Published in final edited form as:

J Microbiol Methods. 2012 July ; 90(1): 29–35. doi:10.1016/j.mimet.2012.04.003.

Identification of non-specific hybridization using an empirical equation fitted to non-equilibrium dissociation curves

Samuel W Baushke^{a,b}, Robert D Stedtfeld^{a,b}, Dieter M Tourlousse^{a,b}, Farhan Ahmad^{a,b}, Lukas M Wick^c, Erdogan Gulari^d, James M Tiedje^b, and Syed A Hashsham^{a,b,*}

^aDepartment of Civil and Environmental Engineering, 3546 Engineering Building, Michigan State University, East Lansing, MI 48824-1226 ^bCenter for Microbial Ecology, 540 Plant and Soil Sciences Building, Michigan State University, East Lansing, Michigan 48824-1325 ^cBiosynth AG, Rietlistrasse 4, 9422 Staad, Switzerland ^dDepartment of Chemical Engineering, 2300 Hayward St., 3074 H.H. Dow Building, University of Michigan, Ann Arbor, MI 48109-2136

Abstract

Non-equilibrium dissociation curves (NEDCs) have the potential to identify non-specific hybridizations on high throughput, diagnostic microarrays. We report a simple method for identification of non-specific signals by using a new parameter that does not rely on comparison of perfect match and mismatch dissociations. The parameter is the ratio of specific dissociation temperature (T_{d-w}) to theoretical melting temperature (T_m) and can be obtained by automated fitting of a four-parameter, sigmoid, empirical equation to the thousands of curves generated in a typical experiment. The curves fit perfect match NEDCs from an initial experiment with an R^2 of 0.998 ± 0.006 and root mean square of 108 ± 91 fluorescent units. Receiver operating characteristic curve analysis showed low temperature hybridization signals (20–48 °C) to be as effective as area under the curve as primary data filters. Evaluation of three datasets that target 16S rRNA and functional genes with varying degrees of target sequence similarity showed that filtering out hybridizations with $T_{d-w}/T_m < 0.78$ greatly reduced false positive results. In conclusion, T_{d-w}/T_m successfully screened many non-specific hybridizations that could not be identified using single temperature signal intensities alone, while the empirical modeling allowed a simplified approach to the high throughput analysis of thousands of NEDCs.

Keywords

Non-equilibrium dissociation curve; specific dissociation temperature; functional gene; hybridization; microarrays

1. Introduction

High-throughput nucleic acid hybridization systems or microarrays have been widely applied in gene expression studies (Lockhart et al. 1996; Schena et al. 1995; Wodicka et al.

© 2012 Elsevier B.V. All rights reserved.

*Corresponding author: Syed A. Hashsham, Department of Civil and Environmental Engineering, A126 Research Complex-Engineering, East Lansing, MI 48824, Phone: (517) 355 8241, Fax: (517) 355 0250, hashsham@egr.msu.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1997), comparative genomics (Behr et al. 1999; Kallioniemi et al. 1992; Oostlander et al. 2004), microbial gene detection (Bodrossy and Sessitsch 2004; Hashsham et al. 2004; Loy et al. 2002), single nucleotide polymorphism analysis (Gerry et al. 1999; Hacia et al. 1999), and sequencing (Bains and Smith 1988; Chechetkin et al. 2000; Drmanac et al. 1989). During hybridization events, binding of target nucleic acids that are not identical in complementary sequence to the immobilized probes is referred as non-specific hybridization. It affects the quality of microarray data, increases the complexity in data analysis, and results in a false positive detection of a given gene or microorganism (Pozhitkov et al. 2007b). Non-specific hybridization is more challenging for samples with uncharacterized background e.g., environmental matrices (Chandler and Jarrell 2004, 2005; Zhou and Thompson 2002). Therefore, discrimination between specific and non-specific hybridization is important especially for microbial identification with high level of confidence in environmental samples (Wick et al. 2006).

Theoretically, high probe specificity is achieved either by designing oligonucleotide probes which are specific to the complementary sequence of targets and significantly different to the non-targets (Liebich et al. 2006) or by *in silico* predictions based on thermodynamic parameters (Feldkamp et al. 2004). Probe design is still a challenging task as the thermodynamic properties of oligonucleotide hybridization are not yet fully understood (Pozhitkov et al. 2006). Therefore, experimental validation is almost always needed to validate the quality of the designed probes. Experimentally, specificity is achieved by optimization of hybridization conditions (Loy et al. 2002; Peplies et al. 2003), hybridization over a range of temperatures (Li et al. 2004; Liu et al. 2001; Mobarry et al. 1996), or by post-hybridization using non-equilibrium dissociation curves (NEDCs) (Khomiakova et al. 2003; Li et al. 2004) or isothermal post-hybridization washing for one or multiple cycles (Binder et al. 2010; Pozhitkov et al. 2010).

NEDCs have the potential to address non-specific hybridizations in high throughput, diagnostic microarrays (Liu et al. 2001; Pozhitkov et al. 2005; Starke et al. 2006). Dissociation curves are developed by subjecting post-hybridization microarrays to increasing temperatures while measuring the decrease in hybridized nucleic acids. Approaches using NEDCs have primarily focused on comparing perfect match (PM) probe/target dissociation profiles to those that have one or more mismatch (MM) bases. Common approaches calculate the temperatures at which 50% of the initial duplexes remain (T_{d-50}) (Li et al. 2004), temperatures at maximum dissociation (Pozhitkov et al. 2005), specific dissociation temperature (T_{d-w}), (Wick et al. 2006) or analysis of fluorescence patterns (Pozhitkov et al. 2007a). Non-specific hybridizations usually wash off at lower temperatures than perfect match hybridizations. The calculated temperatures have been directly compared (Li et al. 2004; Starke et al. 2006), used for determining future wash temperatures, or used as input in neural network classification (Pozhitkov et al. 2005). While these NEDC analyses provide information on the specificity of hybridization events, they all rely on the use of MM probes, complex analyses, or multiple experiments. Additionally, use of these dissociation indices to compare PM and MM dissociation rates can lead to erroneous conclusions (Pozhitkov et al. 2007b; Pozhitkov et al. 2007c).

This study presents a simple method for identification of non-specific hybridizations by using a new parameter that does not rely on comparison of PM and MM dissociations. The parameter is the ratio of a probe's measured maximum specific dissociation temperature (T_{d-w}) to its theoretical melting temperature for a perfectly matching hybridization (T_m). The parameter can be obtained by automated fitting of a four-parameter, sigmoidally-shaped, asymmetric empirical equation. Automated fitting greatly reduces analysis time and potential for error associated with the thousands of curves generated in a typical PM vs. MM hybridization experiment using an NEDC approach. We also tested the ability of T_{d-w}/T_m to

identify non-specific hybridizations with three different datasets including probes designed from 16S rRNA and functional genes with varying degrees of sequence similarity to the targets in solution. This empirical approach can be applied to high throughput hybridization data for reliably identifying microbes in complex environmental matrices.

2. Materials and methods

2.1. Microarray synthesis

Microarrays for each dataset were synthesized by Xeotron Corporation, Houston, TX (now part of Invitrogen, Carlsbad, CA) using a proprietary *in situ* synthesis technology developed at the University of Michigan (Gao et al. 2001). The oligonucleotides were synthesized with an estimated density of 1 molecule per 200 square angstroms.

2.2. Probe design

16S rRNA gene array: Probes for the *Burkholderia xenovorans* strain LB400 were designed with the following steps: i) all possible non-overlapping 20-mer sequences in the entire 16S rRNA gene were used as probes (generating 73 probes). ii) Additional probes were generated by overlapping 19 bases in the regions of the 16S rRNA gene with less than 50 exact matches (determined by performing BLAST for probes from the first step with the RDP-II database). A total of 209 probes (20-mer) with perfect matches to the 16S rRNA gene of LB400 were obtained with this design scheme. Four replicates of each probe were synthesized on the microarray. Mismatch probes were also designed to contain two randomly generated mismatches in positions 7 and 14. A total of 931 PM probes targeting *Burkholderia xenovorans* LB400 16S rRNA gene are on the array. Also included are approximately 6,805 probes with one or more MM to LB400. These probes were designed to target various levels of phylogeny for microorganisms typically in anaerobic communities.

Virulence and marker gene array: Oligonucleotide 18-mer probes were designed for 17 pathogens targeting 93 virulence and marker genes from 671 sequences retrieved from GenBank (May 2004). Probes were designed to be complementary to all sequences of a gene for a given species and contain at least two mismatches to all other sequences. Mismatch probes were designed with one random mismatch in the center of the probe. More on this array can be found in a recently published manuscript by our group (Miller et al. 2008).

E. coli strain fingerprinting array: Sequences of 11 genes from *Escherichia coli* strain O157:H7 RIMD 0509952 (Hayashi et al. 2001) and one gene from *Shigella flexneri* 2A 2457T were used to design 935 18-mer oligonucleotide probes. For 352 of the perfect match probes, three single mismatch 18-mer probes were created randomly with respect to both position and type of mismatch for a total of 1056 mismatch probes. For 578 of the perfect match probes, additional 18-mer mismatch probes with a single mismatch at position 9 were designed. Finally, 20, 25, 35 and 45-mer, are also included. (Wick et al. 2006).

2.3. Preparation of DNA

16S rRNA gene array: Genomic DNA from pure cultures of *B. xenovorans* strain LB400 was extracted, amplified for the 16S rRNA gene, and labeled with amino-allyl-dUTP using Klenow polymerase. Genomic DNA was extracted using the protocol for gram positive bacteria from Qiagen DNeasy Tissue Kit (Qiagen, Valencia, CA). The 16S rRNA genes were amplified from the genomic DNA using a 27F (5'-AGAGTTTGATCMTGGCTCAG-3') and 1525R (5'-AAGGAGGTGWTGCARCC-3') primer pair and Platinum Taq DNA Polymerase (Invitrogen; Carlsbad, CA). Thirty cycles of the following temperature program were used for amplification: denaturation at 94°C for 30

s, annealing at 52°C for 45 s, and elongation at 72°C for 90 s. All PCR amplicons were cleaned using the Qiagen PCR purification kit (Qiagen Inc., Valencia, CA).

Amplified 16S rRNA gene product was labeled using the Bioprime DNA Labeling Kit (Invitrogen, San Diego, CA). The labeling protocol included a 90 min incubation of 250 ng of 16S rRNA amplified product with Klenow polymerase and 5:1 amino-allyl-dUTP:dTTP (Ambion, Austin, TX). All amino-allyl-dUTP-labeled products were cleaned using the Qiagen PCR purification kit with modified wash buffer (5 mM K₂HPO₄, pH 8.0, 80% EtOH) and elution buffer (4 mM K₂HPO₄, pH 8.5). Cyanine dye was attached by incubating 3 to 5 µg of amino-allyl-dUTP-labeled DNA for 1 h in 50:50 mixture of 0.1 M sodium carbonate buffer (pH 9.3) and N-hydroxysuccinimide ester Cyanine dye (prepared in fresh dimethyl-sulfoxide). DNA product from dye coupling was cleaned using the Qiagen PCR purification kit. Before hybridization, 200 pmol of both Cy3 and Cy5 labeled 16S rRNA gene product from LB400 were combined.

Virulence and marker gene array: Target DNA was amplified for 47 genes from 12 pathogens as described in another study (Miller et. al, 2008), amino-allyl-dUTP labeled with the Klenow-based Invitrogen DNA labeling system, coupled with Cy3 or Cy5 dyes, and 100 ng of each gene was hybridized to the array.

E. coli strain fingerprinting array: Target DNA was amplified and labeled as above. Fragments of 600 bp including the sequences targeted by the oligos on the chip were amplified from DNA of *E. coli* strain O157:H7 RIMD 0509952 (36) (*aroE*, *clpX*, *cstA*, *glyA*, *lysP*, *rpoS*, *mdh*, *stx1*, *stx2*, *eae* and *uidA*) and from DNA of *S. flexneri* 2A 2457T (*virA*). For the cross-hybridization experiment, 45 ng of PCR mix (sum of all 12 PCR products, i.e. each at 3.75 ng) and 3.7 µg of genomic DNA were combined. More details on the amplification are described in another study (Wick et al. 2006).

2.4. Hybridization and scanning

Target mixtures were prepared in 50 µl of hybridization solution containing 35% formamide, 0.4% Triton X-100, and 6X SSPE. Target DNA was denatured at 95°C for 3 min, cooled on ice for 1 min, and passed through a 0.22 µm Millex-GS syringe driven filter (Millipore, Billerica, MA) to remove particulates and prevent clogging of microfluidic channels on the microarray. All hybridizations were carried out in triplicate for 16 to 18 h at 20°C using an M-2 microfluidic station (Xeotron Corporation) (Wick et al. 2006). A flow rate of 500 µl per min was used for recirculation of hybridization solution through the microfluidic array during hybridization. After overnight hybridization at 20°C, the microarray was washed using wash buffer 2 (6X SSPE, 0.2% Triton X-100), wash buffer 4 (1 X SSPE, 0.2% Triton X-100), and wash buffer 2 with no Triton X-100 in series for 2.2 minutes each (500 µl per min, 20°C). A non-equilibrium thermal dissociation approach for the 16S rRNA gene array was achieved by a high stringency wash (10 mM Na₂HPO₄, 5 mM Na₂EDTA, pH 6.6; flow rate 500 µl per min) for 2.2 min at increasing temperatures from 20 to 70°C at 2 degree intervals. The other two arrays were washed for 1.4 minutes at increasing temperatures from 25 to 60°C at 1 degree intervals.

After each wash, chips were removed from the microfluidic station and scanned using a GenePix 4000B 16-bit laser scanner (Axon instruments, Union City, CA) with a resolution of 5 µm at Cy5 (635 nm laser) and Cy3 (532 nm laser) wavelengths. The PMT settings were kept constant for each scan. Fluorescence signal intensities were extracted from microarray images using GenePix5.0 (Axon Instruments, Union City, CA), yielding values between 0 and 65,535 arbitrary units for the 16-bit scanner. The median of 90 pixels, for each of 8,000 spots, was extracted by GenePix and stored in a text file.

2.5. Automatic extraction and processing of data using Microsoft Excel

Data from dissociation experiments using the GenePix scanner and software resulted in multiple “GenePix results” text files. A macro was written that extracted the probe labels and median value of spot signal intensities from each temperature file into two Microsoft Excel worksheets (one for each dye). Each temperature file was opened with Excel and three columns were automatically copied and pasted into the final sheet including the feature identification and median pixel intensities scanned at the wavelengths corresponding to the Cy5 (635 nm) and Cy3 dyes (532 nm). The data were arranged so that each row contained the signal intensities at each temperature point for one feature. Replicated features were averaged, extra rows removed, and data from Cy3 and Cy5 separated into two worksheets. The parameter estimates for non-linear regression were found using the Solver add-in with default values and the following constraints; β and γ parameters are less than or equal to zero, range is less than or equal to twice the maximum raw value, and the background is greater than or equal to the minimum raw value. Automatically calculated values included the following: i) start values for the parameter estimates, ii) sum of squared residuals, iii) numerically integrated area under the fitted equation, iv) T_{d-50} , v) T_{d-w} , and vi) temperature at maximum slope. Macros used for automatic processing of the data are provided as supplemental information.

2.6. Presence of targets and ROC curve analysis

Each probe sequence was compared to the gene sequences and reverse complements of the DNA spiked into the samples. Area under the ROC curve was calculated using Analyze-it for Microsoft Excel (Analyze-It Software, Ltd. Leeds, England).

2.7. Theoretical predictions of melting temperatures

Predictions of the T_m for the probe/target duplexes were obtained using Markham and Zuker’s DINAMelt Server, the two-state hybridization module (Markham and Zuker 2005). The probe and its reverse complement were used in the calculation with DNA, 43°C, 0.03 M $[Na^+]$, and 0 M $[Mg^{++}]$ for the energy rules and 0.01 μM for the strand concentration. A higher temperature (43°C) instead of the actual 20°C was used as the buffer contained 35% formamide, which is expected to destabilize duplexes in a way equivalent to increasing the temperature by 21–25°C (Blake and Delcourt 1996; McConaughy et al. 1969; Urakawa et al. 2002).

3. Results and discussion

3.1. An empirical equation to calculate T_{d-w}

Data generated from NEDCs i) show decrease in signal intensity with increasing temperature, ii) are constrained between two horizontal asymptotes, and iii) contain one inflection point. The inflection point is where the change in signal per unit temperature reaches a maximum. Data from NEDCs are not symmetric about the inflection point with the curve being more gradual at lower temperatures and tighter at high temperatures (Fig. 1A). This sigmoidal shape can be fitted with a four-parameter continuous equation (Equation 1) to facilitate calculations and summarize the data (Chechetkin et al. 2000; Ratkowsky 1990).

$$Signal(Temperature) = background + \frac{range}{\beta - \gamma * Temperature} \quad (1)$$

Equation 1 has asymmetric shape with lower root mean square error for residuals between the curve and raw data than a similar equation reported earlier (Chechetkin et al. 2000). The

influences of the equation's four parameters (background, range, β and γ) on the curve are shown in Figs. 1A–1C. The background and range estimates describe the locations of the horizontal asymptotes, while the β and γ estimates are used to match the shape of the non-equilibrium dissociation data. The curve approaches the background at lower temperatures for larger β values and smaller γ values.

Since nonlinear regression requires start values for parameter estimates, satisfactory start values for the four parameters were found to be: the minimum raw signal for background; the range of raw signals for range; -4 for β ; and -0.1 for γ . These start values are automatically calculated before each regression using raw data for each curve.

The ability of the curve to fit data was evaluated using 931 perfect match hybridizations from the 16S rRNA gene array (Figs. 1 and 2). The 95% confidence intervals for one fitted curve are shown in Fig. 1A. The average root mean square for residuals of the 931 perfect match probes is 108 ± 91 fluorescent units. The average R^2 value for the 931 curves is 0.998 ± 0.006 . The range and shapes of the 931 curves are shown below (Fig. 2A) as well as the studentized residuals plotted against temperature (Fig. 2B) and predicted signal (Fig. 2C). When plotted against predicted signal intensity, the residuals appear to be independent and normally distributed with a zero mean and common variance. However when plotted against temperature, the residuals appear to have larger variance at lower temperatures, which may be due to weakly associated non-specific hybridizations that wash away at lower temperatures.

Calculation of previously reported dissociation parameters is straightforward using parameter estimates from the fitted equation. The temperature at which the dissociation signal reaches half its original intensity (T_{d-50}) and the specific dissociation temperature (T_{d-w}) are calculated using Equations 2 and 3, respectively, using estimates for the β and γ parameters. The T_{d-w} estimate is the temperature where the second derivative of the equation fit to log-transformed signal intensities is equal to zero.

$$\widehat{T}_{d-50} = \frac{\widehat{\beta} - \ln(\ln(2))}{\widehat{\gamma}} \quad (2)$$

$$\widehat{T}_{d-w} = \frac{\widehat{\beta}}{\widehat{\gamma}} \quad (3)$$

3.2. Filtering data based on signal to noise ratio

Post-hybridization signal intensities are first used to remove the empty spots from the datasets, where no (or very low concentrations) of perfectly matching or non-specific DNA had hybridized. Since signal intensity was measured at each of 26 temperatures when developing the dissociation curves, an investigation to determine the best parameter for removing empty spots from the datasets was performed. Each of the single temperature intensities as well as the total area under the curve (AUC) were evaluated using receiver operating characteristic (ROC) curve analysis. ROC curves evaluate the ability of a parameter to discriminate between two groups based on true positive, false positive, true negative and false negative counts as the cut-off value of the parameter is changed (Parodi et al. 2005). The closer the area under the ROC curve is to 1.0 the better the parameter is for discrimination. The parameter used in the ROC curve analysis was signal to noise ratio (SNR), or the intensity of each feature divided by the median intensity of designated

background spots, which are randomly distributed spots on each microarray that contain no probes.

ROC curve analysis, as shown in Fig. 3A, indicated that i) SNR for AUC or SNR for 20–48°C are good filters with the area under the ROC curve above 0.96, and ii) high temperature (50–70 °C) SNR, are fair filters with area under the ROC between 0.90 and 0.96.

It was originally thought that using an area under the fitted equation (SNR – AUC in Fig. 3A) would reduce the variability found in single temperature ratios and thereby be a better parameter (Hashsham et al. 2004). The similarity in using SNR - AUC or single, low temperature hybridizations was found to be attributable to the distributions of the parameters. Although the area under the fitted equation does reduce the variability in the hybridization signal, its relative value is about half that of the single low temperature points, rendering its usefulness as a discriminator about equal.

The high discriminating ability of all SNRs (including 50–70°C SNR) is believed to be attributable to the high concentration of target DNA in this particular experiment. Follow up experiments with lower concentrations of target DNA might be able to identify a single SNR to use as a filter.

The observation that SNR at 70°C for the 16S rRNA gene experiment was a fair discriminator prompted a closer inspection of these high temperature signals. A plot of the signal intensities at 20°C vs. 70°C (Fig. 4) showed a positive correlation for probes with high signal, greater than 100 fluorescence units, at 20°C. This is evidence that not all of the duplexes are dissociated after a 70°C wash for this concentration of targets. Since the area under the ROC for SNR at 70°C is lower than that SNR as AUC or SNR at 20–48°C (Fig. 3B), these spots may include: both perfect match and non-specific hybridizations with high melting temperatures; or a higher percentage of perfect match hybridizations as well as a higher rate of false negatives (i.e., both perfect match hybridizations with low melting temperatures and non-specific hybridizations were washed away).

A SNR ratio of two was used in our analyses to separate signal intensities due to hybridization from signal intensities due to background. This was found to be an effective cut-off for an initial filter since the data for these very low signal intensities gave unreliable values for the second filter, T_{d-w}/T_m , which is used to distinguish between perfect match and non-specific hybridizations.

3.3. Identification of non-specific hybridizations using T_{d-w}/T_m

As mentioned in the introduction, this study uses the ratio of a probe's measured maximum specific dissociation temperature T_{d-w} to its theoretical melting temperature T_m , or T_{d-w}/T_m to discriminate between perfect match and non-specific hybridizations. Selection of a cut-off value for T_{d-w}/T_m was found by setting its value so that the false negative and false positive rates for the experiments were equal. This cut-off value was found to be 0.78 and only varied slightly, ± 0.03 , when tested with other experiments.

A comparison of using two filters, T_{d-w}/T_m greater than 0.78 and SNR greater than 2, versus solely using a SNR greater than 2 at single temperature points shows the advantages of using two filters. The ROC analysis in Fig. 3B was adapted for two parameters to show discrimination (percentage of true positives or negatives) by subtracting the average of the false positive and false negative percentages from one hundred percent. Discrimination was plotted for both T_{d-w}/T_m and SNR as well as SNR alone at temperatures from 20 to 70°C. The discrimination was 97% using both T_{d-w}/T_m and SNR, between 80% and 90% for SNR

alone at temperatures from 20–38°C, and below 70% for SNR alone at temperatures from 40–70°C.

3.4. Application of the Method to Three Datasets

Analysis was evaluated using three different microarrays and experimental designs including a: i) 16S rRNA gene or diversity array, ii) virulence and marker gene array, and iii) *E. coli* strain fingerprinting array (Fig. 5). In Fig. 5, the black hollow diamonds represent probes that were designed to perfectly complement the hybridized target and the grey hollow circles represent probes that do not perfectly complement the target. The black lines represent the SNR and T_{d-w}/T_m cut-offs. If a spot falls to the left of the T_{d-w}/T_m cut-off or below the SNR cut-off, it is determined that the target DNA is not in the sample. The insets in Fig. 5 show the correlation between T_{d-w} and T_m for each of the arrays. For VMG and *E. coli* strain fingerprinting array, some of the estimated T_{d-w} values were below 20°C. Since hybridization was performed at 20°C, T_{d-w} values lower than this value should not be possible. The majority of sub-20°C T_{d-w} values were calculated from non-specific dissociation curves that didn't have inflection points. The curve fitting algorithm struggled with these curves and shifted the T_{d-w} values below 20°C.

DNA from the *B. Xenovorans* strain LB400 16S rRNA gene was hybridized to the 16S rRNA gene or diversity array, which had many probes that targeted the 16S rRNA gene for *B. xenovorans* LB400 strain (LB400). For each 20mer probe complementary to the gene, there was a similar probe with mismatches at positions 7 and 14. The array also contained many probes with low similarity to the gene, which were included in discrimination calculations. The discrimination for this experiment was 97.2%, with only 2.8% of probes giving incorrect information about presence or absence (Fig. 5A).

DNA from 48 genes (13 pathogens) was hybridized to the virulence marker gene (VMG) array, which contains probes that target various functional genes from several pathogenic organisms. The array also contains mismatch probes designed with one random mismatch in the center of the probe and many probes that target genes that were not in the sample. The discrimination for this experiment was 96.4% (Fig. 5B).

A complex mixture of DNA as described in the methods was hybridized to the *E. coli* strain fingerprinting array, which contains probes that target various functional genes for *E. coli*. For each probe that targets a gene in the sample, there were three mismatch probes with random type and position. This includes mismatch probes with the mismatch synthesized at the very end of the probe. The discrimination for this experiment was lower at 86.3% (Fig. 5C). The lower discrimination is believed to be due to the high similarity between the mismatch probes and the targets, but may also have to do with concentration dependence of NEDC characteristics in complex target samples (Pozhitkov et al. 2008).

Table 1 shows discrimination (the percentage of true positives over all positives or true negatives over all negatives, which was quantified by setting the cut-off value so false positive and false negative rates are equal) calculated for each of the experiments using three different discrimination parameters; T_{d-w}/T_m , Max Slope Temp/ T_m , and T_{d-50}/T_m . Discrimination using the T_{d-w}/T_m parameter was generally better than discrimination using the other parameters. This advantage of using T_{d-w}/T_m is due to the better correlation between T_{d-w} and T_m for perfect match probes. T_{d-w} may correlate better with T_m due to advantages of fitting the data to an equation (e.g., all data points influence where the curve's inflection point will be whereas with other methods, selection of the temperature at which dissociation is maximum uses less of the data).

The data as shown in Table 1 were obtained from low complexity samples with minimal cross-hybridization. More complex samples will have more cross-hybridization. To test the influence of cross-hybridization on the T_{d-w}/T_m method, genomic DNA from *E. coli* Sakai spiked with 12 PCR products (~600 bp in length) was used for hybridization to the *E. coli* strain fingerprinting array. The ratio was 45 ng of PCR mix (sum of all 12 PCR products, i.e. each at 3.75 ng) and 3.7 μ g of genomic DNA (Wick et al. 2006). Using complex samples, it is difficult to determine whether the entire signal for a spot is due to cross-hybridization or a combination of cross-hybridization and perfect match target. However, using the T_{d-w}/T_m parameter, 75.7% discrimination was believed to be achieved.

Overall, the T_{d-w}/T_m method was remarkably successful in identifying non-specific hybridizations. One particular advantage of this parameter is its lack of correlation with signal intensity (Fig. 5). This lack of dependence on signal intensity makes the parameter equally effective across all signal intensities for screening against non-specific hybridizations. Ordinarily, high signal intensities at a single temperature would be considered as true positive signals and not cross-hybridization. This empirical method is successful in screening a high degree of non-specific discriminations in every data set except non-specific hybridizations with high sequence similarity. Therefore, it can be applied to the high throughput microarray data analysis for microbial detection but may not be very useful for single nucleotide polymorphisms analysis.

3.5. Usefulness of this approach for other platforms

This approach raises an important question i.e., does high signal to noise ratio automatically signify hybridization by the intended target? The answer is obviously no. Thus there is a need to prove that a high SNR originates from specific hybridization. The approach adopted for this determination will depend on the type of platform (e.g. glass slide arrays, NimbleGen, Agilent, and the arrays described herein). Because obtaining NEDCs is not a routine procedure and rather cumbersome for most platforms, the question is how to incorporate this information in other platforms. One alternative is to carry out specificity validation at varying temperatures to obtain T_{d-w}/T_m , similar to presented here and demonstrate that the high SNRs are indeed specific. A second alternative is to use increasing formamide concentration as an equivalent for increase in temperature and establish the specificity (similar to fluorescent in situ hybridization optimization studies). Overall, array data with probes optimized at a single temperature to obtain high SNRs should be interpreted with caution especially when such arrays are used for analyzing complex microbial community samples.

4. Conclusions

In conclusion, the new T_{d-w}/T_m parameter was successful for screening against many non-specific hybridizations that could not be identified using single temperature signal intensities alone even in complex samples. Furthermore, empirical modeling allowed a simplified approach to the high throughput analysis required for microarray experiments.

Acknowledgments

This work was supported in part by grants from Superfund Basic Research Program grant P42 ES004911-17 from the US National Institutes of Environmental Health Sciences, NIH (1RO1 RR018625 03), Michigan Economic Development Corporation (GR467 PO085P3000517) and MSU Foundation.

References

- Bains W, Smith GC. A novel method for nucleic-acid sequence determination. *J Theor Biol.* 1988; 135:303–307. [PubMed: 3256722]

- Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, Small PM. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science*. 1999; 284:1520–1523. [PubMed: 10348738]
- Binder H, Krohn K, Burden CJ. Washing scaling of GeneChip microarray expression. *BMC Bioinformatics*. 2010; 11
- Blake RD, Delcourt SG. Thermodynamic effects of formamide on DNA stability. *Nucleic Acids Res*. 1996; 24:2095–2103. [PubMed: 8668541]
- Bodrossy L, Sessitsch A. Oligonucleotide microarrays in microbial diagnostics. *Curr Opin Microbiol*. 2004; 7:245–254. [PubMed: 15196491]
- Chandler DP, Jarrell AE. Automated purification and suspension array detection of 16S rRNA from soil and sediment extracts by using tunable surface microparticles. *Appl Environ Microbiol*. 2004; 70:2621–2631. [PubMed: 15128511]
- Chandler DP, Jarrell AE. Taking arrays from the lab to the field: trying to make sense of the unknown. *Biotechniques*. 2005; 38:591–600. [PubMed: 15884677]
- Chechetkin VR, Turygin AY, Proudnikov DY, Prokopenko DV, Kirillov EV, Mirzabekov AD. Sequencing by hybridization with the generic 6-mer oligonucleotide microarray: an advanced scheme for data processing. *J Biomol Struct Dyn*. 2000; 18:83–101. [PubMed: 11021654]
- Drmanac R, Labat I, Brukner I, Crkvenjakov R. Sequencing of megabase plus DNA by hybridization - Theory of the method. *Genomics*. 1989; 4:114–128. [PubMed: 2737674]
- Feldkamp U, Wacker R, Schroeder H, Banzhaf W, Niemeyer CM. Microarray-based in vitro evaluation of DNA oligomer libraries designed in silico. *Chemphyschem*. 2004; 5:367–372. [PubMed: 15067873]
- Gao XL, LeProust E, Zhang H, Srivannavit O, Gulari E, Yu PL, Nishiguchi C, Xiang Q, Zhou XC. A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res*. 2001; 29:4744–4750. [PubMed: 11713325]
- Gerry NP, Witowski NE, Day J, Hammer RP, Barany G, Barany F. Universal DNA microarray method for multiplex detection of low abundance point mutations. *J Mol Biol*. 1999; 292:251–262. [PubMed: 10493873]
- Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, Brody LC, Wang D, Lander ES, Lipshutz R, Fodor SPA, Collins FS. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet*. 1999; 22:164–167. [PubMed: 10369258]
- Hashsham SA, Wick LM, Rouillard JM, Gulari E, Tiedje JM. Potential of DNA microarrays for developing parallel detection tools (PDTs) for microorganisms relevant to biodefense and related research needs. *Biosens Bioelectron*. 2004; 20:668–683. [PubMed: 15522582]
- Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157: H7 and genomic comparison with a laboratory strain K-12. *DNA Res*. 2001; 8:11–22. [PubMed: 11258796]
- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*. 1992; 258:818–821. [PubMed: 1359641]
- Khomiakova EB, Livshits MA, Sharonov A, Prokopenko DV, Mirzabekov AD. Analysis of perfect and mismatched DNA duplexes by a generic hexanucleotide microchip. *Mol Biol*. 2003; 37:726–741.
- Li ES, Ng JK, Wu JH, Liu WT. Evaluating single-base-pair discriminating capability of planar oligonucleotide microchips using a non-equilibrium dissociation approach. *Environ Microbiol*. 2004; 6:1197–1202. [PubMed: 15479252]
- Liebich J, Schadt CW, Chong SC, He ZL, Rhee SK, Zhou JZ. Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. *Appl Environ Microbiol*. 2006; 72:1688–1691. [PubMed: 16461729]

- Liu WT, Mirzabekov AD, Stahl DA. Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. *Environ Microbiol.* 2001; 3:619–629. [PubMed: 11722542]
- Lockhart DJ, Dong HL, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang CW, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol.* 1996; 14:1675–1680. [PubMed: 9634850]
- Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J, Schleifer KH, Wagner M. Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl Environ Microbiol.* 2002; 68:5064–5081. [PubMed: 12324358]
- Markham NR, Zuker M. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.* 2005; 33:W577–W581. [PubMed: 15980540]
- McConaughty BL, Laird CD, McCarthy BJ. Nucleic acid reassociation in formamide. *Biochemistry.* 1969; 8:3289–3295. [PubMed: 4980190]
- Miller SM, Turlousse DM, Stedtfeld RD, Baushke SW, Herzog AB, Wick LM, Rouillard JM, Gulari E, Tiedje JM, Hashsham SA. In situ-synthesized virulence and marker gene biochip for detection of bacterial pathogens in water. *Appl Environ Microbiol.* 2008; 74:2200–2209. [PubMed: 18245235]
- Mobarry BK, Wagner M, Urbain V, Rittmann BE, Stahl DA. Phylogenetic probes for analyzing abundance and spatial organization of nitrifying bacteria. *Appl Environ Microbiol.* 1996; 62:2156–2162. [PubMed: 8787412]
- Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. *Clin Genet.* 2004; 66:488–495. [PubMed: 15521975]
- Parodi S, Izzotti A, Muselli M. Re: The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *J Natl Cancer Inst.* 2005; 97:234–235. [PubMed: 15687368]
- Peplies J, Glockner FO, Amann R. Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. *Appl Environ Microbiol.* 2003; 69:1397–1407. [PubMed: 12620822]
- Pozhitkov A, Chernov B, Yershov G, Noble PA. Evaluation of gel-pad oligonucleotide microarray technology by using artificial neural networks. *Appl Environ Microbiol.* 2005; 71:8663–8676. [PubMed: 16332861]
- Pozhitkov A, Noble PA, Domazet-Loso T, Nolte AW, Sonnenberg R, Staehler P, Beier M, Tautz D. Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucleic Acids Res.* 2006; 34:e66. [PubMed: 16707658]
- Pozhitkov AE, Bailey KD, Noble PA. Development of a statistically robust quantification method for microorganisms in mixtures using oligonucleotide microarrays. *J Microbiol Meth.* 2007a; 70:292–300.
- Pozhitkov AE, Boube I, Brouwer MH, Noble PA. Beyond Affymetrix arrays: expanding the set of known hybridization isotherms and observing pre-wash signal intensities. *Nucleic Acids Research.* 2010; 38
- Pozhitkov AE, Rule RA, Stedtfeld RD, Hashsham SA, Noble PA. Concentration dependency of nonequilibrium thermal dissociation curves in complex target samples. *J Microbiol Methods.* 2008; 74:82–88. [PubMed: 18471911]
- Pozhitkov AE, Stedtfeld RD, Hashsham SA, Noble PA. Revision of the nonequilibrium thermal dissociation and stringent washing approaches for identification of mixed nucleic acid targets by microarrays. *Nucleic Acids Res.* 2007b; 35:e70. [PubMed: 17430966]
- Pozhitkov AE, Tautz D, Noble PA. Oligonucleotide microarrays: Widely applied--poorly understood. *Brief Funct Genomic Proteomic.* 2007c; 6:141–148. [PubMed: 17644526]
- Ratkowsky, DA. *Handbook of Nonlinear Regression Models.* Marcel Dekker, Inc; New York, New York: 1990.
- Schena M, Shalon D, Davis RW, Brown PO. Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray. *Science.* 1995; 270:467–470. [PubMed: 7569999]

- Starke EM, Smoot JC, Smoot LM, Liu WT, Chandler DP, Lee HH, Stahl DA. Technology development to explore the relationship between oral health and the oral microbial community. *BMC Oral Health*. 2006; 6(Suppl 1):S10. [PubMed: 16934111]
- Urakawa H, Noble PA, El Fantroussi S, Kelly JJ, Stahl DA. Single-base-pair discrimination of terminal mismatches by using oligonucleotide microarrays and neural network analyses. *Appl Environ Microbiol*. 2002; 68:235–244. [PubMed: 11772632]
- Wick LM, Rouillard JM, Whittam TS, Gulari E, Tiedje JM, Hashsham SA. On-chip non-equilibrium dissociation curves and dissociation rate constants as methods to assess specificity of oligonucleotide probes. *Nucleic Acids Res*. 2006; 34:e26. [PubMed: 16478712]
- Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol*. 1997; 15:1359–1367. [PubMed: 9415887]
- Zhou JZ, Thompson DK. Challenges in applying microarrays to environmental studies. *Curr Opin Biotechnol*. 2002; 13:204–207. [PubMed: 12180093]

Highlights

- Automated empirical modeling approach for analysis of high-throughput microarrays.
- Identification of non-specific microarray signals using a new parameter (T_{d-w}/T_m).
- Method does not rely on comparison of perfect match and mismatch dissociations.
- Evaluated three datasets with the new parameter.

Fig. 1.

An empirical, sigmoidally shaped curve to model dissociation data. The empirical curve described in Equation 1. (A) The curve fitted to data from a perfect match probe/target duplex as well as visualization of the inflection point, asymptotes, and influences of background and range parameters. The 95% confidence limits are shown in grey. (B) The influence of the β parameter on the equation. More negative β values fit to curves that melt at higher temperatures. (C) The influence of the γ parameter on the equation. More negative γ values are fit to curves that melt at lower temperatures.

Fig. 2. The range and shape of fitted curves and residuals. (A) The range and shapes of 931 curves fit to perfect match dissociations. (B) The studentized residuals for the same curves plotted against temperature (C) The studentized residuals for the same curves plotted against predicted signal intensity.

Fig. 3. ROC curve analysis. (A) Area under the ROC curve is shown for SNR as AUC and SNR from 20–70°C. (B) Discrimination is shown for SNR and T_{d-w}/T_m from 20–70°C.

Fig. 4.
Correlation between signal to noise ratios at 20 °C and 70 °C.

Fig. 5. Identification of non-specific hybridizations using T_{d-w}/T_m . Discrimination of (A) 97.2%, (B) 96.4%, and (C) 86.3% are obtained for three datasets with T_{d-w}/T_m cut-off equal to 0.78. Dotted lines are at SNR equal to 2 and solid lines are at the T_{d-w}/T_m cut-off. Black, hollow diamonds represent probes that were designed to perfectly complement the hybridized target. Light-grey, hollow circles represent probes that do not perfectly complement the target. Inset: T_{d-w} vs. T_m with the solid T_{d-w}/T_m cut-off line.

Table 1

Discrimination for Three Experiments using Different Dissociation Parameters

	T_{d-w}/T_m	Max Slope Temp/ T_m	T_{d-50}/T_m
16S rRNA gene array	97.2	94.6	94.1
VMG array	96.4	96.3	96.8
<i>E. coli</i> array	86.3	78.2	77.8