc score differ by 10, then E is 148, so the evidence is overwhelmingly in favor of the model with the lower AICc score.

Identifying the most important inputs for predicting outputs. The relative contributions of inputs to predicting outputs were determined by using the Measure Importance of Inputs procedure in the Neuroet package (38, 46). Fifteen NN models were generated for each input variable (e.g., SI value). Models that fell into the lower 25th percentile based on their ranked SS were removed from the analysis because they were considered fixed in local error minima. The AICc scores of the remaining 11 models were averaged. The procedure was then repeated for each input variable. The AICc scores of the inputs were ranked by their value. The probability score and evidence ratio of the ranked inputs (discussed above) were calculated to determine the probability that one input (or a combination of inputs), was better than another.

Developing the Quality and Shape calculator. Melting profiles of 4800 probe-target duplexes from data set 1 were manually scored for their Quality and Shape values by comparing the profiles to predetermined standards as shown in Fig. 1. The Quality scoring was based on the disjointedness of adjacent points in the melting profile. A "smooth" melting profile was scored 0 (Fig. 1A and 1C) while a melting profile that had one to several disjointed points was scored 0.25 to 1.0 (Fig. 1B, 1D-F), respectively, depending on the amount of noise. The Shape scoring was based on the overall shape of the melting profile: an ideal shape was assigned a score of 0, a questionable shape was assigned a score of 0.5, and a random or not-interpretable shape was assigned a score of 1. The final data set used for training and testing the NN consisted of a subset of 1500 records (from the 4,800 melting profiles; i.e., 500 ideal, 500 uncertain, and 500 poor melting profiles) and 500 records that were generated with random numbers having a range of 0 to 1.

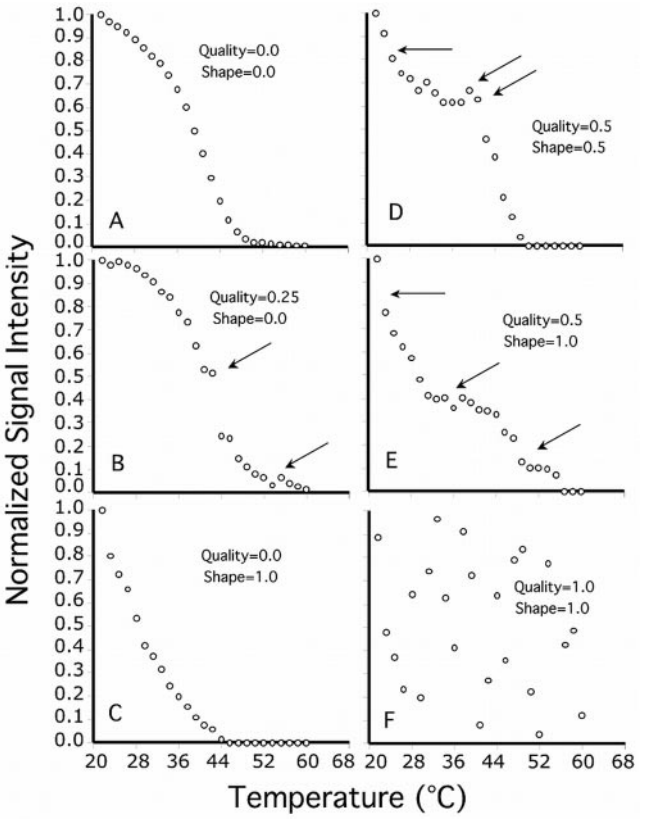

FIG. 1. Example criteria used to classify melting profiles. A to E, Actual melting profiles; F, Random profile. Quality criterion: A and C have low scores because they are smooth; B and D to F have disjointed points as indicated by the arrows and have high scores. Shape criterion: A and B have low scores because they have ideal shapes.

For each melting profile, the SI values were normalized to a maximum value of 1 and a minimum of 0. The first 25 SI values were used as input data to train NNs to predict Quality scores while the first 25 SI values and the Quality scores were used as input data to predict Shape scores. Hence, the Shape score considered both the disjointedness of adjacent points as well as the shape of melting profiles.

Quantifying the effects of noise on Quality and Shape scores. To determine the effects of noise on Quality and Shape scores, 10 melting profiles that were considered ideal by visual interpretation were selected from a single microarray experiment. Fixed amounts of randomly generated noise were then added to each of the ten profiles at levels of approximately 0, 0.1, 0.5, 1.0, 5.0, 10, and 50%. The Quality and Shape scores were computed for each melting profile.

Multivariate statistical analysis. Principal-component analysis (PCA) was employed to examine the distribution of melt characteristics relative to duplex types (e.g., P, I, or N) and to construct ordination plots.

Thermodynamic calculations. The thermodynamic properties (ΔG^0) of each perfect-match duplex were calculated using OligoAnal (31). With exception to temperature and oligonucleotide length, we used all default parameters. For our calculations, the temperature was set to 20°C and oligonucleotide length was adjusted accordingly.

TABLE 2. *R*-squared values of observed versus predicted scores

Variable	R^2 by data set (n = 2,000)		
	Training (n = 1,600)	Testing (n = 400)	Validation (n = 400)
Quality 1	0.82	0.81	0.78
Quality 2	0.82	0.81	0.78
Shape 1	0.89	0.90	0.93
Shape 2	0.89	0.89	0.89

RESULTS

Quality and Shape scores of melting profiles. Results obtained by using the Optimize hidden neurons procedure in the Neuroet package revealed that the optimal number of hidden neurons needed to predict Quality and Shape scores was two (data not shown). The NNs accounted for approximately 78 to 93% of the variability in the data, indicating that they provided reasonable predictions (Table 2). The concordance across rows of Table 2 (i.e., among training, testing, and validation data sets) indicated that none of the NNs were under- or overtrained, and that the architecture of the NNs was appropriate for the data. Note that there was little difference in the R-square values of Shape scores indicating that having two Quality score as inputs (Shape 2) rather than one (Shape 1) did not substantially improve shape predictions. The equations of the trained NNs were extracted, checked for accuracy using a spreadsheet (MS Excel), incorporated into a C^{++} program, and made into a web-accessible tool (i.e., Melt Quality and Shape Calculator, http://noble.ce.washington.edu/programpage.jsp).

Figure 2 shows the relative importance of NN inputs for predicting Quality and Shape scores. Sudden increases in P_{model} values (and decreases in E ratios) of the ranked inputs were used as limits for interpreting the importance of inputs (data not shown). There were no differences in the importance of inputs between the two Quality scores indicating that different NNs yielded consistent results. The first 8 SI values, representing SI values at 20 to 34°C, the 11th SI value, representing 40°C, and the last SI value (i.e., input 25), representing the SI value at 68°C, were more important for predicting Quality scores than SI values at other temperatures (e.g., 36, 38°C, and 42 to 66°C) (Fig. 2, upper panel). These findings indicated that Quality scores were based on SI values at the beginning, middle, and end of the melting profile. However, this was not the case for Shape scores (Fig. 2, lower panel), since the first nine SI values, representing SI values at 20 to 36°C, and the Quality score(s) used as NN inputs were more important for predicting the Shape scores than the SI values at other temperatures (e.g., 38 to 68°C). These findings indicated that the beginning of the melt and the Quality score(s), as interpreted from the disjointedness among adjacent SI values, are important for predicting the shape of melting profiles.

To quantify the amount of noise needed to affect Quality and Shape predictions, incremental amounts of random noise were added to ideal melting profiles. As anticipated, adding noise to melting profiles increased the value and variability of the Shape and Quality scores (Fig. S1 in the supplemental material). These experimental results indicate that Quality and Shape scores derived from NNs were able to correctly classify the quality and shape of melting profiles.

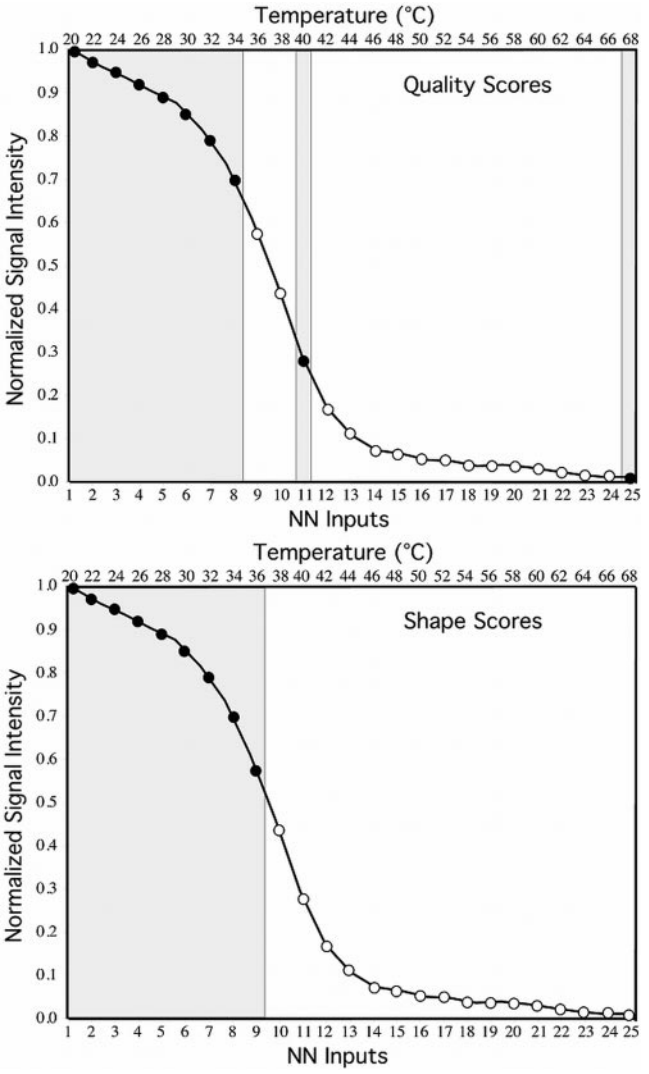

FIG. 2. Importance of NN inputs to predict Quality (top panel) and Shape (bottom panel) scores relative to an idealized melting profile. The relative importance of each input was based on rank order and their corresponding P_{model} and E ratio values (see text). Shaded areas (and solid black circles) indicate inputs that were statistically more important than the other inputs (white circles). Note that for Shape scores (lower panel), NN inputs 26 and 27 were also found to be important for predicting Shape scores. Inputs 26 and 27 correspond to Quality 1 and 2 scores, respectively.

Equations defining the relationship between SI values and Quality and Shape scores were extracted from the trained NNs and incorporated into an application on the web. The Melt Quality and Shape Calculator and documentation files are available at http://noble.ce.washington.edu/programpage.jsp.

Melting profile performance (MPP) calculator. Although the Melt Quality and Shape Calculator provides information on variability and shape of melting profiles, it does not consider other melt characteristics such as the area under the curve or initial SI values. To account for these other melt characteristics, we developed a calculator that predicts melting profile performance using melt characteristics shown in Table

TABLE 3. Correlation coefficients of variables relative to PCA axes (n = 2,928)

Variable	Pearson correlation by PC axis:	
	PC1	PC2
Initial signal intensity	−0.74	0.54
True signal intensity area	−0.67	0.70
Normalized signal intensity area	−0.70	0.07
Number of regressions used	−0.73	−0.07
Quality 1	0.86	0.38
Quality 2	0.83	0.40
Shape 1	0.95	0.16
Shape 2	0.91	0.14

3 as inputs and using information from PCA as a guide to classify ideal from poor melting profiles. Note that ideal and poor terms are not the same as high and low scores for Quality and Shape. A balanced data set was constructed from data set 1 and analyzed by PCA. The balanced data set consisted of equal number (n = 976) of perfectly matched (P) duplexes, duplexes containing internal mismatches (I), and duplexes containing more than two mismatches (N) extracted from data set 1. The following characteristics were used for PCA: (i) initial signal intensity, (ii) true signal intensity area (i.e., the area under the curve without normalization), (iii) normalized signal intensity area (i.e., area under the curve with min/max normalization), (iv) number of regression lines used to estimate the dissociation temperature, (v) Quality scores 1 and 2, and (vi) Shape scores 1 and 2.

Results from PCA revealed that 81% of the total matrix variance was explained by two principal axes, with PCA1 explaining 65% of the total matrix variance and was correlated strongly with the following variables: initial signal intensity, the normalized area under the curve, number of regressions used and Quality and Shape scores (Table 3), and PCA2 explaining 16% of the total matrix variance, was strongly correlated to the area under the curve (Table 3). An ordination plot revealed that one large group distributed along the principal component 1 (PCA1) had considerable coherence. This coherence was further examined by dividing the ordination plot into subplots based on duplex type (e.g., P, I, N) (Fig. 3). Most of the P-type (84.2%), some of the I-type (49.7%), and few of the N-type 13.0%) melting profiles occurred on the left side of −2.5 on the x axis. Melting profiles in this region had low Quality ($X\pm$ Std; 0.08 ± 0.06) and Shape scores (0.05 ± 0.07) indicating that they were suitable for statistical interpretation. Few of the P-type (12.1%), some of the I-type (45.6%), and most of the N-type (81.0%) melting profiles occurred on the right side of −2.0 on the x axis. These profiles had high Quality (0.75 ± 0.10) and Shape (1.0 ± 0.01) scores indicating that they are highly variable. These results indicate that the position of melting profiles on the ordination plot was related to melt characteristics of the duplex types (e.g., P, I, and N).

The melting profile performance calculator assigned each melting profile a score based on its distribution in the ordination plot and, in some cases (e.g., depending on its position) by examining the melt characteristics of individual duplexes. An MPP score of 1 was manually assigned to the cloud of profiles distributed between approximately −1.5 and −6.0 in the x axis and −2.0 and 11.0 in the y axis (Fig. 3). Melting profiles immediately to the right and/or surrounding the cloud were assigned a MPP score of 0.5 (since they were difficult to classify them as either 1 or 0), and the remaining profiles were assigned a score of 0. An NN was trained to establish the relationship between the eight melting profile characteristics (initial SI, area under the curve, normalized area under the curve, number of regression lines used to calculate the T_d, and Quality and Shape scores) and their corresponding MPP score.

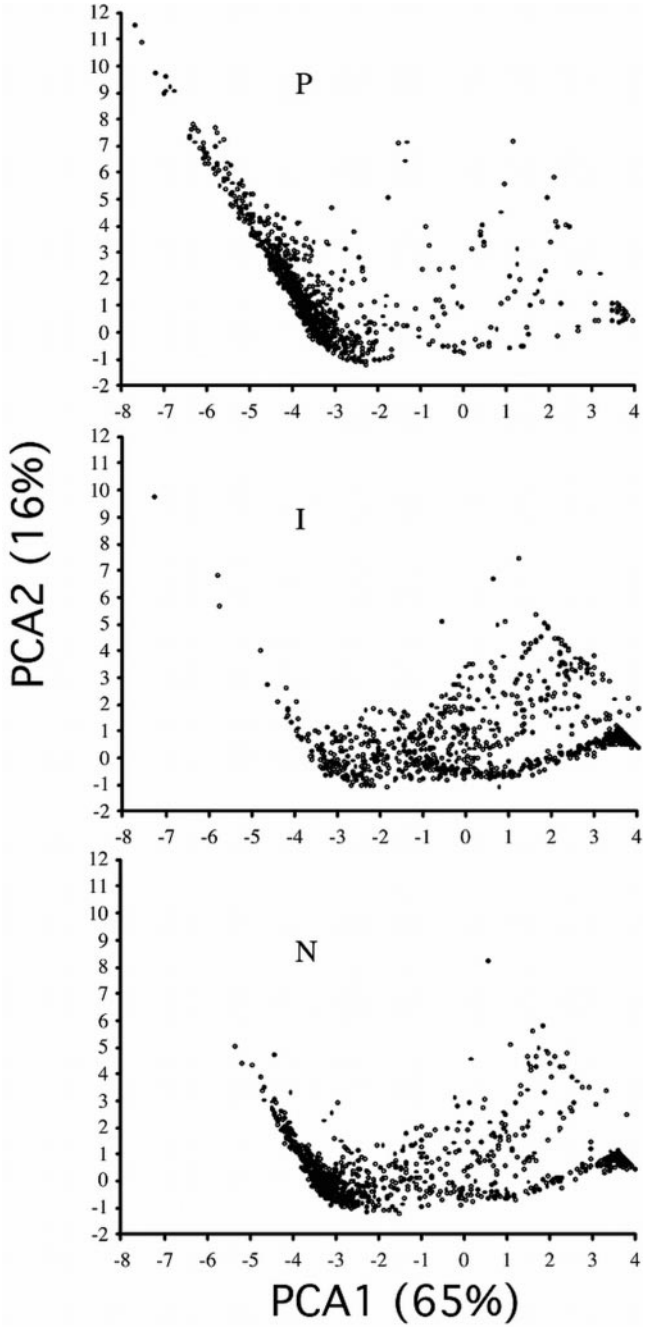

FIG. 3. Ordination plots produced by PCA of melting profile variables. P, perfectly matched probe target duplexes; I, duplexes containing an internal mismatch; N, duplexes containing more than two mismatches.

The optimal number of hidden neurons was determined by training NNs (multiple times) using one to eight hidden neurons, and identifying the architecture yielding the lowest AICc score. The lowest AICc score occurred for NNs having five hidden neurons, thus the optimal architecture used to train the NN was eight inputs (one for each variable), five hidden neurons, and one output neuron (i.e., one for the MPP score). Eighty percent of the data were used to train the NN, 10% of the data were used to test the NN and the remaining 10% were used to validate the NN. The final R-squared values of train, test, and validation data sets were close to 1 (e.g., 0.96), indicating considerable similarity. We determined the importance of melting profile characteristics to predict MPP scores and found that the initial SI value, the number of regression lines used to calculate the T_d, and the Shape1 score provided the best combination of characteristics to predict the MPP score, accounting for approximately 90% of the variability (data not shown). The weights and biases extracted from the NN were used to calculate the MPP score from melting profile characteristics.

The relationship between the predicted MPP score and positions on the ordination plot for the balanced data set are shown in Fig. 4 (upper panel). Approximately 48.5% of the balanced data set was classified as ideal melting profiles, 5.7% could not be classified, and the remaining 45.6% was classified as poor (Fig. 4, lower panel).

The relationships between MPP and duplex type are shown for data sets 1 and 2 (Table 4). Both data sets contained disproportionately more Is and Ns than the balanced data set. The MPP scores showed a general trend that a greater number of perfectly matched probe-target duplexes were classified as ideal than duplexes with two or more mismatches (from left to right in Table 4). Similar results were obtained for both data sets indicating that the balanced data set used to develop the equations relating melt characteristics to MPP scores was able to generalize predictions.

Figure 5 shows MPP calculator results for perfectly matched duplexes involving probes 62 (Univ 1390) and 438 (S-P-Grpos-1200-a-A-13). For probe 62, 89.5% (145/162) were classified as ideal, 3.1% (5/162) were classified as uncertain, and 7.4% (12/162) were classified as poor. For probe 438, 33.7% (33/98) were classified as ideal, 6.1% (6/98) were classified as uncertain), and 60.2% (59/98) were classified as poor. Not all melting profiles are shown in Fig. 5 for clarity. Relative to probe 62, the initial SI values of probe 438 (perfect-match duplexes) were consistently low (Fig. 6), indicating that (in general) probes with low binding constants tended to be classified as having poor melting profiles by the MPP calculator. Conversely, probes with high binding constants tended to have ideal melting profiles (Fig. 6). Negative SI values are due to the method used for background subtraction (i.e., Fotin et al. [13]; discussed below).

To resolve the relationship between initial signal intensities of probes and their corresponding binding constants, we calculated a proxy for the binding constant, i.e., the ΔG^{0}_{20} for all perfect-match duplexes. Figure 7 shows the relationship between initial SI values and ΔG^{0}_{20} for duplexes that were replicated at least 40 times. Approximately 83% of the variability in the data was explained by mean initial SI values and ΔG^{0}. Note that probe 438 had a higher ΔG^{0} (−21.8 kcal/mol) than probe 62 (−33.8 kcal/mol), which supports the notion that duplexes with high binding constants (negative ΔG^{0}_{20}) tend to have ideal melting profiles.

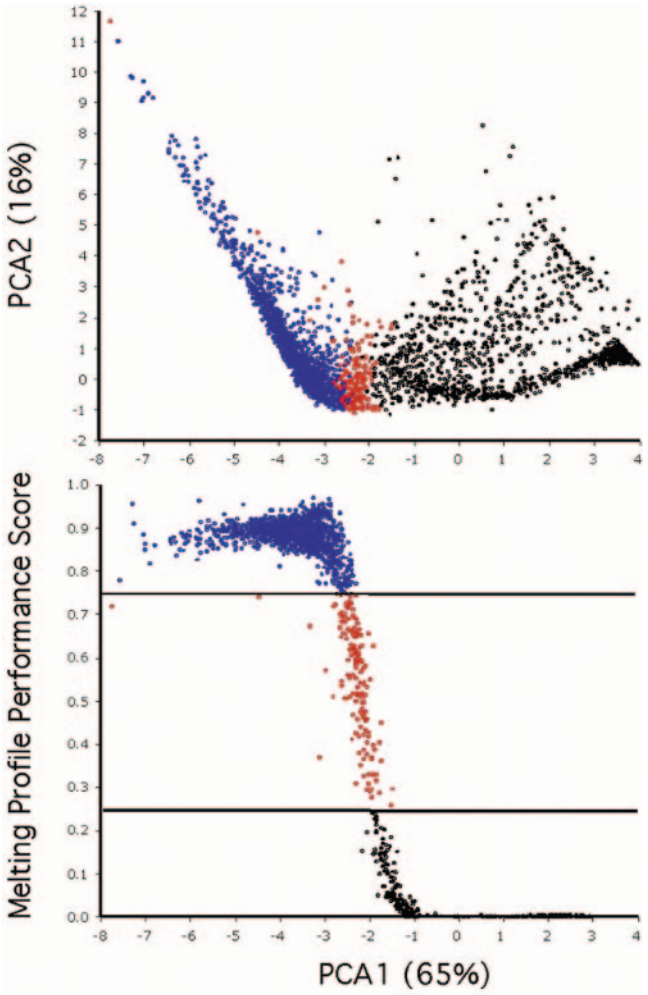

FIG. 4. Melting profile performance scores by position on the ordination plot. Top panel, two principal components; bottom panel, absolute duplex performance scores (predicted by MPP calculator) relative to PCA1. Blue dots represent melting profiles with duplex performance scores of >0.75; red dots represent profiles with scores between 0.25 and 0.75; black dots represent profiles with scores of <0.25.

The distribution of initial SI values of duplexes with two or more internal mismatches (from data set 1) is shown in Fig. 8. Most of these duplexes were classified as having poor melting profiles (82.1%, Table 4) and low initial SI values, indicating that they were close to the detection limit of the system. Of the 14.9% (Table 4) in data set 1 that were classified as having ideal melting profiles, most duplexes had initial SI values that were comparable to those of perfect-match duplexes (Fig. 6). Lexicographical analysis of these probes did not reveal any particular aspect of their sequence (e.g., length, GC content) which accounted for the preponderance of ideal melting profiles (data not shown).

These results are consistent with: (i) our previous finding that the initial SI values are one of the three critical factors used by the MPP calculator to classify melting profiles, and (ii)

TABLE 4. Classification of melting profile performance by duplex type

Data set	Melting profile performance	% Classification type[a] by probe target duplex type (no. of samples):				
1		P (1,017)	T (110)	I1 (1,185)	I2 (1,084)	N (12,188)
	Ideal	85.8	87.3	72.0	43.2	14.9
	Undefined	1.7	2.7	2.0	4.2	3.0
	Poor	12.5	10.0	26.0	52.7	82.1
2		P (276)	T (21)	I1 (322)	I2 (282)	N (4,559)
	Ideal	66.3	42.9	56.2	33.7	19.4
	Undefined	2.9	0.0	5.9	10.3	5.8
	Poor	30.8	57.1	37.9	56.0	74.8

[a] P, perfect match; T, terminal mismatch; I1, one internal mismatch; I2, two internal mismatches, N, more than two mismatches.

the notion that melting profiles with low initial signal values approach the detection limit of the system, and for this reason, produce inconsistent results which are difficult to interpret.

Effects of image processing on melting profiles. Given the significant number of perfectly matched duplexes yielding poor melting profiles, we investigated the effects of image processing on melt profiles by comparing three background subtraction methods. Raw image stacks (n = 120 images) were collected from thermal dissociations (20°C to 70°C) of three different oligonucleotide duplexes. Each image in a stack represented the SI value of a probe-target duplex collected at one temperature. To simulate subtle variations in placement of the frame used to collect averaged image SI values from the gel-pad, 20 different frame placements of the same image were produced by randomly displacing the x and y coordinates of the frame 20 times (i.e., a change of ±4 pixels representing a variation of 0 to 6.7% total gel-pad area). This displacement never removed the gel pad from the inner frame. All three background subtraction methods were applied to the same 20 modified image stacks.

The Fotin et al. (13) background subtraction method involves subtracting background using the average SI values immediately surrounding the frame and can be defined by the equation:

$$\mathrm{SI} = \frac{I_{\mathrm{inner}} - B_{\mathrm{outer}}}{B_{\mathrm{outer}}} \quad (4)$$

where I_{inner} refers to the averaged intensity of the inner frame (180 μm by 180 μm) and the B_{outer} refers to the averaged intensity of the space between the inner and outer frames (230 μm by 230 μm) (Fig. 9). Specifically, in this study the size of the inner frame was 31 by 31 pixels (approximately 150 by 150 μm)

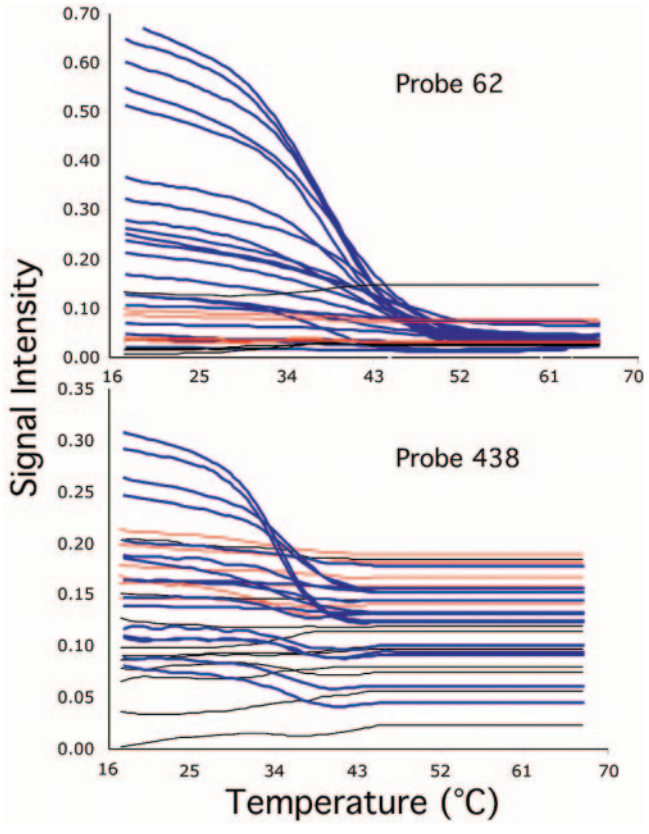

FIG. 5. Interpretation of melting profiles for perfectly matched probe-target duplexes by the MPP calculator. Top panel, a probe (Univ 1390) that tends to yield high initial signal intensity values; lower panel, a probe (S-P-Grpos-1200-a-A-13) that tends to yield low initial SI values. Ideal profiles, blue lines; uncertain profiles, red lines; and poor profiles, black lines. Not all melting profiles are shown for clarity.

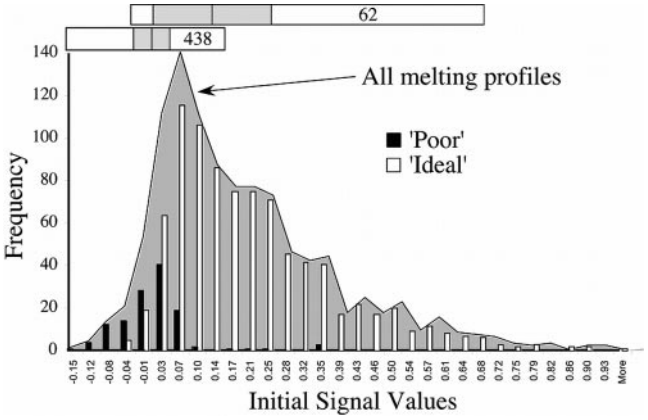

FIG. 6. Distribution of initial SI values of perfectly matched probes from data set 1 by MPP type. Initial SI values of all perfectly matched melting profiles are shown as shaded background (i.e., including uncertain profiles). The range, mean ± standard deviations (gray) of initial signal intensities for probes 438 (S-P-Grpos-1200-a-A-13, ΔG^0 = −21.8 kcal/mol) and 62 (Univ 1390, ΔG^0 = −33.8 kcal/mol) are presented in the horizontal bars above.

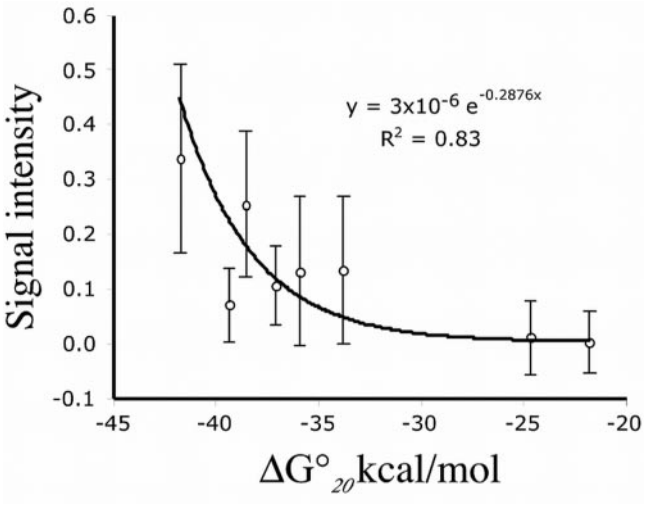

FIG. 7. Relationship between initial SI values of perfectly matched duplexes and their corresponding binding constants (ΔG^0_{20}) for solution. SI values were determined by the Fotin et al. (13) method. Mean and standard deviation ($n > 40$) of SI values are shown for each duplex. Probes in order from lowest to highest ΔG^0_{20} are (number, name): 63, Univ 907; 64, Eub 927; 65, Eub 338; 390, Eub 336; 74, S-P-Grpos-1192-a-A-22; 62, Univ 1390; 75, S-P-Grpos-1199-a-A-15; and 438, S-P-Grpos-1200-a-A-13.

and that for the outer frame was 39 by 39 pixels (in accordance with Fotin et al. [13]), since these sizes are routinely used in the laboratory and were the default settings for data acquisition. The Yershov (52) background subtraction method involves subtracting background using the total SI values immediately surrounding the frame and can be defined by the equation:

$$\mathrm{SI} = I_{\mathrm{inner}} - B_{\mathrm{outer}} \quad (5)$$

where I_{inner} refers to the average SI of the inner frame (141 μm by 141 μm) and the B_{outer} refers to the average SI of the space between the inner and outer frames (200 μm by 200 μm). A third background subtraction method (developed by us) involves subtracting the average or total SI value of the last

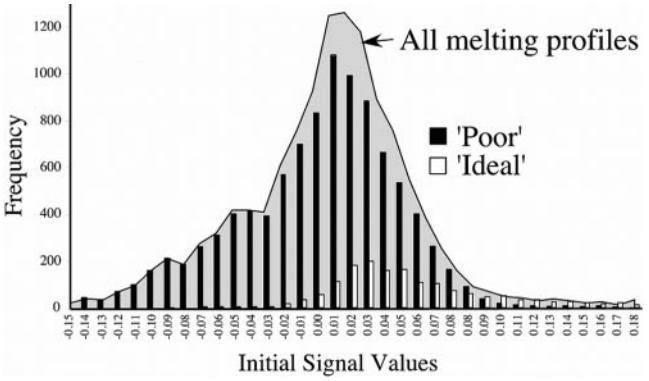

FIG. 8. Distribution of initial SI values of probes having more than two mismatches to target sequences from data set 1 by MPP type. Initial SI values of melting profiles of all probes with more than two mismatches are shown as shaded background (i.e., including uncertain profiles).

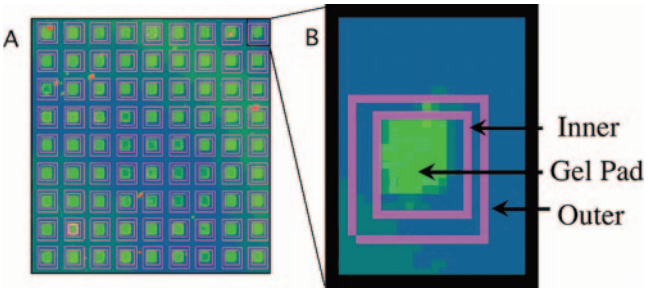

FIG. 9. Color-enhanced image of a portion of a gel-pad microarray showing the inner and outer grids framing the gel pads used by the Fotin et al. (13) image processing software. Gel-pads, green; frames, pink; background, blue. Panel B is a magnified image of single gel-pad from panel A.

image (I_{last}) of the pad from that of each image in a stack (I_{pad}) and can be defined by the equation:

$$\mathrm{SI} = I_{\mathrm{pad}} - I_{\mathrm{last}} \quad (6)$$

Background subtraction methods significantly affected the shape of melting profiles (Fig. 10). The Fotin et al. (13) method resulted in erratic variations in the melting profiles of all image stacks, indicating that the placement of the frame significantly affected the fidelity of melting profile readout (Fig. 10A). The corresponding normalized, i.e., $(X - \mathrm{min})/(\mathrm{max} - \mathrm{min})$, melting profiles are shown in Fig. 10B, 10D, and 10F. Melting profiles analyzed by the Fotin et al. (13) method yielded a much greater range in the variability of T_{d} values (47 to 53°C) (Fig. 10B) than the other methods (51 to 52°C) (Fig. 10D and 10F), indicating that the background subtraction method can drastically affect interpretation of melting profiles. Although the Fotin et al. (13) method accounted for most of the variation in SI values (Fig. 10A), a spike was clearly visible at 25°C for melting profiles obtained using the alternative methods (Fig. 10C and E). The source of this variation is not known, but may be due to changes in shape of the coverslip of the reaction chamber which altered the target fluorescence in the gel-pads. Spikes in SI values occurred at 25°C in numerous experiments (i.e., $n > 30$) (data not shown; Zack McMurry, personal communication). The other two original raw stack images analyzed in the same manner yielded similar results (not shown).

Results from displacement experiments obtained using native 16S rRNA sequences from *Bacteroides forsythus* also yielded similar results, although variation of initial SI values was less than those obtained using oligonucleotide targets. However, the range of T_{d}s was about the same (data not shown). Figures 10C and 10E show that initial signal intensity values remained relatively constant when just I_{inner} (equation 5) or I_{pad} (equation 6) was used. Apparently it is the B_{outer} that leads to nonuniformed signal intensity values presumably because it is most affected by the diffusion of the target out of the gel pad and into solution. When I_{inner} is divided by B_{outer} in Fotin's method (equation 4), the noise of both intensities are multiplied together increasing the noise of signal intensity values. Hence, increased variation in initial SI values for oligonucleotide targets might be due to Fotin's equation.

Data sets 1 and 2 were produced using Fotin et al. (13)

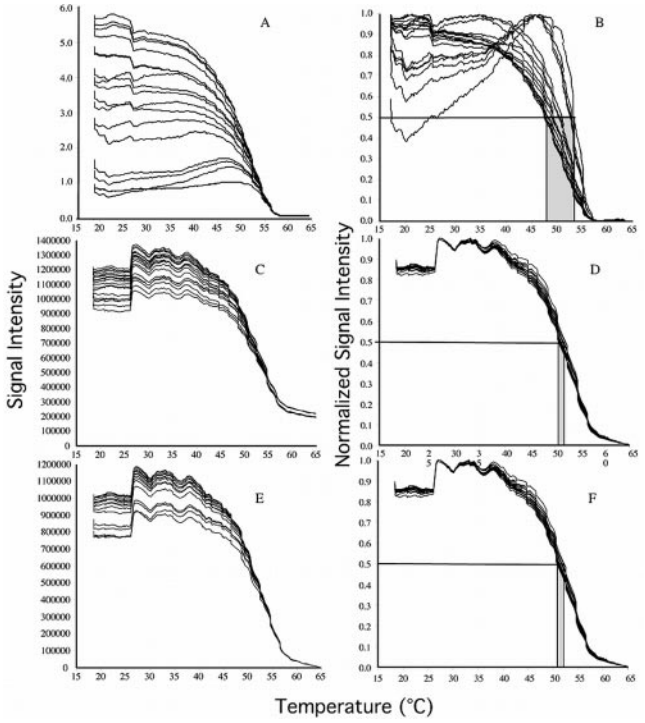

FIG. 10. Composite melting profiles derived from the analysis of 20 displaced image stacks from a single gel-pad microarray experiment. Placement of the original stack image was displaced in the *by* and *y* coordinates of the frame (see text for details). Panels A, C, and E represent stack images processed by using different background subtraction methods: in-out/out (equation 4) (13), the in-out (equation 5) (52), and the in method (equation 6) (this study), respectively. Panels B, D, and F represent the normalized melting profiles (used to calculate the dissociation temperature) from panels A, C, and E, respectively. Gray boxes indicate the range of T_d values.

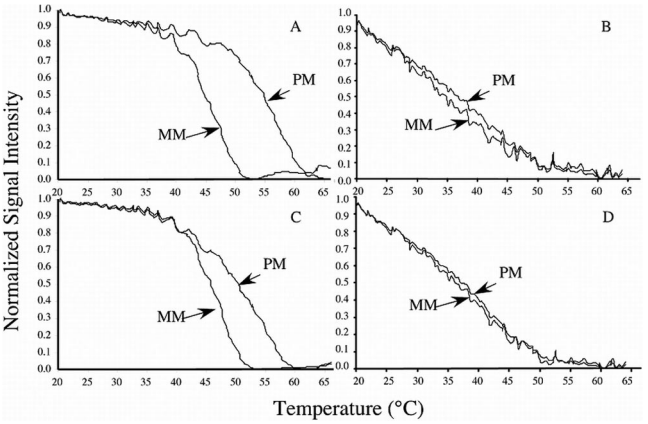

FIG. 11. Differences in melting profiles obtained using synthetic (i.e., oligonucleotide) (panels A and C) and native rRNA target (panels B and D). Images were obtained using equation 6. Panels A and B represent probes 62 (Univ 1390, perfect match, PM) and 399 (Univ 1390-c13, single internal mismatch, MM), respectively. Panels C and D represent probes 63 (Univ 907, PM) and 401 (Univ 907-c9, MM), respectively.

method. Therefore, one can expect the background subtraction method and placement of the window framing the gel pad to have a substantial effect on the quality of these datasets. Unfortunately, we could not regenerate the datasets because the image acquisition software (13) did not store the original images.

Interestingly, most of the RNA profiles observed with our method (equation 6) were linear rather than sigmoid. When melting profiles were calculated in terms of the Fotin et al. (13) (equation 4) method, they turned out to be curved, indicating that the method alters the form of melting profiles to approximate sigmoid melting profiles in solution. To more thoroughly examine the extreme effects of equation 4 on the shape of melting profiles, in silico simulations were performed using two straight lines that approximated the values of I_{inner} and B_{outer} along the temperature course of a duplex melt. The simulations produced curved melting profiles resembling sigmoid melting profiles in solution. Moreover, increasing the slope of B_{outer} (to simulate the diffusion of labeled targets into solution), substantially increased the curvature of the melt, indicating that target size might have significant affects on the form of melting profiles calculated by this method.

Effects of diffusion on melting profiles. The motion of large RNA fragments in the gel pad and the aqueous solution is determined by diffusion, which in turn, is governed by the size of the molecule. The effect of target fragment size on melting profiles was investigated to determine if diffusion was a major determinant in the observed melting profiles since diffusion affects B_{outer} in equations 4 and 5. The alternative image processing and background subtraction method (equation 6) was used for these experiments. Diffusion could be a major determinant if various-sized targets hybridizing to identical probes yielded different melting profiles. In this case, short synthetic targets (i.e., 18- to 22-nucleotide oligonucleotides) should diffuse more rapidly from the gel-pad into solution during the temperature course of the experiment than native fragmented 16S rRNA (approximately 100 to 150 nucleotides). Alternatively, if observed melting profiles were similar, we could conclude that melting profiles were independent of target fragment size. To compensate for possible dye effects (due to differences in the labeling of target oligonucleotides and native RNA), we compared the differences in T_d between perfect-match and mismatched duplexes (Fig. 11) labeled with the same dye.

In two replicated experiments, there was no difference in the T_ds for pairs 62 and 399, or pairs 63 and 401 when native rRNA was used. However, there were significant differences in T_ds for pairs 62 and 399 (5°C for two experiments) and pairs 63 and 401 (10°C for two experiments) when target oligonucleotides were used. Experiments conducted using lissamine-rhodamine B ethylenediamine dye, rather than Cy3, yielded similar results (Zack McMurry, personal communication), indicating that the differences in dye used for labeling the target did not significantly contribute to diffusion. These findings are consistent with the notion that the shorter synthetic targets readily diffuse into solution when they dissociate from probes in the gel pads to solution while longer targets do not diffuse as rapidly, presumably due to their larger size and secondary structure.

It is important to recognize that initial SI values for perfect-match and mismatched duplexes were significantly different (data not shown). Scaling (in order to compare T_ds) of the

melting profiles, as currently practiced, masked the difference between perfect-match and mismatch duplexes when RNA was used as the target. Therefore, for long RNA fragments, melting may not provide additional power for discriminating perfect-match and mismatched duplexes.

DISCUSSION

NNs have traditionally been used to recognize patterns in data that are impossible to substantiate by linear predictive modeling methods (30). In particular, NNs are useful for analyzing fuzzy, noisy, chaotic, and/or unpredictable nonlinear data. For example, NNs have been used to identify the restriction enzyme patterns of *Escherichia coli* O157:H7 (7), stable low molecular weight rRNA banding patterns of microbial communities (36), and to predict SI values and dissociation temperatures of probe-target duplexes on microarrays (47), the pyrolysis mass spectra of *Mycobacterium tuberculosis* complex species (14), the promoter sites of *E. coli* (18), protein binding sequence motifs (39), and the fatty acids of microbial communities (37).

In this study, NNs were trained to recognize patterns in gel-pad microarray data (e.g., the variability and shape of melting profiles) and to determine the variable (i.e., input), or combinations of variables (i.e., inputs), contributing to observed patterns in the data. The NN program, Neuroet (38) was employed because, in contrast to available NN packages (free and commercial), it allowed us (i) to automatically determine the optimal number of hidden neurons using an iterative approach, and (ii) to easily extract equations from trained NNs in order to link together several NNs that were used to recognize various aspects of melting profiles (e.g., Quality, Shape, and MPP scores), and (iii) to identify the input or combinations of inputs which are important for making predictions.

Before training the NN, the optimal architecture must be determined (5, 30). We optimized the number of hidden neurons (i.e., architecture) for training NNs and compared the predictions of test and validation data sets to ensure that the architecture and training conditions lead to accurate and reliable predictions. The automated procedure in Neuroet that calculated the number of hidden neurons substantially sped up (~100 times) the process of determining the optimum architecture of the NN because the user was not required to manually compare the performance of different NNs (e.g., R-square values of observed versus predicted outputs of test and validation data sets). Rather, the best NN model was selected based on AICc scores, with the lowest score containing the optimum number of hidden neurons for data analyses. Information on the predictions for testing and validation data (Table 2) demonstrated that the final NNs were not under- or overtrained and that the predictions were quite accurate-even for melting profiles not used for training. Equations extracted from these NNs were used to build the MPP calculator.

The MPP calculator consisted of equations extracted from four trained NNs (i.e., NNs that predicted Quality 1, Quality 2, Shape 1, and Shape 2 scores). They were linked together in the following fashion: (i) the predicted outputs from equations predicting the Quality scores were fed into the inputs of the equations predicting the Shape scores of melting profiles, and (ii) the predictions of Quality and Shape scores, as well as other melt characteristics, were used as inputs to an equation that predicted MPP scores. Our approach of linking together different characteristics of melting profiles into one calculator was successful since we were able to effectively discriminate between ideal and poor melt profiles as demonstrated in Fig. 4 and 5. Moreover, we were able to account for most (90%) of the variability in the data.

Just knowing that an NN accurately recognizes specific patterns in complex data does not provide explanatory insight into the contributions of input variables to the prediction process. For this reason, we used the Measuring Importance of Inputs procedure in the Neuroet package. This procedure yielded consistent results when repeated numerous times ($n = 5$), indicating that the approach was statistically robust and in agreement with the rigorous testing conducted in a similar study (38). To our knowledge, this is the first study to demonstrate the utility of this procedure using biological data. The finding that SI values at the beginning of the melt were important to predict Quality and Shape scores was anticipated since SI values are highest at the beginning of the melt and more variable along the temperature course as target sequences dissociate, and SI values approach the detection limit of the system (Fig. 5). The finding that SI values at the beginning, middle, and the end of the melting profile were used to predict Quality scores suggests that, in addition to considering the disjointedness of adjacent SI values at the beginning of the melt, the NN also considered SI values at the maximum slope of the melt, and at the end of the melt.

The MPP calculator provided a way to consistently classify melting profiles based on the change of SI values with temperature. Key findings obtained by analyzing the aggregated data with the MPP calculator were that: (i) approximately 18% of the perfect-match duplexes yielded poor melting profiles, (ii) approximately 20% of duplexes with two or more mismatches were classified as ideal and (iii) these results were similar for two independent data sets. Visual evaluation of these profiles confirmed that indeed they were classified correctly, in contrast to our expectations. Gel-pad microarray technology was developed to serve as an analytic method that would be able to identify specific nucleic acids in a sample. It turns out that the method itself imposes very high complications for interpreting its results (see Discussion about overlapping processes). Hence, these findings identified systematic problems intrinsic to interpretation of melting profiles and gel-pad technology. The observation that certain mismatched duplex types produced ideal melts was anticipated (41) since these duplexes presumably have high binding constants and their SI values were within the dynamic range of the camera. It was for this reason that we investigated the effects of image processing, binding constants, and diffusion on melting profiles.

Interpretation of melting profiles. Classical DNA melting experiments using spectrophotometric analysis have shown that a sigmoid melting profile is precisely determined by the temperature course of the binding constant (9, 29). The hybridization process in a gel-pad microarray might be regarded as an equilibrium process (given adequate time for hybridization). The equilibrium process can be described by the Law of Mass Action (15) and expressed by the following equation:

$$K_p = e^{-\frac{\Delta G^0}{RT}} \quad (7)$$

where ΔG^0 is the change in the standard Gibbs free energy, R is the universal gas constant, T is the absolute temperature, and K_p is the binding constant of the nucleic acid strands. According to equation 7, the binding constant exponentially increases as ΔG° becomes more negative-therefore, strong binding of a probe to a target (i.e., high K_p) indicates high duplex stability (i.e., negative ΔG°). Binding constants (and therefore the ΔG^0s) are sequence-dependent (9) since duplexes formed from different perfect-match probes yielded different initial SI values (Fig. 7). These findings are consistent with the notion that differences in initial SI values of duplexes (determined by using equation 4) are mainly (~83%) attributable to differences in the binding constants. In contrast, the melting process is not an equilibrium process because prior to the melting of duplexes, the microarray was washed with a stringent buffer that removes all nonhybridized material. Hence, duplexes in gel-pads are not in equilibrium with the original single-stranded nucleic acids.

The dissolution process is determined by the rate with which duplexes dissociate, the diffusion of the target within the pad, and the diffusion of the target into the solution outside of the pad. Therefore, the observed (nonequilibrium) melting profile in a gel-pad is an overlap of the true melting of duplexes, the diffusion of the target into solution, and the temperature-dependence of the fluorescent dye used in these experiments (33). We demonstrated the affects of target size on melting profiles by comparing the melting profiles of short synthetic targets (oligonucleotides) to those of native rRNA fragmented using the currently established protocol (Materials and Methods). While this and a previous study (48) demonstrated that duplexes formed using synthetic targets in the range of 20 to 40 nucleotides allowed the effective discrimination of perfectly matched and mismatched duplexes (Fig. 11), fragmented native rRNA targets did not. Hence, discrimination of perfectly matched from single mismatched duplexes (12) with rRNA fragmented using the current protocol (producing fragments in the range of 100 to 150 nucleotides) appears to be problematic presumably because of reduced diffusivity (a function of size and possibly secondary structure).

The effects of diffusion are especially relevant to our study since the observed melting profiles were analyzed by image acquisition software that used the Fotin et al. (13) background subtraction method. Fotin et al. (13) originally proposed this method, which is just a division of the local background corrected SI value over its local background, because it partially compensated for the temperature dependence of the fluorescence dye and variations in the intensity of the exciting light. Moreover, this method has been widely used in numerous studies (8, 10, 12, 20, 24, 45, 47, 48, 49). The space immediately surrounding the gel-pad is the region most affected by the diffusion of the dissociated target into solution, which can be observed by the significant increase in SI values of the background (i.e., B_{outer}) during the temperature course of the experiment (data not shown). For this reason, minor differences (i.e., 6 to 7%) in placement of the rigid grid that frames gel-pads relative to the background (i.e., B_{outer}) significantly influence the form of melting profiles and their initial SI values produced by this method (Fig. 10).

The existing image analysis software was not sufficiently flexible to allow ideal placement of the grid to all pads on a microarray (centering the inner window on each gel pad). For instance, placement of the grid to pads in one region (e.g., corners of the microarray) was always associated with different placement of the grid to pads in other regions of the microarray (data not shown). Therefore, diffusion of the target and placement of the grid frames were major sources of experimental variation in this study (as is likely true of others) (8, 12, 24, 26, 47, 48) that affect our ability to interpret microarray data.

Although some erratic melting profiles may be attributed to the Fotin et al. (13) background subtraction method, uncertain or poor melting profiles might also be due to probes with low binding constants. For example, more than 60% of the targets that were complementary to probe 438 (S-P-Grpos-1200-a-A-13) were classified as having poor melting profiles, because probe 438 has a low binding constant ($\Delta G^0 = -21.8$ kcal/mol., see Fig. 7) to complementary targets as clearly shown in Fig. 5. Fluorescence originating from duplexes with low binding constants presumably approaches the detection limit of the system. These results also demonstrate that the concentration of fragmented target sequences and the binding constant of the probes, combined with the dynamic range of the camera, can sometimes confound our ability to detect fluorescent signals of melting profiles.

Our working hypothesis that the general form of a melting profile in gel-pad microarrays (within the dynamic range of the detection system) should be sigmoid-shaped and not change from one duplex to the next, and that observed differences are due to experimental variations is supported because we found ideal melting profiles for various perfect-match and mismatched duplexes. Therefore, melting of duplexes that have various types and compositions have the same general form of melting profiles. Deviations from the general form can be attributed to experimental artifacts (e.g., Fotin's equation, diffusion of the target, SI values approaching the detection threshold of the camera, etc.).

In accordance with equation 7, the binding constant should exponentially decrease (and ΔG^0 become more positive) for duplexes with mismatched base pairs, making the duplexes unstable. Table 4 supports this general trend: duplexes with 2 or more mismatched base pairs yielded a lower number of ideal melting profiles and a higher number of poor melting profiles. Most duplexes containing mismatches were classified as having poor melting profiles because they have low binding constants, and consequently, were below the detection limit of the system.

Some duplexes with mismatches (~18%) consistently yielded ideal melt profiles (Table 4). The MPP calculator classifies melting profiles (i.e., ideal, uncertain, or poor) based on subjective judgment about the form of melting profiles. Smooth sigmoid shape and absence of erratic glitches were the primary criteria for classifying a profile as ideal as shown by the majority of melting profiles from perfect-match duplexes. This finding indicates that the simple observation of melting profiles was not enough to interpret gel-pad microarray data, as used in a previous study (12). Indeed, some duplexes with multiple mismatches behaved the same way as perfect-match duplexes: reasonable T_d and ideal shape. Hence, a melting profile, by

itself, does not provide useful information for distinguishing between specific and nonspecific hybridizations. These results are very problematic for the application of gel-pad microarray technology to environmental samples of unknown composition (both sequence- and concentration-wise). Perhaps development of a robust physical model that considers the effects of diffusion, dissociation kinetics, thermodynamics of hybridization, sequence dependencies, target concentration, and rate of temperature increase would drastically improve the situation.

In summary, the NN approach outlined in this study was especially useful for examining data affected by overlapping factors (e.g., placement of the grid frames, background subtraction method) which were not obvious at the beginning of the study. The analysis of 1,293 perfect-match duplexes by the MPP calculator revealed that ~18% were correctly classified as having poor melting profiles. Visual examination of all of these profiles showed that they indeed looked poor having a linear shape and very low SI values. Presumably, this result was due to image processing artifacts (see above) and/or low binding constants of the probes, despite the fact that they were perfect-match duplexes.

Towards developing a diagnostic tool for microbial identification. To address the problems associated with the image acquisition and processing, we have developed a new image analysis system that captures the images of all gel pads on a microarray as they change with temperature. The images are stored as image stacks. Image processing software has also been developed that allows users: (i) to place grid frames, (ii) to convert image stacks to SI values, and (iii) to implement a variety of background subtraction methods including those in equations 4, 5, and 6. We have used this software to analyze melting profiles as shown in Fig. 10 and 11. Unfortunately, data sets 1 and 2 could not be reanalyzed because the original image analysis software did not save images. Our new image acquisition and processing system does save images.

Although single-base-pair discrimination is possible with gel-pad microarrays using oligonucleotide targets, this level of resolution might not be generally possible with native rRNA unless the target rRNA is fragmented to a very short length, similar to that of oligonucleotides used in this and other studies (47, 48). As shown in Fig. 11, we were not able to clearly distinguish between perfect-match and mismatched duplexes with a small study set of probes examined with the current fragmentation protocol, which produces fragments of 100 to 150 nucleotides in length, and the modified image analysis software. Also, the secondary structure of RNA may play a critical role in discrimination. Although fragmenting rRNA to shorter lengths might help resolve single nucleotide mismatches, one has to be aware that extensive fragmentation by any currently existing hydrolytic method using metal ions or Brønsted acids and bases induces bias towards the fragments involved in secondary structure (helices) (32). This bias will affect quantification of microorganisms from environmental samples since some of the informative single stranded portion of rRNA will be destroyed by hydrolysis.

Application of the MPP calculator and subsequent experiments revealed that the main variables affecting the form of melting profiles using the current gel-pad microarray are: the placement of the grid frames and the background subtraction method which are both used by the image analysis system, and duplexes with low binding constants. A key variable affecting the use of dissociation analysis to improve mismatch discrimination, relative to more conventional use of an average hybridization stringency, was shown to be dependent on target molecule size. Although single mismatch discrimination has been demonstrated for certain probes using oligonucleotide targets with the existing technology, an assessment of the more general utility of the gel-pad format for dissociation analysis will require reevaluation of both the image processing and fragmentation/labeling protocols.

ACKNOWLEDGMENTS

We thank John Kelly, Heidi Gough, Jim Smoot, Seana Davidson, and David Stahl for thoroughly reading the manuscript and providing valuable suggestions. We also thank Travis Krick and Zack McMurry for their technical assistance and Laura Smoot for providing microbial cultures. We thank three anonymous reviewers for their helpful insights. We are grateful to Hauke Smidt and Martin Könnecke for providing their microarray data.

This work was supported by grant 1U01DE014955-01 from NIH/NIDCR to H.S. and P.A.N. and grant R-82945801 from EPA-CEER-GOM to P.A.N. Funding through DARPA provided the gel pad microarrays.

REFERENCES

1. **Aleksander, I., and H. Morton (ed.).** 1991. An introduction to neural computing, p. 1–20. Chapman & Hall, Ltd., London, United Kingdom.
2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
3. **Basheer, I. A., and M. Hajmeer.** 2000. Artificial neural networks: fundamentals, computing, design, and application. J. Microbiol. Methods. **43:**3–31.
4. **Bavykin, S. G., J. P. Akowski, V. M. Zakhariev, V. E. Barsky, A. N. Perov, and A. D. Mirzabekov.** 2001. Portable system for microbial sample preparation and oligonucleotide microarray analysis. Appl. Environ. Microbiol. **67:** 922–928.
5. **Bishop, C., M.** 1995. Neural networks for pattern recognition. Oxford University Press, London, United Kingdom
6. **Bodrossy, L., and A. Sessitsch.** 2004. Oligonucleotide microarrays in microbial diagnostics. Curr. Opin. **7:**245–254.
7. **Carson, C. A., J. M. Keller, K. K. McAdoo, D. Wang, B. Higgins, C. W. Bailey, J. G. Thorne, B. J. Payne, M. Skala, and A. W. Hahn.** 1995. *Escherichia coli* O157:H7 restriction pattern recognition by artificial neural networks. J. Clin. Microbiol. **33:**2894–2898.
8. **Chechetkin, V. R., A. Y. Turygin, D. Y. Proudnikov, D. V. Prokepenko, E. V. Kirillov, and A. D. Mirzabeko.** 2000. Sequencing by hybridization with generic 6-mer oligonucleotide microarray: an advanced scheme for data processing. J. Biomol. Struct. Dynam. **18:**83–101.
9. **Doktycz, M. J., M. D. Morris, S. J. Dormady, K. L. Beattie, and K. B. Jacobson.** 1995. Optical melting of 128 octamer DNA duplexes. J. Biol. Chem. **270:**8439–8445.
10. **Drobyshev, A., N. Mologina, V. Shik, D. Pobedimskaya, G. Yershov, and A. Mirzabekov.** 1997. Sequence analysis by hybridization with oligonucleotide microchip: identification of *B-thalassemia* mutations. Gene **188:**45–52.
11. **Drobyshev, A. L., A. S. Zasedatelev, G. M. Yershov, and A. D. Mirzabekov.** 1999. Massive parallel analysis of DNA-Hoechst 33258 binding specificity with a generic oligodeoxyribonucleotide microchip. Nucleic Acids Res. **27:** 4100–4105.
12. **El Fantroussi, S., H. Urakawa, A. E. Bernhard, P. A. Noble, J. J. Kelly, and D. A. Stahl.** 2003. Direct profiling of environmental microbial populations by thermal dissociation analysis of native ribosomal RNAs hybridized to oligonucleotide microarrays. Appl. Environ. Microbiol. **69:**2377–2382.
13. **Fotin, A. V., A. L. Drobyshev, D. Y. Proudnikov, A. N. Perov, and A. D. Mirzabekov.** 1998. Parallel thermodynamic analysis of duplexes on oligodeoxyribonucleotide microchips. Nucleic Acids Res. **26:**1515–1521.
14. **Freeman, R., R. Goodacre, P. R. Sisson, J. G. Magee, A. C. Ward, and N. F. Lightfoot.** 1994. Rapid identification of species within the *Mycobacterium tuberculosis* complex by artificial neural network analysis of pyrolysis mass spectra. J. Med. Microbiol. **40:**170–173.
15. **Greiner, W., L. Neise, and H. Stocker.** 1995. Thermodynamics and statistical mechanics. Springer-Verlag, New York, N.Y.
16. **Guggenheim, M., S. Shapiro, R. Gmur, and B. Guggenheim.** 2001. Spatial arrangements and associative behavior of species in an in vitro oral biofilm model. Appl. Environ. Microbiol. **67:**1343–1350.
17. **Guschin, D. Y., B. K. Mobarry, D. Proudnikov, D. A. Stahl, B. E. Rittmann, and A. D. Mirzabekov.** 1997. Oligonucleotide microchips as genosensors for

determinative and environmental studies in microbiology. Appl. Environ. Microbiol. **63:**2397–2402.
18. **Horton, P. B., and M. Kanehisa.** 1992. An assessment of neural networks and statistical approaches for prediction of *E. coli* promoter sites. Nucleic Acids Res. **20:**4331–4338.
19. **Hurvich, C. M., and C. L. Tsai.** 1989. Regression and time series model selection in small samples. Biometrika **76:**297–307.
20. **Koizumi, Y., J. J. Kelly, T. Nakagawa, H. Urakawa, S. El Fantroussi, S. AlMuzaini, M. Fukui, Y. Urushigawa, and D. A. Stahl.** 2002. Parallel characterization of anaerobic toluene-and ethylbenzene-degrading microbial consortia by PCR-denaturing gradient gel electrophoresis, RNA-DNA membrane hybridization, and DNA microarray technology. Appl. Environ. Microbiol. **68:**3215–3225.
21. **Kolenbrander, P. E.** 2000. Oral microbial communities: Biofilms, interactions, and genetic systems. Annu. Rev. Microbiol. **54:**413–437.
22. **Kolenbrander, P. E., R. N. Andersen, K. Kazmerzak, R. Wu, and R. J. Palmer, Jr.** 1999. Spatial organization of oral bacteria in biofilms. Methods Enzymol. **310:**322–332.
23. **Krylov, A. S., O. A. Zasedateleva, D. V. Prokopenko, J. Rouviere Yaniv, and A. D. Mirzabekov.** 2001. Massive parallel analysis of the binding specificity of histone-like protein HU to single-and double-stranded DNA with generic oligodeoxyribonucleotide microchips. Nucleic Acids Res. **29:**2654–2660.
24. **Lebed, J. B., V. R. Chechetkin, A. Y. Turygin, V. V. Shick, and A. D. Mirzabekov.** 2001. Comparison of complex DNA mixtures with generic oligonucleotide microchips. J. Biomol. Struct. Dynam. **18:**813–823.
25. **Li, J., E. J. Helmerhorst, C. W. Leone, R. F. Troxler, T. Yaskell, A. D. Haffajee, S. S. Socransky, and F. G. Oppenheim.** 2004. Identification of early microbial colonizers in human dental biofilm. J. Appl. Microbiol. **97:**1311–1318.
26. **Liu, W.-T., A. D. Mirzabekov, and D. A. Stahl.** 2001. Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. Environ. Microbiol. **3:**619–629.
27. **Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, A. B. Yadhukumar, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lubmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K.-H. Schleifer.** 2004. ARB: a software environment for sequence data. Nucleic Acids Res. **32:**1363–1371.
28. **Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker, Jr., P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje.** 2001. The RDP-II (Ribosomal Database Project). Nucleic Acids Res. **29:**173–174.
29. **Marky, L. A., and K. J. Breslauer.** 1987. Calculating thermodynamic data for transition of any molecularity from equilibrium melting curves. Biopolymers **2:**1601–1620.
30. **Masters, T.** 1993. Practical neural network recipes in C++. Academic Press, New York, N.Y.
31. **Matveeva, O. V., S. A. Shabalina, V. A. Nemtsov, A. D. Tsodikov, R. F. Gesteland, and J. F. Atkins.** 2003. Thermodynamic calculations and statistical correlations for oligo-probes design. Nucleic Acids Res. **31:**4211–4217.
32. **Mikkola, S., U. Kaukinen, and H. Lönnberg.** 2001. The effect of secondary structure on cleavage of phosphodiester bonds of RNA. Cell Biochem. Biophys. **34:**95–119.
33. **Molecular Probes Handbook.** 14 January 2005, posting date. [Online.] http://www.probes.com/handbook/.
34. **Moore, W. E., and L. V. Moore.** 1994. The bacteria of periodontal diseases. Periodontol. 2000 **5:**66–77.
35. **Motulsky, H., and A. Christopoulos.** 2002. Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting. GraphPad Software, Inc., Indianapolis, Ind.
36. **Noble, P. A., K. D. Bidle, and M. Fletcher.** 1997. Natural microbial community compositions compared by a back-propagating neural network and cluster analysis of 5S rRNA. Appl. Environ. Microbiol. **63:**1762–1770.
37. **Noble, P. A., J. S. Almeida, and C. R. Lovell.** 2000. Application of neural computing methods for interpreting phospholipid fatty acid profiles of natural microbial communities. Appl. Environ. Microbiol. **66:**694–699.
38. **Noble, P. A. and E. Tribou.** Neuroet: an easy-to-use artificial neural network for ecological and biological modeling. Ecol. Modelling, in press.
39. **O'Neill, M. C.** 1998. A general procedure for locating and analyzing protein-binding sequence motifs in nucleic acids. Proc. Natl. Acad. Sci. USA **95:**10710–10715.
40. **Proudnikov, D., E. Timofeev, and A. Mirzabekov.** 1998. Immobilization of DNA in polyacrylamide gel for the manufacture of DNA and DNA-oligonucleotide microchips. Anal. Biochem. **259:**34–41.
41. **Raskin, L., J. M. Stromley, B. E. Rittmann, and D. A. Stahl.** 1994. Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. Appl. Environ. Microbiol. **60:**1232–1240.
42. **Rumelhart, D. E., G. E. Hinton, and R. J. Williams.** 1986. Learning internal representation by error back propagation, p. 318–362. *In* D. E. Rumelhart and J. L. McCleland (ed.), Parallel distributed processing. M.I.T. Press, Cambridge, Mass.
43. **Socransky, S. S., A. D. Haffajee, M. A. Cugini, C. Smith, and R. L. Kent, Jr.** 1998. Microbial complexes in subgingival plaque. J. Clin. Periodontol. **25:**134–144.
44. **Stahl, D. A., B. Flesher, H. R. Mansfield, and L. Montgomery.** 1988. Use of phylogenetically based hybridization probes for studies of ruminal microbial ecology. Appl. Environ. Microbiol. **54:**1079–1084.
45. **Timofeev, E., and A. Mirzabekov.** 2001. Binding specificity and stability of duplexes formed by modified oligonucleotides with 4096-hexanucleotide microarray. Nucleic Acids Res. **29:**2626–2634.
46. **Tribou, E., and P. A. Noble.** 1 June 2004, posting date. Neuroet: a simple artificial neural network for scientists, University of Washington, Seattle, WA 98195. [Online.] http://noble.ce.washington.edu/Neuroet.htm.
47. **Urakawa, H., P. A. Noble, S. El Fantroussi, J. J. Kelly, and D. A. Stahl.** 2002. Single-base-pair discrimination of terminal mismatches by using oligonucleotide microarrays and neural network analyses. Appl. Environ. Microbiol. **68:**235–244.
48. **Urakawa, H., S. El Fantroussi, H. Smidt, J. C. Smoot, E. H. Tribou, J. J. Kelly, P. A. Noble, and D. A. Stahl.** 2003. Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. Appl. Environ. Microbiol. **69:**2848–2856.
49. **Vasiliskov, V. A., D. V. Prokopenko, and A. D. Mirzabekov.** 2001. Parallel multiplex thermodynamic analysis of coaxial base stacking in DNA duplexes by oligodeoxyribonucleotide microchips. Nucleic Acids Res. **29:**2303–2313.
50. **Wecke, J., T. Kersten, K. Madela, A. Moter, U. B. Göbel, A. Friedmann, and J. P. Bernimoulin.** 2000. A novel technique for monitoring the development of bacterial biofilms in human periodontal pockets. FEMS Microbiol. Lett. **191:**95–101.
51. **Wilson, K. H., W. J. Wilson, J. L. Radosevich, T. Z. DeSantis, V. S. Viswanathan, T. A. Kuczmarski, and G. L. Andersen.** 2002. High-density microarray of small-subunit ribosomal DNA probes. Appl. Environ. Microbiol. **68:**2535–2541.
52. **Yershov, G., V. Barsky, A. Belgovskiy, E. Kirillov, E. Kreindlin, I. Ivanov, S. Parinov, D. Guschin, A. Drobishev, S. Dubiley, and A. Mirzabekov.** 1996. DNA analysis and diagnostics on oligonucleotide microchips. Proc. Natl. Acad. Sci. USA **93**:4913–4918.