1-2002

Single-Base-Pair Discrimination of Terminal Mismatches by Using Oligonucleotide Microarrays and Neural Network Analyses

Hidetoshi Urakawa

Peter A. Noble

Said El Fantroussi

John J. Kelly
Loyola University Chicago, Jkelly7@luc.edu

David A. Stahl

Recommended Citation
Single-Base-Pair Discrimination of Terminal Mismatches by Using Oligonucleotide Microarrays and Neural Network Analyses

Hidetoshi Urakawa,1 Peter A. Noble,1 Said El Fantroussi,1 John J. Kelly,2 and David A. Stahl1∗

Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington 98195,1 and Department of Civil and Environmental Engineering, Northwestern University, Evanston, Illinois 602082

Received 4 June 2001/Accepted 25 October 2001

The effects of single-base-pair near-terminal and terminal mismatches on the dissociation temperature (Td) and signal intensity of short DNA duplexes were determined by using oligonucleotide microarrays and neural network (NN) analyses. Two perfect-match probes and 29 probes having a single-base-pair mismatch at positions 1 to 5 from the 5′ terminus of the probe were designed to target one of two short sequences representing 16S rRNA. Nonequilibrium dissociation rates (i.e., melting profiles) of all probe-target duplexes were determined simultaneously. Analysis of variance revealed that position of the mismatch, type of mismatch, and formamide concentration significantly affected the Td and signal intensity. Increasing the concentration of formamide in the washing buffer decreased the Td and signal intensity, and it decreased the variability of the signal. Although Tds of probe-target duplexes with mismatches in the first or second position were not significantly different from one another, duplexes with mismatches in the third to fifth positions had significantly lower Tds than those with mismatches in the first or second position. The trained NNs predicted the Td with high accuracies (R2 = 0.93). However, the NNs predicted the signal intensity only moderately accurately (R2 = 0.67), presumably due to increased noise in the signal intensity at low formamide concentrations. Sensitivity analysis revealed that the concentration of formamide explained most (75%) of the variability in Tds, followed by position of the mismatch (19%) and type of mismatch (6%). The results suggest that position of the mismatch at or near the 5′ terminus plays a greater role in determining the Td and signal intensity of duplexes than the type of mismatch.

DNA microarray technology provides parallel nucleic acid hybridizations for hundreds to thousands of oligonucleotides or larger DNA fragments on a small surface area (35). In applied and environmental microbiology, this technology has been used for assessing gene expressions (31, 33, 40), characterizing whole genomes (8), and identifying bacteria (11). This technology assumes an adequate discrimination between probes and their targets, which is determined in part by the stability of probe-target duplexes and by hybridization and wash conditions.

The stability of any DNA duplex can be determined by monitoring the nonequilibrium dissociation rate (i.e., melting profile) and then calculating the dissociation temperature (Td) of the probe-target duplex. Acquiring information on how Td changes for a particular DNA duplex under different hybridization and wash conditions allows us to maximize discrimination by minimizing nonspecific hybridizations. However, complete discrimination is often difficult to achieve, particularly for short probe-target duplexes (~18 to 25 nucleotides) with single-base-pair mismatches. Accurately predicting the Td of short probe-target duplexes for defined washing conditions, probe and target lengths, and nucleotide composition and order is a substantial challenge. Understanding the rules governing nucleic acid hybridizations of short probe-target duplexes is necessary and fundamental for the application of microarray technology to routine environmental microbiology because it would facilitate the design of good probes (i.e., those which have high specificity), minimize the chances of nonspecific hybridizations, and improve our ability to confidently interpret microarray data.

There is a paucity of data dealing with the Td of short DNA duplexes, and even less is known about duplexes containing terminal mismatches (34). In general, a mismatch near or at the terminus of a short duplex is less destabilizing than an internal mismatch (38). However, experimental observations have shown that in at least some instances, the type of a mismatch (i.e., the potential loss of hydrogen bonds due to incorrect base pairing) can override the effects of position (39). Hence, there are no absolute rules for predicting the influence of mismatch position on duplex stability (38). The effects of the position of a single-base-pair mismatch and the type of mismatch on predicting Td of a probe-target duplex are nonlinear, with position being more important in some instances and type of mismatch being more important in others. Determining the rules controlling nonlinear factors is further hampered by conventional statistical methods (e.g., analysis of variance [ANOVA]) which use linear approaches to examine nonlinear data. An approach that provides the advantages of conventional statistics while accounting for the nonlinear nature of the experimental data and facilitating a more detailed examination of the data would be very advantageous.

* Corresponding author. Mailing address: Civil and Environmental Engineering, University of Washington, 302 More Hall, Box 352700, Seattle, WA 98195. Phone: (206) 685-3464. Fax: (206) 685-9185. E-mail: dastahl@u.washington.edu.
Artificial neural networks (NNs) provide a useful tool for recognizing patterns in complex, nonlinear data sets such as those associated with predicting the $T_m$ of specific probe-target duplexes. NNs are particularly advantageous over conventional statistical methods because they can deal with the inherent variability associated with biological data. NNs are constructed by using computer software and consist of layers of neurons that make independent computations and pass on their outputs to other neurons (28). Each neuron in a layer is connected to neurons in the next layer, so that the output of each neuron affects the activation of all neurons to which it is connected. Neurons are adaptable and, through the process of learning from examples, store knowledge and make it available for use (1). In a training technique called back-propagation (32), a pattern is presented to an input layer of a network and the network produces output based on the sum of the weighted inputs. When the pattern of the output layer is compared to target values, the errors between them are computed. An error function is used to adjust the weights and biases of each neuron. The adjusted weights of a trained network can be used to recognize and predict patterns such as the $T_m$ of probe-target duplexes. The adjusted weights can also be used to provide information on functional relations between variables and an output. For example, sensitivity analysis of the adjusted weights can be used to determine the relative contribution of individual input neurons (i.e., position of the mismatch, type of mismatch, and formamide concentration) to an output neuron (e.g., $T_m$).

The objective of this study was to investigate the effects of single-base-pair mismatches on $T_m$ and signal intensity of oligonucleotide duplexes using polyacrylamide gel-based DNA microarrays. Specifically, we examined the effects of mismatch type, position of the mismatch relative to the 5’ terminus of the probe, and formamide concentration in the washing buffer on $T_m$ and signal intensity using 31 oligonucleotide probes which target DNA sequences for 16S rRNA (rDNA). A second objective was to extract general rules establishing the relative contribution of each of these variables to duplex stability as needed for more informed probe design. To this end, we report on the use of a back-propagating NN for dealing with the high variability and nonlinearity of this data set. Sensitivity analysis of the trained NNs identified the factors that most contributed to training the NNs.

**MATERIALS AND METHODS**

**Synthesized target DNA.** The 16S rDNA gene sequences of *Staphylococcus epidermidis* (accession numbers L37605 and X75943), which is known as an important opportunistic pathogen associated with infections of synthetic medical devices (14), and *Nitrosomonas europaena* (accession number MB46402), which is a well-studied lithoautotrophic nitrifying bacterium (15, 39), were obtained from GenBank at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/GenBank). A short section of these sequences was custom-designed by using the probe design function of ARB software (http://www.mikro.biologie.tu-muenchen.de). The specificity of the probe for the target was checked with the probe check function in the ARB software, the BLAST search (3) at the National Center for Biotechnology Information, and the Probe Match program in Ribosomal Database Project II (17). Self-complementarities were also examined by Ribosomal Database Project II. An 18-base oligonucleotide probe (5’-GATAGTTATTTGCTGTCGAC-3’) targeting halotolerant and obligately halophilic *Nitrosomonas*, which was designed and previously reported as NEU, was also used (43). Oligonucleotide probes were synthesized with an amino linker at the 3’ end at Argonne National Laboratory (5). Mismatch probes were designed to have various single-base-pair mismatches in the first to fifth positions from the 5’ terminus of the probe (Table 1). The microarray matrix, containing 100- by 100- by 20-μm polyacrylamide gel pads placed 100 μm from each other and fixed to a glass slide, was manufactured by photopolimerization (10) and activated as described previously (27). A total of 3 ml of 1 mM amino-oligonucleotide solutions was applied to each gel element containing aldehyde groups (42) which were designed and implemented by a robot arrayer (44). A total of 31 oligonucleotide probes were immobilized through reductive coupling of the 3’ amino group of the oligonucleotide with the aldehyde group of the activated gel pad on the microarrays (27).

**Hybridization and washing protocols.** Hybridizations were carried out at room temperature (20°C) for 12 h in 40 μl of hybridization buffer containing 1 μg of each target DNA (final concentration, 50 ng/μl), 0.9 M NaCl, 20 mM Tris-HCl (pH 8.0), and 40% formamide. Following 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 0, 10, 20, or 30% formamide. After the final wash, 100 μl of washing buffer was added to the washing chamber (Grace BioLabs, Bend, Ore.) for image and melting profile analyses.

**Image and melting profile analyses.** To generate melting profiles, the microarray was fixed on a thermobar mounted on the stage of a custom-designed epifluorescence microscope (State Optical Institute, St. Petersburg, Russia) and connected with a thermoelectric temperature controller (LFI-3751; Wavelength Electronics, Inc, Bozeman, Mont.) and a water bath (Cole Parmer Instruments Co., Chicago, Ill.). The microscope was equipped with appropriate fluorescence filters (Omega Optical, Brattleboro, Vt.) and a cooled charge-coupled device camera (Princeton Instruments, Trenton, N.J.) and manipulated with a software program which allows image acquisition, processing, and analysis (LabVIEW version 5.1; National Instruments Co., Austin, Tex.) (16). Melting profiles for all probes were monitored and recorded at 2°C intervals between 18 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.

**T_m software.** The $T_m$ software was designed to automatically calculate the experimentally determined $T_m$ and signal intensity at the $T_m$ 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://stahl.ce.washington.edu. This interface contains a Web-based interface for this software is available at http://stahl.ce.washington.edu. This interface contains a Web-based interface for this software is available at http://stahl.ce.washington.edu.

**Normalized data analysis.** The data were normalized to compare various combinations of probe-target duplex profiles and to standardize the $T_m$ calculations. The data were normalized with the following equation: normalized value = (actual value − minimum value)/(maximum value − minimum value).

The temperature of each profile was normalized to a minimum value of 0, corresponding to 10°C, and a maximum value of 1, corresponding to 70°C. The intensities of every sample were normalized to a minimum of 0, corresponding to the lowest intensity in a profile, and a maximum of 1, corresponding to the highest intensity in a profile. If the difference between the maximum and minimum signal intensity was less than 0.1, the normalized value was set to 0.8.

Preliminary experiments revealed difficulties in determining the intensity mid-point needed to calculate the $T_m$ because in some cases (approximately 30% of 369 samples), the maximum signal intensity was slightly greater than the intensity observed at the beginning of the experiment (Fig. 1), even when care was taken to avoid pixel saturation. In other cases, maximum intensity equaled the initial intensity (data not shown). For this reason, three $T_m$ were calculated: one based on maximum intensity, another based on the initial intensity, and the mean of the two intensities.

Normalized intensities in the range of 0.35 to 0.65 were used to calculate slopes around the $T_m$ (Fig. 1). This range was chosen to maximize the number of points used to calculate the slope. The program was designed to consider all possible slopes within the specified range and to select the slope and intercept
TABLE 1. Probes used in this study and effect of formamide concentration, position, and type of mismatch on $T_d$

<table>
<thead>
<tr>
<th>Probe(^a)</th>
<th>Sequence(^b)</th>
<th>$T_d$ with formamide concn (%) in washing buffer(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S. epidermidis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPM</td>
<td>TCGCACATCGCTGCACTGTT</td>
<td>45.4 ± 1.8</td>
</tr>
<tr>
<td>s1aa</td>
<td>aCGCACATCGCTGCACTGTT</td>
<td>43.1 ± 0.8</td>
</tr>
<tr>
<td>s1ga</td>
<td>gCGCACATCGCTGCACTGTT</td>
<td>46.3 ± 1.8</td>
</tr>
<tr>
<td>s1ea</td>
<td>TCGCACATCGCTGCACTGTT</td>
<td>44.2 ± 1.1</td>
</tr>
<tr>
<td>s2ag</td>
<td>TaGCACATCGCTGCACTGTT</td>
<td>47.8 ± 1.4</td>
</tr>
<tr>
<td>s2gg</td>
<td>TgGCACATCGCTGCACTGTT</td>
<td>45.7 ± 1.2</td>
</tr>
<tr>
<td>s2tg</td>
<td>TgGCACATCGCTGCACTGTT</td>
<td>44.4 ± 1.1</td>
</tr>
<tr>
<td>s3cc</td>
<td>TCACATCGCTGCACTGTT</td>
<td>40.8 ± 0.4</td>
</tr>
<tr>
<td>s3tc</td>
<td>TCACATCGCTGCACTGTT</td>
<td>40.7 ± 0.5</td>
</tr>
<tr>
<td>s3cc</td>
<td>TCACATCGCTGCACTGTT</td>
<td>43.5 ± 0.4</td>
</tr>
<tr>
<td>s4gg</td>
<td>TCGgACATCGCTGCACTGTT</td>
<td>43.5 ± 0.4</td>
</tr>
<tr>
<td>s4tg</td>
<td>TCGaACATCGCTGCACTGTT</td>
<td>42.8 ± 0.5</td>
</tr>
<tr>
<td>s4ag</td>
<td>TCGaACATCGCTGCACTGTT</td>
<td>42.0 ± 0.6</td>
</tr>
<tr>
<td>s5gt</td>
<td>TCGgACATCGCTGCACTGTT</td>
<td>44.5 ± 1.2</td>
</tr>
<tr>
<td>s5ct</td>
<td>TCGgACATCGCTGCACTGTT</td>
<td>40.6 ± 0.6</td>
</tr>
<tr>
<td>s5t</td>
<td>TCGgACATCGCTGCACTGTT</td>
<td>43.2 ± 1.2</td>
</tr>
<tr>
<td>N. eutropha</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPM</td>
<td>CCCCTCTGCTGCACTCTA</td>
<td>43.0 ± 1.4</td>
</tr>
<tr>
<td>n1gg</td>
<td>gCCCTCTGCTGCACTCTA</td>
<td>45.4 ± 1.4</td>
</tr>
<tr>
<td>n1ag</td>
<td>aCCCTCTGCTGCACTCTA</td>
<td>45.3 ± 1.8</td>
</tr>
<tr>
<td>n1tg</td>
<td>tCCCTCTGCTGCACTCTA</td>
<td>44.0 ± 1.8</td>
</tr>
<tr>
<td>n2gg</td>
<td>CgCCCTCTGCTGCACTCTA</td>
<td>43.5 ± 1.5</td>
</tr>
<tr>
<td>n2ag</td>
<td>CaCCCTCTGCTGCACTCTA</td>
<td>44.5 ± 0.8</td>
</tr>
<tr>
<td>n2tg</td>
<td>CCCTCTGCTGCACTCTA</td>
<td>44.2 ± 0.9</td>
</tr>
<tr>
<td>n3gg</td>
<td>CgCTCTGCTGCACTCTA</td>
<td>43.6 ± 0.9</td>
</tr>
<tr>
<td>n3ag</td>
<td>CaCTCTGCTGCACTCTA</td>
<td>43.1 ± 1.3</td>
</tr>
<tr>
<td>n3tg</td>
<td>CCCTCTGCTGCACTCTA</td>
<td>41.0 ± 1.3</td>
</tr>
<tr>
<td>n4gg</td>
<td>CgCTCTGCTGCACTCTA</td>
<td>41.7 ± 0.5</td>
</tr>
<tr>
<td>n4ag</td>
<td>CaCTCTGCTGCACTCTA</td>
<td>41.4 ± 0.8</td>
</tr>
<tr>
<td>n4tg</td>
<td>CCCTCTGCTGCACTCTA</td>
<td>40.4 ± 1.0</td>
</tr>
<tr>
<td>n5ga</td>
<td>CCCgCTGCTGCACTCTA</td>
<td>39.5 ± 0.8</td>
</tr>
<tr>
<td>n5aa</td>
<td>CCCgCTGCTGCACTCTA</td>
<td>41.5 ± 0.9</td>
</tr>
</tbody>
</table>

\(^a\) Probe names incorporate the type of target (s, *Staphylococcus*; n, *Nitrosomonas*), position of the mismatch (positions 1 to 5 from 5' end of probe), and the type of mismatch (probe/target). SPM and NPM, perfect match probes for *Staphylococcus* and *Nitrosomonas* targets, respectively.

\(^b\) Mismatches are in lower case.

\(^c\) Values are means ± standard deviations (n = 3).

with the highest Pearson correlation coefficient. The selected slope and intercept were then used to calculate the normalized $T_d$. The normalized $T_d$ was used to determine the equation defining the slope and intercept between normalized and actual temperatures and back-calculating the actual value. Similarly, the intensities at the $T_{ds}$ were calculated by determining the slope and intercept defining the equation between normalized and actual intensities and back-calculating the signal intensity at the $T_d$. For Fig. 1, $T_{ds}$ was calculated using the maximum normalized intensity in the profile, $T_{ds}$ was calculated using the initial intensity (i.e., lowest temperature), and $T_{ds}$ was calculated using the mean of the maximum normalized intensity and the initial intensity.

**Data for the NN.** The complete input data set consisted of the following variables: the position of the single-base-pair mismatch, the type of mismatch, and the concentration of formamide used in the washing buffer. The position of the mismatch was given a value of 0 for a perfect match between probe and target and a value of 1 to 5 for mismatches occurring at one to five positions from the 5' terminus of the probe, respectively. The type of mismatch refers to the number of hydrogen bonds which would be potentially lost from the target DNA due to incorrect base pairing. If the number of hydrogen bonds potentially lost from the target DNA due to incorrect base pairing was three (i.e., a G or C mismatch), the mismatch type was coded as 1. If the number of hydrogen bonds potentially lost from the target DNA was two (i.e., an A or T mismatch), the mismatch type was coded as 0. Perfect matches could not be coded as 0 or 1 for the type of mismatch. Since the NN required a value for type of mismatch, the number of perfect match records in the data set was duplicated, one-half of the records were coded as 0 for mismatch type, and the remainder were coded as 1. Formamide concentrations were coded as 0, 10, 20, and 30, corresponding to the percent formamide used in the washing buffer.

The output data set for determining the $T_d$ consisted of one variable: the experimentally determined $T_d$. The output data set for determining the signal intensity consisted of one variable: the calculated signal intensity at the experimentally determined $T_d$ as determined by the $T_d$ software.

For NN and sensitivity analyses, the entire data set was normalized, with a minimum value of the data set being set at 0.1 and a maximum value set at 0.9 (18). At the end of the analyses, the NN predictions were converted back to their
actual values by defining the equation between the normalized and actual data and back-calculating the value.

**Statistical software.** The NN and sensitivity software was custom-designed by using C++ software and based on the recommendations of Bishop (6), Hagan et al. (12), Masters (18), and Basheer and Hajmeer (4). Stand-alone applications of the NN and sensitivity analysis used in this study are available on request (panoble@washington.edu), and a Web-based user interface is available at http://noble.cc.washington.edu. This Web-based interface contains neuralreadme and sensereadme files, which describe how to use the software, and specifies the files needed to implement the back-propagating NN and sensitivity analysis (e.g., data and datay).

For NN analysis, a logistic equation, \( f(x) = 1/(1 + e^{-x}) \), was used as the transfer function and error back-propagation was used to optimize the connection weights and biases. Full interconnection between the layers was used. The learning rate ranged from 0.1 to 0.5 U depending on the training run and was adjusted accordingly by the user. The input and output architecture of the NN consisted of three input neurons representing the position of the mismatch, the type of mismatch, and the concentration of the formamide and one output neuron representing the normalized \( T_d \) or signal intensity at the \( T_d \).

For NN analysis, the order of the data records was randomized, and 90% of the data (\( n = 347 \)) were selected for training the NN. The remaining 10% (\( n = 39 \)) were used for testing the trained NN. In order to minimize the effective number of degrees of freedom in the network, training was stopped when the error measured with the independent test data started to increase (6, 24). This criterion was also used to select the optimal number of hidden neurons, which was determined to be three for all NNs.

Sensitivity analysis was used to determine which of the NN inputs (i.e., position of the mismatch, type of mismatch, and concentration of formamide) significantly contributed to predicting the \( T_d \) or signal intensity at \( T_d \). A C++ program was custom-designed and is similar to the sensitivity programs described by Masters (18) and Noble et al. (24). Briefly, the relative sensitivity of the normalized \( T_d \) was defined as the change in normalized \( T_d \) relative to the change of an input. The sensitivity of each input was determined by increasing the minimum value of an input to its maximum value by using a step function for every possible combination and recording the change in normalized \( T_d \). In this study, four step intervals were used for each input value. Changes in \( T_d \) for all variables in a sample were adjusted to a total value of 1, with each variable contributing to a portion of the total change in \( T_d \). Sensitivity analysis for this study involved a total of 4,632 combinations (3 inputs \( \times \) 4 intervals \( \times \) 386 samples). The overall relative sensitivity of the inputs was determined by calculating the mean change in normalized \( T_d \) for each input (386 samples).

**Conventional statistical analysis.** Pearson product-moment correlation was used to determine the degree of association between variables. Linear regressions were used to estimate the relationship of one variable to another (36). An ANOVA was used to determine the source of variability in the experimental data. The Student-Newman-Keuls (SNK) test was employed to determine whether the difference between any two means in a set of means was significant (20). ANOVA and SNK tests were conducted by using the GLM procedure in an SAS program (release 6.12; SAS Inc., Cary, N.C.). Student t tests were tabulated in MS Excel 98 (Microsoft, Inc., Redmond, Wash.) using a Macintosh 9.1 operating system.

**RESULTS**

**\( T_d \) software.** Two-tailed \( t \) tests of \( T_{d1} \) and \( T_{d2} \) using the entire data set (\( n = 386 \)) revealed that there was no statistical difference in the mean \( T_d \), suggesting that differences in the maximum and initial intensities had no affect on the calculated \( T_d \). However, it was necessary to consider the difference in \( T_d \) for samples not treated with formamide, since two-tailed \( t \) tests of \( T_{d1} \) and \( T_{d2} \) (\( n = 96 \)) yielded significant differences (\( P < 0.04 \)). Preliminary results suggested that this difference may have resulted from pixel saturation in samples containing low concentrations of formamide and decreased background signal.

---

**FIG. 1.** Effect of maximum and initial signal intensities on the calculated \( T_d \). Three \( T_{d8} \) were calculated based on the midpoints of initial, (maximum + initial)/2, and maximum intensities corresponding to the transition from DNA duplex to random coil. Shown is an intensity-temperature profile for the \( S. \) epidermidis target probe \( s2ag \) and the target sequence. The normalized \( T_{d8} \) were calculated to be 0.63, 0.64, and 0.64, which correspond to \( T_{d8} \) of 47.9, 48.3, and 48.1°C, respectively.

![Diagram](https://via.placeholder.com/150)
originating from nonspecific binding. The results (presented below) support this, since high concentrations of formamide in the washing buffer decreased background signal intensity. For this reason, T_{d3} was used for all subsequent analyses.

**Effect of formamide, position, and type of mismatch on T_{d}**

Preliminary studies using 0 to 70% formamide in the hybridization buffer revealed that 40% formamide was optimal because it yielded the highest signal intensities (data not shown). To further investigate how formamide affects signal intensity, we included formamide in the washing buffer. At room temperature (20°C), none of the probes was washed off the microarray at 0% formamide (Fig. 2). However, at 30% formamide, some of the probes with mismatches at positions 3 to 5 had reduced signal intensities. Figure 3 shows the effects of 0 and 30% formamide on melting profiles of a probe-target duplex with a single-base-pair mismatch at the fourth position from the 5’ terminus of the probe. Formamide shifted the melting profiles to the left, reducing the T_d.

Table 1 lists the mean T_d as a function of the concentration of formamide in the washing buffer, the type of mismatch, and the position of the mismatch from the 5’ terminus of the probe. ANOVA revealed that the formamide concentration in the washing buffer, the type of mismatch, and the position of the mismatch significantly affected the T_d (Table 2). In general, increasing the formamide concentration by 1% decreased the T_d by approximately 0.6°C (Fig. 4a). The regression model indicated that there was a significant interaction between the concentration of formamide and the type of mismatch and T_d (Table 2). At high formamide concentrations (e.g., 30%), the potential loss of three hydrogen bonds due to incorrect base pairing in a probe-target duplex (i.e., G or C mismatches) [T_d = 42.1°C − (0.58 × % formamide); 250 samples; R^2 = 0.90] had a greater effect on the T_d than the potential loss of two hydrogen bonds (i.e., A or T mismatches) [T_d = 42.1°C − (0.52 × % formamide); n = 119 samples; R^2 = 0.83]. These results provide evidence for the nonlinear relationships between T_d, formamide concentration, and mismatch type. SNK results (α = 0.05) indicated that the loss of three hydrogen bonds significantly decreased the T_d by 0.9°C (mean ± stand-

**TABLE 2. ANOVA of T_d as a function of formamide concentration, position of the mismatch, and type of mismatch (369 samples)**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Mean square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formamide concentration (%)</td>
<td>3</td>
<td>4,971.1</td>
<td>1,846.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Position of the mismatch from 5’ end</td>
<td>5</td>
<td>151.5</td>
<td>56.3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Type of mismatch (A/T or G/C)</td>
<td>1</td>
<td>58.3</td>
<td>21.7</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Formamide concentration (%) × type of mismatch (A/T or G/C)</td>
<td>3</td>
<td>15.8</td>
<td>5.9</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

FIG. 2. (a) Typical image of a DNA microarray; (b) fluorescence images of perfect match (PM) duplexes and duplexes with a single-base-pair mismatch. Single-base-pair mismatches are located at the first to fifth positions of the 5’ terminus of the probe. After hybridization, the microarrays were washed with a washing buffer containing 0 or 30% formamide (see Materials and Methods). 1, *S. epidermidis*, 0% formamide; 2, *S. epidermidis*, 30% formamide; 3, *N. eutropha*, 0% formamide; 4, *N. eutropha*, 30% formamide.

FIG. 3. Effect of formamide concentration on the melting profiles of n4gg and *N. eutropha* target. The single-base-pair mismatch occurs at the fourth position from the 5’ terminus of the probe. Open circles, perfect-match *Nitrosomonas* probe in 0% formamide; closed circles, n4gg in 0% formamide; open squares, perfect-match *Nitrosomonas* probe in 30% formamide; closed squares, n4gg in 30% formamide.
In general, moving the position of the mismatch away from the 5’ terminus to the center of the probe decreased the $T_d$ of the probe-target duplex (Fig. 5a). The $T_d$ of probes with a perfect match to the target and those with a single-base-pair mismatch at the first or second position from the 5’ terminus of the probe were not significantly different from one another (Fig. 5a) (SNK test). Similarly, the $T_d$ of probes having mismatches located at the third to fifth positions (Fig. 5a) (SNK test) were not significantly different from one another. Overall, probes containing a single-base-pair mismatch at positions 3 to 5 ($32.6 \pm 6.6^\circ C, n = 212$) decreased the $T_d$ by 2.8°C compared to probes with a single-base-pair mismatch at the first or second position ($35.4 \pm 6.6^\circ C, n = 168$).

**Effect of formamide concentration on $T_d$ and signal intensity.** ANOVA revealed that formamide concentration, type of mismatch, and position of the mismatch from the 5’ terminus of the probe significantly affected signal intensity of the probes (Table 3). Signal intensities at $T_d$ were not

![Graph of $T_d$ vs. Formamide concentration](image)

**TABLE 3. ANOVA of signal intensity at $T_d$ as a function of formamide concentration, position of the mismatch, and type of mismatch (369 samples)**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Mean square</th>
<th>$F$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formamide concentration (%)</td>
<td>3</td>
<td>11.4</td>
<td>2.6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Position of the mismatch from 5’ end</td>
<td>5</td>
<td>0.5</td>
<td>10.6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Type of mismatch (A/T or G/C)</td>
<td>1</td>
<td>0.2</td>
<td>3.6</td>
<td>0.0011</td>
</tr>
</tbody>
</table>
significantly different at 0 and 10% formamide. However, at 20 and 30% formamide, the signal intensity decreased linearly as well as the variability associated with the signal (Fig. 4b). The statistically insignificant increase in signal intensity at 10% formamide was presumably due to decreased nonspecific binding on the gel pads. Conditions favoring nonspecific binding (e.g., low formamide concentration) might have increased the background noise, decreasing the contrast and increasing the background correction of the image-processing software.

In general, moving the position of a single-base-pair mismatch away from the 5’ terminus of the probe (i.e., to the center of the probe) decreased signal intensity (Fig. 5b). Probes with perfect matches to the target had significantly higher signal intensities than those with a single-base-pair mismatch at the first to fifth positions from the 5’ terminus of the probe. Single-base-pair mismatches at the second position had significantly higher signal intensities than mismatches at the first, third, fourth, or fifth position. Presumably, the low signal intensity at the first position was an artifact of the data set, because mismatches in the first position included AT and GC mismatches while the second to fourth positions included only GC mismatches. That is, two-hydrogen-bond mismatches were not adequately represented at the second to fourth positions in the targets. Clearly, a variety of targets is needed to elaborate and verify these findings.

NN and sensitivity analysis. Comparison of the predicted versus the actual $T_d$ of three trained NNs yielded similar results with high regression coefficients ($R^2 = 0.93$). Figure 6a shows the relationship between predicted and actual $T_d$s for one of the three NNs. The trained NNs were able to predict the $T_d$ given the position of the mismatch, the type of mismatch, and the formamide concentration in the washing buffer within approximately 2°C from the actual $T_d$ (mean errors ± standard deviations: NN 1, 1.4 ± 1.2°C; NN 2, 1.4 ± 1.1°C; NN 3, 1.4 ± 1.2°C; $n = 386$).

The analysis of the sensitivity of individual inputs to the $T_d$ was repeated for three separately trained NNs to determine the contribution of position of the mismatch, type of the mismatch, and formamide concentration to the $T_d$. Figure 7 shows the results of the sensitivity analysis of one sample from one trained NN. In these experiments, the concentration of formamide had a much greater effect on $T_d$ than the position of the mismatch or the type of mismatch. Table 4 provides a summary of the sensitivity analysis for three separately trained NNs. Sensitivity analysis was performed on several trained NNs because NNs with low correlations between actual and predicted outputs may yield inconsistent results. However, all three sensitivity analyses yielded similar and consistent results. More than 75% of the variation in the data set was explained by the formamide concentration, consistent with the expectation that formamide played a key role in predicting the $T_d$. Approximately 19% of the variation of the data was explained by the position of the mismatches, and less than 6% of the variation was explained by the type of mismatch.

Comparison of the predicted and actual signal intensities of three trained NNs yielded similar results with moderate regression coefficients ($R^2 = 0.67$). Figure 6b shows the relationship between predicted and actual $T_d$s for one of the three NNs. The trained NNs were able to predict the signal intensity at low signal intensities (as depicted by the 99% confidence limits of the predictions). However, at high signal intensities the predictability decreased significantly, presumably due to increased noise in the signal intensity. All data points having actual signal intensities greater than 0.9 U in Fig. 6b represented samples treated with 0 to 10% formamide. More predictable signal intensities were obtained in samples containing 20 to 30% formamide. No other input factors were correlated with the variability of the predictions.

It is also important to note that the relationship between predicted and actual signal intensities does not appear to be linear (Fig. 6b). This finding suggests that the NNs had difficulties in finding patterns in the data presumably due to intrinsic noise in the signal intensity, particularly in samples...
having low concentrations of formamide in the washing buffer. Noise in the signal intensity clearly limited the ability of the NN to make reproducible predictions.

The analysis of the sensitivity of individual inputs to the signal intensity at the \( T_d \) was repeated for three separately trained NNs to determine the contribution of the position of the mismatch, the type of the mismatch and concentration of formamide to the signal intensity. Table 5 is a summary of the sensitivity analyses conducted with three separately trained NNs. All three sensitivity analyses yielded similar and consistent results. It is important to note that we did not anticipate identical results from the separately trained NN, since random weights used during the first training step are different for each NN and the weights are individually adjusted during training. Rather, we anticipated that different NNs would yield similar and consistent results if they were able to adequately recognize patterns in the data. More than 69% of the variation in the data set was explained by the formamide concentration, indicating that the formamide in the washing buffer played a key role in determining the signal intensity. Approximately 20% of the variation in the data was explained by the position of the mismatch, and less than 7% of the variation was explained by the type of mismatch.

### DISCUSSION

The use of DNA probes in any format for bacterial or gene identification relies on good discrimination between probe-target duplexes with perfect matches and duplexes with mismatched base pairing. Since complete discrimination between perfectly matched duplexes and duplexes with mismatches is often difficult to achieve, the rules governing oligonucleotide duplex stability must be determined in order to establish the relative contribution of nonspecific hybridizations to signal intensity and to facilitate the development of probes having good discriminating capabilities. We chose \( T_d \) as the discriminating measure in this study because \( T_d \) has been extensively used in studies employing conventional membrane hybridization techniques (21, 29, 30, 37), and because calculating the \( T_d \) using the nonequilibrium dissociation approach (i.e., melting profiles) and oligonucleotide microarrays has been successfully demonstrated (16). Moreover, determining the \( T_d \) by using microarrays provides rapid and reproducible data, which facilitates rigorous statistical analyses.

**NN and sensitivity analyses.** The application of NNs to the analysis of complex data in microbiology is relatively new (2), and to our knowledge, no study has used back-propagation of NNs to analyze microarray data. NNs have been used to identify the restriction enzyme profiles of *E. coli* O157:H7 (7), the pyrolysis mass spectra of *Mycobacterium tuberculosis* complex species (9), the promoter sites of *E. coli* (13), protein-DNA binding sites (25), fatty acid profiles of microbial communities (24), nifH-specific binding patterns from denaturing gradient gel electrophoresis analysis (26), and stable low-molecular-weight rRNA profiles (23). However, there is a paucity of studies which have clearly demonstrated the utility of NNs over conventional statistical approaches, and even fewer studies have demonstrated the utility of sensitivity analysis.

NNs are able to recognize nonlinear patterns in complex data which cannot be discerned by using conventional statistical approaches (18). Two independent studies have recently shown that NNs outperformed and provided better predictive power than conventional statistical approaches (22, 24). Based on these studies, we reasoned that better recognition of patterns should improve our ability to predict outputs (such as \( T_d \)), provided that the trained NN has been appropriately optimized (as defined in reference 6). In this study, trained NNs accurately predicted the \( T_d \) and signal intensity at \( T_d \) given the

---

**TABLE 4. Contribution of sources to changes in normalized \( T_d \) based on sensitivity analyses of three independently trained NNs**

<table>
<thead>
<tr>
<th>Source</th>
<th>Relative sensitivity(^a) by NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position of the mismatch for 5' end</td>
<td>0.20 ± 0.02 0.19 ± 0.02 0.19 ± 0.02</td>
</tr>
<tr>
<td>Type of mismatch</td>
<td>0.06 ± 0.03 0.05 ± 0.01 0.06 ± 0.03</td>
</tr>
<tr>
<td>Formamide concentration</td>
<td>0.74 ± 0.03 0.76 ± 0.04 0.75 ± 0.03</td>
</tr>
</tbody>
</table>

\(^a\) Each sensitivity value is based on the contribution of the variable to changing the normalized \( T_d \). Values are means ± standard deviations (\( n = 386 \)).

---

**TABLE 5. Contribution of sources to changes in signal intensity at \( T_d \) based on sensitivity analyses of three independently trained NNs**

<table>
<thead>
<tr>
<th>Source</th>
<th>Relative sensitivity(^a) by NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position of the mismatch for 5' end</td>
<td>0.23 ± 0.04 0.23 ± 0.06 0.19 ± 0.06</td>
</tr>
<tr>
<td>Type of mismatch</td>
<td>0.08 ± 0.04 0.07 ± 0.05 0.07 ± 0.04</td>
</tr>
<tr>
<td>Formamide concentration</td>
<td>0.69 ± 0.05 0.71 ± 0.09 0.74 ± 0.07</td>
</tr>
</tbody>
</table>

\(^a\) Each sensitivity value is based on the contribution of the variable to changing the signal intensity at \( T_d \). Values are means ± standard deviations (\( n = 386 \)).
position and type of mismatch and concentration of formamide. However, despite the considerable correspondence in the results of linear (e.g., ANOVA) and nonlinear (e.g., NN) approaches, we were unable to clearly resolve any differences. This finding is presumably due to the limited range of input factors (i.e., position of mismatch, mismatch type, and formamide concentration) used to predict the outputs (i.e., \( T_d \) or signal intensity at \( T_d \)) and/or because all NNs were optimized to produce generalized predictions and their performance is ultimately limited by the intrinsic noise of the data (6). Nonetheless, on all occasions the NNs yielded results which were consistent with those obtained by ANOVA.

Studies now in progress are considering more complicated interactions, such as the contribution of stacking energies, secondary structure of DNA duplexes, GC content, and length of the probes to \( T_d \) prediction. These studies will likely yield more definitive differences between linear and nonlinear approaches and demonstrate that NNs provide more detailed examination of the data than conventional linear statistical approaches.

Sensitivity analysis of trained NNs yielded information on which individual inputs (in this study, the position of the mismatch, the type of mismatch, and the formamide concentration) significantly contributed to the target output (i.e., \( T_d \) and signal intensity at \( T_d \)). To our knowledge, this is the second study that has demonstrated the utility of sensitivity analysis. A previous study by Noble et al. (24) compared the results of sensitivity analysis to those obtained by principal-component analysis (a linear analysis method). They showed that sensitivity analysis was reproducible and attributed differences in the results to the nonlinearity of the data. In this study, the results of sensitivity analyses were consistent and reproducible, demonstrating that sensitivity analysis was statistically valid and robust.

**Effects of signal intensity on \( T_d \).** An unexpected finding of this study was the relative importance of signal intensity for determining the \( T_d \). As shown in Fig. 6b, the NNs had considerable difficulties predicting signal intensities at low formamide concentrations, presumably owing to increased noise in the signal intensity, as clearly depicted in Fig. 4b. High variability in signal intensities occurred at 0% and 10% formamide. Yet the NNs were able to successfully predict the \( T_d \) regardless of noise in the signal intensities (Fig. 6a). Although \( T_d \) predictions were not affected in this study, it is possible that intrinsic noise in signal intensity could affect \( T_d \) predictions when low concentrations of formamide are used in the washing buffer. Future studies will need to consider signal intensity as an important factor contributing to the variability in \( T_d \) predictions. It is also important to emphasize that variability in the signal intensity can be minimized by using high concentrations of formamide (Fig. 4b). Hence, adding formamide to the washing buffer may improve \( T_d \) prediction.

**Rules governing probe-target duplexes having single-base-pair mismatches.** Currently, this technology has limited use in routine environmental microbiology because we know little about the extent of nonspecific hybridizations of probe-target duplexes and the diversity of 16S rRNA and other gene targets in nature. Clarifying the rules governing probe-target duplexes which have perfect matches and those with a single-base-pair mismatch near or at the 5’ terminus is a necessary and fundamental task for the development of DNA microarray technology.

One rule discovered in this study is that position of the mismatch is more important than the type of mismatch for duplexes with a near-terminus or terminus single-base-pair mismatch. This finding is important for probe design for two reasons: (i) probes having a single-base-pair mismatch in the first or second position from the 5’ terminus may provide an effective means of identifying closely related species or identifying individual species which have high variability in a specific region of the 16S rRNA, and (ii) probes with a single-base-pair mismatch at positions 3 to 5 have greater discriminating abilities than those with a mismatch at or near the 5’ terminus. We anticipate that these are general rules. However, further studies using additional targets and probes are needed to refine and expand these findings.

Another rule discovered in this study is that formamide in the washing buffer decreased noise in the signal intensity. Minimizing noise in the signal intensity is very important for the development of microarray technology because too much noise increases the uncertainty of hybridization events and clouds our ability to accurately discriminate between probe-target duplexes with perfect matches and duplexes with mismatches. This is particularly relevant for analysis of melting profiles because of the problems associated with pixel saturation. Pixel saturation may affect \( T_d \) determination in samples containing low concentrations of formamide because of significant differences between initial and maximum signal intensities (Fig. 1). This problem may be minimized by using appropriate exposure times. However, if this format is to be used for environmental samples, one can expect that the dynamic range will be much larger than our experimental results. Consequently, pixel saturation may be a substantial problem that needs to be addressed in future studies.

**Effect of formamide in the washing buffer.** It is well recognized that differentiation between probe-target duplexes with perfect matches and those with mismatches is affected by the washing conditions (e.g., buffers, salt concentration, and temperature) (29, 41). We examined the effects of formamide in the washing buffer because formamide, being a mild denaturant for nucleic acids, disrupts hydrogen bonding and increases the specificity of hybridizations (38). Thus, we originally hypothesized that formamide should have a greater effect on the type of single-base-pair mismatch than the position of the mismatch in internal mismatches. However, our hypothesis was not supported by the data on near-terminus or terminal mismatches, since we found that the position of the mismatch was more important than the compositional effects of the mismatch, at least for mismatches occurring at and near the terminus of the probe. Nonetheless, formamide decreased the \( T_d \) of G or C mismatches at a slightly higher rate than A or T type mismatches.

Formamide decreased the \( T_d \) at a rate of 0.6°C per 1% formamide (Fig. 4a) in our microarray. This rate is similar but not identical to that reported by McConaughy et al. (19), who showed that \( T_d \) decreases at a rate of 0.7°C per 1% formamide in solution. Salt concentration and subtle differences in duplex stability such as length, GC content, secondary structure, and composition and number of mismatches of probe-target duplexes presumably account for the observed differences.
In summary, oligonucleotide microarrays coupled with NN and sensitivity analysis were used to assess the effects of single-base-pair mismatches located at and near the 5′ termini of probes on discrimination of probe-target duplexes. NNs were able to accurately predict the $T_d$ given the position and type of mismatch and concentration of formamide. The relevant contributions of formamide concentration, position of the mismatch, and type of mismatch for the NN predictions were identified by sensitivity analysis.

ACKNOWLEDGMENTS

We thank G. Yershov, A. Kukhtin, and A. Gemmell for their efforts in manufacturing the oligonucleotide microarrays and S. Surzhikov for synthesis of the oligonucleotide probes. We thank E. Tribou for his computer expertise. We also thank the College Inn Environmental Microbiology Journal Club, who critically reviewed the manuscript. This work was supported by grants from DARPA and NASA to D.A.S. and NSF DEB-0088879 to P.A.N.

REFERENCES