Journal of Microbiological Methods 76 (2009) 188–195

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Use of hidden correlations in short oligonucleotide array data are insufficient for accurate quantification of nucleic acid targets in complex target mixtures

Rebecca A. Rule [a], Alex E. Pozhitkov [b], Peter A. Noble [a,*]

[a] Civil and Environmental Engineering, University of Washington, Seattle, WA, 98195 USA
[b] College of Marine Sciences, University of Southern Mississippi, 703 E. Beach Drive, Ocean Springs, MS, 39566 USA

ARTICLE INFO

Article history:
Received 15 September 2008
Received in revised form 17 October 2008
Accepted 20 October 2008
Available online 30 October 2008

Keywords:
Microbial ecology
Optimization algorithm
Cross-hybridization
Free energy
rRNA
Nonspecific hybridizations
Oligonucleotide arrays

ABSTRACT

Nonspecific target binding (*i.e.*, cross-hybridization) is a major challenge for interpreting oligonucleotide microarray results because it is difficult to determine what portion of the signal is due to binding of complementary (specific) targets to a probe versus that due to binding of nonspecific targets. Solving this challenge would be a major accomplishment in microarray research potentially allowing quantification of targets in biological samples. Marcelino et al. recently described a new approach that reportedly solves this challenge by iteratively deconvoluting 'true' specific signal from raw signal, and quantifying ribosomal (rRNA) sequences in artificial and natural communities (*i.e.*, "Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data", Proc. Natl. Acad. Sci. 103, 13629–13634). We evaluated their approach using high-density oligonucleotide microarrays and Latin-square designed experiments consisting of 6 and 8 rRNA targets in 16 different artificial mixtures. Our results show that contrary to the claims in the article, the hidden correlations in the microarray data are insufficient for accurate quantification of nucleic acid targets in complex artificial target mixtures.

1. Introduction

Many fields of biomedical and environmental science depend upon accurate quantification of specific nucleic acid targets because these targets indirectly reflect the condition or state of a biological system. Oligonucleotide microarrays offer a significant potential to accurately quantify multiple nucleic acid targets in a biological sample because they typically contain hundreds of thousands of different immobilized probes, with each probe acting as an individual sensor with its own specificity and sensitivity to different nucleic acid targets in solution (Bishop et al., 2008). The intensity of a signal from the probes provides a measure of the amount of bound nucleic acid target in a sample. However, interpreting the signal in terms of bound targets is problematic because there is a multitude of possible interactions that can occur between nucleic acid targets in solution and an oligonucleotide probe on a microarray surface (Zhang et al., 2005, Pozhitkov et al., 2006). Several approaches have been used to remove or minimize the effects of nonspecific target binding: (i) subtracting nonspecific signal from specific signal by using oligonucleotide probes with one- or two-internal mismatches (Lipshutz et al., 1999; DeSantis et al., 2005), (ii) attempting to design oligonucleotide probes that have high specificity and sensitivity to the nucleic acid target of interest (Mei et al., 2003; Pozhitkov et al., 2005a; Rouillard et al., 2002; He et al., 2005; Li et al., 2005; Binder and Preibisch, 2005; Matveeva et al., 2003; Lipshutz et al., 1999), and (iii) scanning through a wide range of stringency conditions so that nonspecific targets are removed (Pozhitkov et al., 2005b, 2007a, 2008a). So far, not one of the above approaches has been shown to be effective at avoiding or accounting for nonspecific target binding, particularly in mixed target samples (Pozhitkov et al., 2007b).

A recent study by Marcelino et al. (2006) described a new approach that iteratively deconvolutes 'true' probe-target signal from raw signal affected by nonspecific target binding. The 'true' probe-target signal was then used to quantify microbial targets in a biological sample. At face value, the new approach, henceforth referred to as the hidden-correlation-based quantification (HCQ) approach, seems quite appealing since it promises to alleviate problems associated with interpreting the signal from a microarray. These problems arise from the fact that an immobilized oligonucleotide probe naturally has different binding energies to different targets, and targets typically occur at different concentrations in a biological sample. Hence, the physical meaning of probe signal intensity in microarray data is very difficult to interpret because it depends on both the binding energies of hybridized probe-target duplexes and the concentration of the targets in solution. The ability to calculate the probability of cross-hybridization for each probe on a microarray to every potential target, if proven valid, would ensure the specificity of a probe to a target and thus aid in determining the concentration of targets in a biological sample. The HCQ approach is especially appealing because it seems to be based on physicochemical understanding of the hybridization phenomenon.

* Corresponding author. Tel.: +1 206 685 7583; fax: +1 206 685 3638.
E-mail addresses: rule77@u.washington.edu (R.A. Rule), Alexander.Pozhitkov@usm.edu (A.E. Pozhitkov), panoble@washington.edu (P.A. Noble).

0167-7012/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2008.10.011

Given that the HCQ approach might represent a significant breakthrough in microarray research, we evaluated the approach using data sets from a previous study (Pozhitkov et al., 2008b). These data sets are based on 3160 probes (in duplicate) and are ideal for evaluating the HCQ approach for 5 reasons: (i) one data set contains individual target hybridization patterns (which throughout this manuscript we refer to as fingerprints at 1 nM conc.), that could be used to calibrate the cross-hybridization predictor, (ii) the other two sets contain Latin-square designed mixtures (n=16 different mixes), i.e., known targets at known concentrations, that could be used to evaluate the accuracy of predicted target concentrations, (iii) only *in vitro* transcribed rRNAs were used as targets in these data sets, so problems associated with the purity of target rRNA extracted from cells are avoided, (iv) the array probes were directly synthesized on the array surface, avoiding potential problems associated with the quality of spotting (probe attachment), and (v) the data sets have been previously used to demonstrate the utility of another approach that accurately quantified multiple nucleic acid targets in complex target mixtures (Pozhitkov et al., 2008b).

The fingerprint data set might also be useful to determine if the length of immobilized array probes affects the probability of cross-hybridization. While the Marcelino et al. (2006) study reported use of long probes (50 to 75 nt), our previous study used short probes (25 nt only). Hence, the current study is based on 25-mer probes as well. Marcelino et al. (2006) claimed extensibility of the HCQ to such short oligonucleotides. In addition, it is well established that short probes have a higher specificity and lower sensitivity to targets than long probes (Suzuki et al., 2007; Relogio et al., 2002; Letowski et al., 2004). Thus, we reasoned that short probes would have a lower probability of cross-hybridization than long probes for the same sequence identity (i.e., sequence similarity), which we would be able to test.

The specific objectives for this study are: (i) to repeat the HCQ approach using specified target and probe sequences in order to ensure proper understanding and to confirm previous HCQ results, (ii) to use the fingerprint data set to calibrate the cross-hybridization predictor in order to compare hybridization probabilities of microarrays with different probe lengths, and (iii) to compare the predicted target concentrations obtained using the HCQ approach to actual target concentrations using data from the Latin-square designed mixture experiments.

2. Materials and methods

In the Marcelino et al. (2006) study, two different variables were referred to as b, and variables s and s_c are also different variables. To minimize confusion, in our study, we refer to one of the b variables as d (see Eqs. (3) and (4)), and variable s_c as c (see Eqs. (1) and (2)).

2.1. Analytical cross-hybridization predictor

The binding energies (Gs) of mismatch target, k, and perfect match target, j, to each microarray probe, j, were calculated by using mfold (http://dinamelt.bioinfo.rpi.edu/twostate.php). The G_{jk}/G_{jj} ratio can be approximated by comparing sequence identity using the following equation:

$$\frac{G_{jk}}{G_{jj}} = s \quad 1 \quad 1 \quad s^2 \quad \text{when s was z c; else} \frac{G_{jk}}{G_{jj}} = c \quad 1 \quad 1 \quad c^2 \qquad (1)$$

where s equals the number of matches for a probe-target duplex divided by probe length, and is related to the microarray-corrected similarity, c by the following equation:

$$c = \frac{(\quad 1)}{8} + \frac{1}{2} \qquad (2)$$

The original cross-hybridization ($_{jk}$) predictor, as described by Marcelino et al. (2006), is:

$$_{jk} = \frac{1}{d}\left(1 \quad \frac{1}{1 + (1 \quad d)^2 e^{h \frac{G_{jk}}{G_{jj}} 1}}\right) \qquad (3)$$

where d and h are values that are fitted to experimental data, and the ratio of the free energies is used as specified by rules outlined in Eq. (1).

The correct cross-hybridization ($_{jk}$) between probe and targets can be estimated using the following equation (Marcellino, personnal communication):

$$_{jk} = \frac{1}{d}\left(1 \quad \frac{1}{1 + (1 \quad d)^2 \quad 1 \quad e^{h \frac{G_{jk}}{G_{jj}} 1}}\right) \qquad (4)$$

The signal intensity of a given probe on an array spot can be modeled by the following equation:

$$Y_j = \ln(+ b_{jj} \quad _{jk} \; _k) + \; _j \qquad (5)$$

where Y_j is the mean natural logarithm of the observed hybridization intensity for a given target species j, in the presence of background v, b is the unequal specific response of perfect match probe-target pairs, is the noise, $ø_k$ is the abundance of target k in the original sample, and $_{jk}$ describes the cross-hybridization between probe j and target k for each probe-target pair.

2.2. Optimization algorithm

All optimization of the data (Eqs. (1)–(5)) was accomplished using Microsoft (MS) Excel Solver rather than best-linear-unbiased-estimation (BLUE) used in the HCQ study, because the data was already in a MS Excel spreadsheet. Unless specified otherwise, initial values for Eqs. (1)–(4) were set to those used by the HCQ study. In the case of Eq. (5), initial values of unknowns v, b, , and , were based on values obtained in the fingerprint data set. Specifically, the mean background v, based on measured signal intensity of 142 negative control spots per array experiment, was 576±143 a.u. (n=1144). The signal intensity of a perfect match probe-target pair, b, was set to a constant due to unequal specific response among different probe-target pairs (i.e., as defined by HCQ study as the difference in binding affinities of probes to targets yielding different signal intensities). The average signal intensities of perfect match probe-target pairs based on individual target hybridizations (fingerprints) was found to be 4861±6,777 a.u. (n=1751). The initial rRNA target abundance, °, was estimated to be half of the maximum target concentration in the Latin-square experiment, which in the first Latin-square experiment was 0.048 nM and in the second Latin-square experiment was 0.090 nM. The noise, °, associated with higher signal intensities was set to zero. The criterion to determine convergence of the algorithm was the relative difference between consecutive iterations b0.001%. Typically 20 to 50 iterations were sufficient to reach convergence.

2.3. Calibration of the cross-hybridization predictor

Calibration of $_{jk}$ was accomplished using the fingerprint data. By comparing signal intensities of probes across six arrays, each array representing a different target, it was possible to assess the amount of cross-hybridization for each probe. We only selected those probes that were identified to be a perfect match to one target and mismatched to

the other 5 targets. For each of the selected probes, we calculated the ratio of signal intensity, which was defined as the signal intensity of a mismatch target divided by the signal intensity of the perfect match target.

Detailed experimental procedures are available in Supporting Text in Appendix A.

3. Results and discussion

As an aid to help readers understand both the equations above, and the HCQ approach, we have provided 18 supplementary files at: http://staff.washington.edu/pozhit/default.htm. Four of the supplementary files provide instructions that guide the reader through the accompanied MS Excel spreadsheet files containing the equations and microarray data.

The reason for reproducing the results of the HCQ study was to ensure that we correctly interpreted the equations. To minimize confusion in the comparison of the HCQ study to ours (i.e., this study), the results are presented in the same order as the HCQ study. The HCQ requires first establishment of a cross-hybridization predictor, which in turn depends on the variables of , *d* and *h*. Next, quantification can be performed using Eq. (5). Below, we have divided the Results and discussion into six sub-sections: (i) the determination of , (ii) calibration of the cross-hybridization predictor (jk), (iii) determining of target concentrations (k) in Latin-square designed experiments using HCQ, (iv) attempts to improve HCQ, (v) reasons for the poor quantification, and (vi) comparison of the results obtained by the HCQ approach with the previously published approach by our group.

3.1. Determination of

3.1.1. Unexplained variables

The HCQ study provided the following equation to determine based on values of T_c and T':

$$= \frac{1}{T_c \quad T^0} \qquad \text{ð6Þ}$$

However, neither T_c nor T' were defined. The HCQ study provides two examples of corresponding to different hybridization temperatures (Table 1). According to the original study, at a hybridization temperature of 60 °C, =1.26, while at a hybridization temperature of 37 °C, =0.83. Since the values for T_c and T' are not explicitly stated and their meanings are not defined, we attempted to determine the values of each by assuming that T_c is the hybridization temperature. When we solved for T', given the example values for and hybridization temperatures provided, the solution resulted in different values of T' for the two examples. This result suggests that T' is not a constant. Clearly, there was not enough information provided to determine the value of at other specific hybridization temperatures. This finding is important and relevant to our study because our experimental data set was based on a hybridization temperature of 45 °C. As a consequence, we were not able to use Eq. (6) to determine for either our data set or the "recreated" HCQ study data set. Instead, we determined by calculating the G_{jk}/G_{jj} ratios of probe-target duplexes, averaging the ratios by sequence identity, and fitting for using Solver and Eq. (1) (see Sup1.doc and Sup2.xls). Here, we define the recreated data set as target sequences that we best guessed, based on the original description in the Marcelino et al. paper, because of the unavailability of actual target sequences.

In addition, the HCQ approach requires a *c* value for making a cross-hybridization predictor. Apparently, *c* is a threshold value below which differences in sequence similarity do not result in any significant changes in G_{jk}/G_{jj} ratios. The original study used to calculate *c* (see Eq. (2)). As we stated above, could not be calculated from the intensive variable of temperature alone. We initially assumed *c*=0.48 and did not consider any duplexes having sequence similarity below 48%. The fitted value of determined the final value of *c* through Eq. (2).

Of note, one might argue that a linear interpolation of the two examples and hybridization temperatures would be sufficient to calculate at 45 °C. The value of the linear interpolation is shown in Table 1, and was not close to the value obtained by other means (see discussion below).

3.1.2. Analysis of values obtained using the recreated data set and our own experimental data set

Fig. 1b shows the results obtained by (i) determining all possible hybridizations that could take place between the probes and rRNA targets used in the HCQ study having 48% sequence similarity, (ii) calculating their corresponding binding energies using mfold (setting of 60 °C and RNA) and then (iii) determining the G_{jk}/G_{jj} ratios. The two theoretical lines are based on Eqs. (1) and (2). The black line was made by Solver, and represents the average G_{jk}/G_{jj} ratios by sequence identity for the recreated data set. For the black line, we found that =0.90. The grey line represents the value of average G_{jk}/G_{jj} by sequence identity as reported by the original HCQ study, where the value of was suggested to be 1.26. Although there is a prominent difference in the curves, the difference did not affect the fitted cross-hybridization prediction shown in Fig. 1a. Here, compensatory changes occurred in the value of *d* and *h*.

For our experimental fingerprint data set, we found that the value of =0.66 (mfold settings of 45 °C and RNA for the nucleic acid type) (Fig. 2b), which turned out to be less than: (i) that reported for HCQ study (=1.26), and (ii) that determined by our fitting of the recreated HCQ data set (=0.90) (Fig. 1b). Apparently, fitting for on our own dataset results in a value dissimilar to those reported by Marcelino et al. (2006).

3.2. Determination of jk

The equation used to produce the probability of cross-hybridization curve (jk) in the HCQ study (Eq. (3)) did not yield unity for perfect match hybridization, and therefore was incorrect. For example, setting the free energy ratios to 1 and using their coefficient values of *d*=0.75 and *h*=8.47, yielded a probability of cross-hybridization that was greater than 1. Through communications with the corresponding author, a revised equation was provided (L. Marcelino, pers. comm., 2008) (Eq. (4)). We applied the revised equation to the digitized data from Fig. 1a of the HCQ study, and were able to confirm that the values stated in their paper (=1.26, *d*=0.75, *h*=8.4) matched the revised equation (see Fig. 1a).

3.2.1. Determination of jk using the recreated data set

Note that the values for *d* and *h* for =0.90 were different than those reported by HCQ study. We have emphasized the relationship among these variables because *d* and *h* were defined to represent specific physicochemical entities that should not change between the HCQ study and our attempt to reproduce their same data set. These results suggest that neither *d* nor *h* represent the physicochemical entities that were declared by Marcelino et al., or else our diligent attempts to reproduce the method described in the HCQ study were not successful due to insufficient detail in the original manuscript.

Table 1
Summary of values by probe length, temperature, and method

Probe length	Temperature	Method	Reference	
50–75 nt	37 °C	Eq. (6)	Marcelino et al. (2006)	0.83
50–75 nt	60 °C	Eq. (6)	Marcelino et al. (2006)	1.26
25 nt	45 °C	Interpolation	This study	0.98
25 nt	45 °C	Eqs. (1) and (2)	This study	0.66

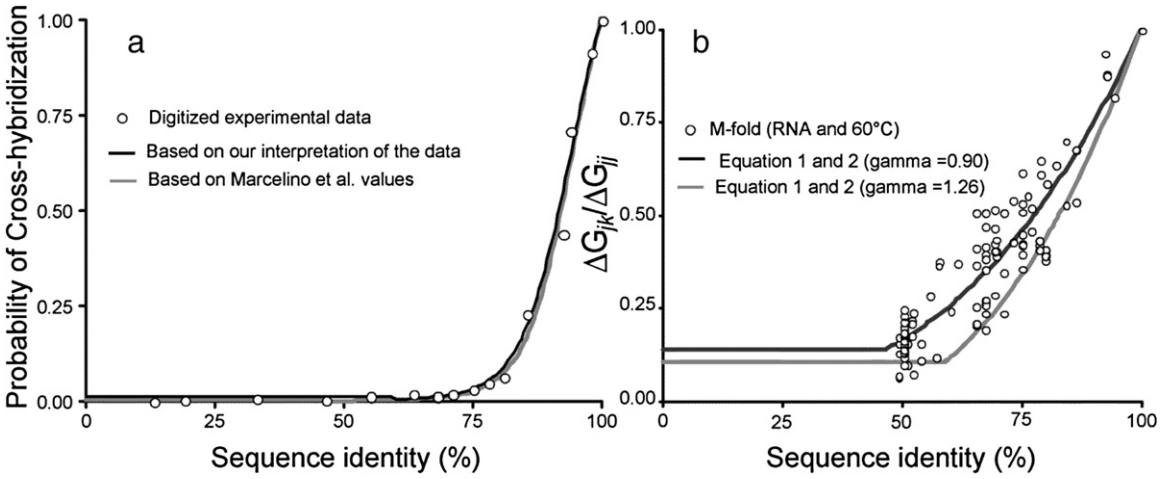

Fig. 1. Reproduction of the HCQ study. (a) Cross-hybridization curve (black line) obtained using the Eq. (4) with values =1.26, d=0.75, h=8.47 as stated in the HCQ study and our curve with values =0.90, d=1×10^{S7}, h=10.72. Digitized experimental data (open circles) were obtained from Fig. 1a in the HCQ study. (b) Comparison of G_{jk}/ G_{jj} ratios, as a function of sequence identity (n=139 probes). Grey line, derived from HCQ study (=1.26); black line, based on our fitting of the data (=0.90). Target sequences were 23S and 16S rRNA of *Vibrio vulni cus, Escherichia coli, V. cholera, Listonella anguillarum, V. alginolyticus*.

3.2.2. Determination of $_{jk}$ using the ngerprint data set

Before calibrating the cross-hybridization predictor, we established the relationship between the average G_{jk}/ G_{jj} ratios and sequence identity for 6582 probe-target duplexes (Fig. 2b). In total, cross-hybridization values (signal intensity ratios) were calculated for 1012 probes, which yielded 6582 values (n=1,097 perfect match probes with 5 mismatch targets). The raw and averaged signal intensity ratios for each probe as a function of the sequence identity are shown in Fig. 2a. Note that there was no defined pattern that could be visually resolved in the raw data, as depicted by the grey dots. However, when the signal intensity ratios were averaged by their specific sequence identity, the 'hidden correlations' between cross-hybridization and sequence identity (as defined by the HCQ study) were revealed, as depicted by the white circles in Fig. 2a. Thirteen cross-hybridization values were calculated for sequence identities ranging from 48 to 96%, with each value representing the average of more than 194 signal intensity ratios (see Sup3.doc and Sup4.xls). We determined the value of d and h by fitting Eq. (4) to the 13 cross-hybridization values. We constrained the minimum value of d to 1×10^{S7} because the authors of the HCQ study stated that the "coefficient d measures the probability of a perfect match target-probe pair remaining hybridized after the reaction", otherwise Solver would produce a value that was less than zero, which is non-physical. The resulting theoretical curve (black line in Fig. 2a) was highly correlated to the cross-hybridization values, with R^2=0.97. Visual inspection showed that the curve did not go through all of the experimental data because the probability of cross-hybridization curve approached zero. With regard to coefficient h, there is neither support in the cited literature, or in the HCQ study for the statement that: "coefficient h is the ratio of energy of binding and energy of breaking of the target-probe pair".

We obtained a slightly higher correlation by fitting the same data to a 5th order polynomial (R^2=0.98) (dashed line in Fig. 3). Note that the curve of the polynomial function raised the probability of cross-hybridization to ~5% for all targets, which is more in line with the experimental data than the curve generated by Eq. (4).

There are many possible factors that could contribute to the differences in the shapes of the curves (e.g., microarray platform, buffers, temperature, probe length). For example, the HCQ study used array probes ranging in length from 50 to 75 nt (average=52.6±4.6 nt), while our study used 25 nt probes. The thin dashed horizontal line in Fig. 3 depicts the 25% probability of cross-hybridization. For a long probe having 87% sequence similarity to a target, there is a 25% probability of

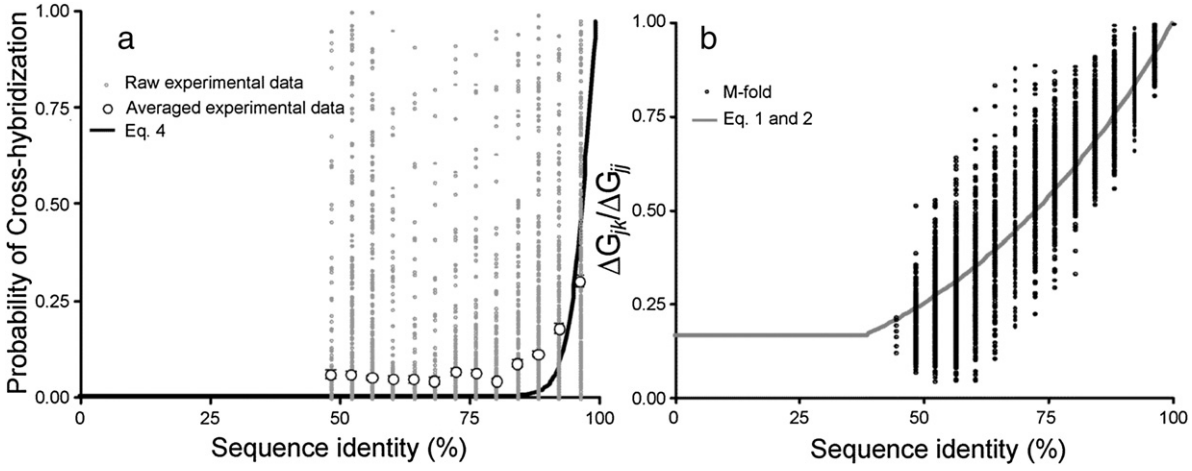

Fig. 2. Performance of the analytical cross-hybridization predictor based on our fingerprint data set. (a) Comparison of raw (grey dots) and average cross-hybridization values by sequence identity (white circles±standard error) to the theoretical curve predicted by the cross-hybridization predictor (Eq. (4) using =0.90, d=1×10^{S7}, h=10.72; n=13 averaged cross-hybridization values). Note that 380 PM/MM probes (~6% of total) were excluded because intensity of the MM probes exceeded those of the PM probe. (b) Comparison of G_{jk}/ G_{jj} ratios calculated by the analytical predictor (Eqs. (1) and (2)) as a function of sequence identity (n=6582). Target sequences were 23S rRNA sequences from six target species and 3160 probes were used. Grey line, derived from fitting the data (=0.66).

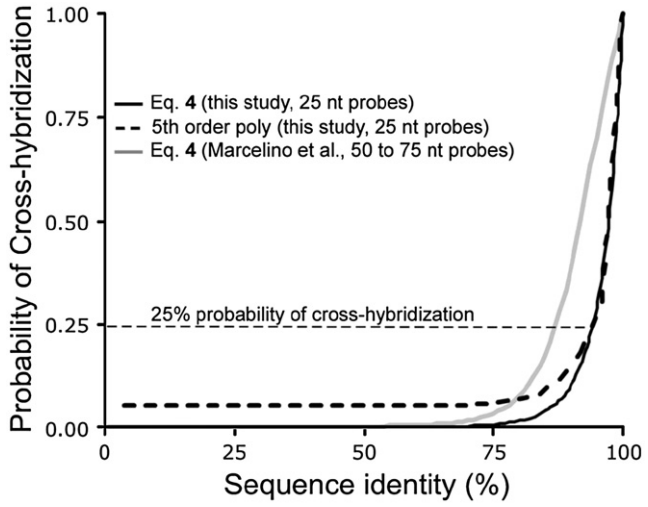

Fig. 3. Comparison of probability of cross-hybridization curves from: HCQ study (grey line; =1.26, d=0.75, h=8.47) based on 50 to 75 nt probes, this study (black line; =0.90, $d=1\times10^{-57}$, h=10.72) based on Eq. (4) and 25 nt probes, and this study (dashed black line) based on 5th order polynomial and 25 nt probes. Note that dashed line suggests at least ~5%probability of cross-hybridization for all probes regardless of sequence identity.

cross-hybridization for that probe and target. For a short probe, the target would require 95% sequence similarity to cross-hybridize to a target at the same probability level. These findings are consistent with the hypothesis that, in general, long probes are more affected by cross-hybridization than short probes, and that short probes have higher specificity to targets than long probes, which is coherent with research findings of Suzuki et al. (2007), Relogio et al. (2002), and Letowski et al. (2004). Nonetheless, more controlled experiments are needed to rigorously test this hypothesis.

3.3. Application of HCQ on two different Latin-square data sets

Given the values of parameters , d, and h for our data set, we assessed the utility of the predictor (i.e., $_{jk}$) and Eq. (5) to accurately quantify targets (i.e., $_{k}$) using data from two Latin-square designed experiments that were based on previously published data (Pozhitkov et al., 2008b). Latin-square design is $n\times n$ table filled with n different concentrations in such a way that each concentration occurs exactly once in each row and exactly once in each column. The first Latin-square experiment used a range of target concentrations from 0.0 to 0.08 nM, while the second experiment used a range of concentrations from 0.0 to 0.25 nM. The exact amount of each target dispensed into each mixture and experiment is available in Table 1 of Pozhitkov et al. (2008b).

For the first Latin-square data set, the predicted target concentrations poorly correlated to actual target concentrations, with R^2=0.33 (Fig. 4a) (see Sup5.doc and Sup6.xls). When we used the 5th order polynomial function as a proxy for $_{jk}$ in Eq. (5), the predicted concentrations yielded a poorer correlation between predicted and actual concentrations, with R^2=0.20 (Fig. 4b) (see Sup11.xls). Similarly, for the second Latin-square data set, predicted target concentrations were

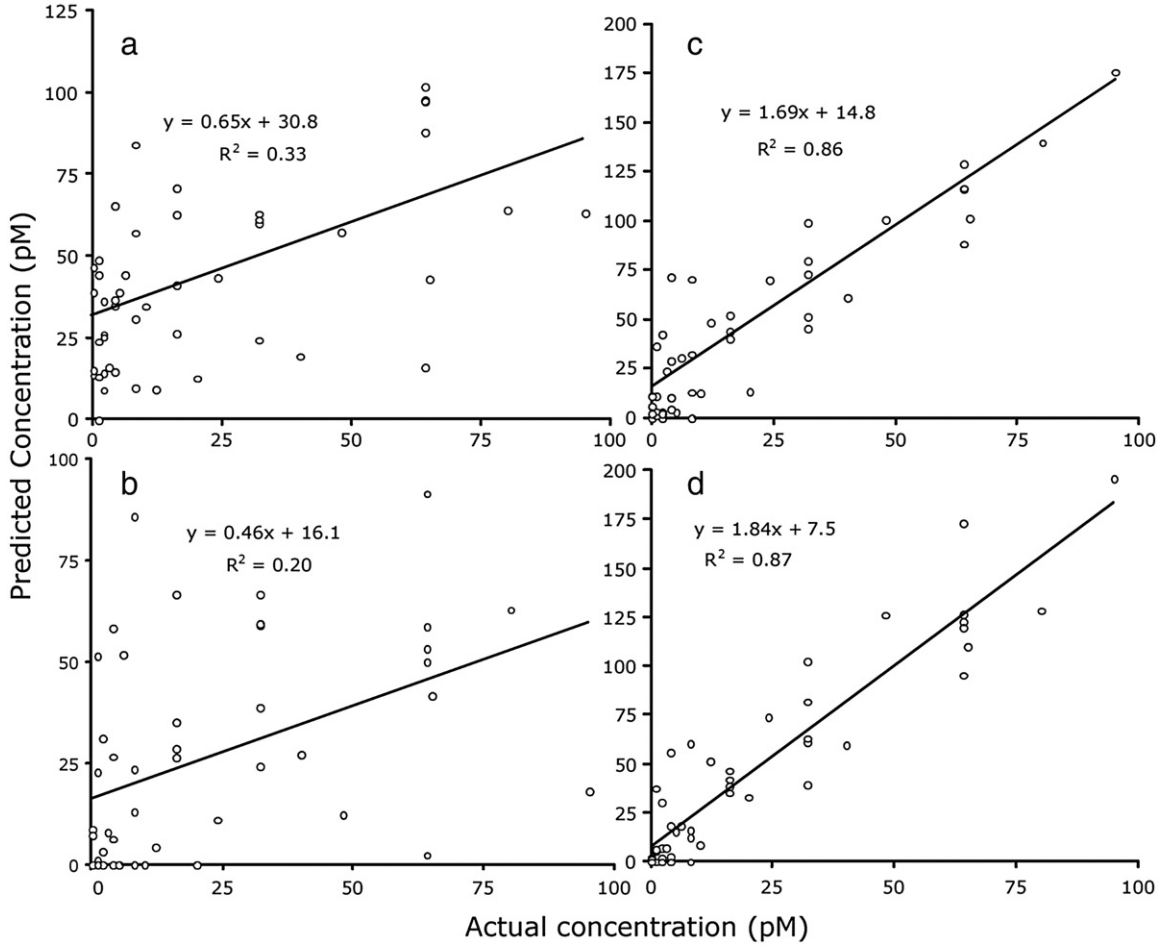

Fig. 4. Relationship between predicted and actual concentrations based on HCQ approach and the first Latin-square data set (6 targets×8 different mixtures). (a) Predicted values based on Eq. (5) and the following settings: =0.90, $d=1\times10^{-57}$, h=10.72. Unless otherwise specified, initial values for each of the six $_{k}$s was set to 48 pM, b=4861 a.u., v=576 a.u., and =0.0. (b) Predicted values based on Eq. (5) except that the 5th order polynomial function was a proxy for $_{jk}$. (c) Same as (a) except specific target binding, b, was accounted for by the fingerprint data. (d) Same as (b) except accounted for the fingerprint data in b (note that 5th order polynomial function was used).

moderately correlated to actual concentrations for the HCQ approach (R^2=0.66) (see Sup7.xls), but when the 5th order polynomial function was used as a proxy for $_{jk}$, the approach yielded a lower correlation (R^2=0.65) (figure not shown; see Sup12.xls).

Presumably, the less than satisfactory results obtained using the HCQ approach was due to the fact that averaging the experimentally-determined cross-hybridization values by sequence identity grossly underestimated the real variability in the microarray data, as shown by the grey dots in Fig. 2b. We initially hypothesized that the 5th order polynomial curve might improve the prediction of cross-hybridization because the curve was visually more in line with the experimental data than the curve produced by Eq. (5) (Fig. 3), and not dependent on variables , *d* and *h* and their associated inherent error. The results showed otherwise.

3.4. Attempts to improve the HCQ approach

In the HCQ study, the initial value of the "unequal specific response", *b*, was set to a constant for all probes because they found that signal intensities only varied by an average of 20% among different perfect match probes. In contrast, we found that the intensities of perfect match probe-target duplexes varied by N130% of the average intensity. Given the poor correlations between predicted and actual target concentrations, we attempted to improve the approach by forcing Eq. (5) to account for the different binding energies of specific (perfect match) probes, which might be another major source of variability in the approach. To account for this variability, we set the value of *b* for each perfect match probe-target duplex to the signal intensity of the probes in the fingerprint data set. This modification would allow Eq. (5) to account for differences in the binding energies among different probes.

For the first Latin-square experiment, setting the value of *b* to the perfect match intensity of the recorded fingerprint *did* improve the correlation between predicted and actual target concentrations (from R^2=0.33 to 0.86) using the modified Eq. (5). (Fig. 4c) (see Sup8.doc and Sup9.xls). Evidently, there was no significant improvement using the 5th order polynomial function (from R^2=0.86 to 0.87) (Fig. 4d) (see Sup13.xls). For the second Latin-square experiment, the results were inconclusive because setting the value of *b* to the perfect match intensity of the recorded fingerprint decreased the correlation between predicted and actual target concentrations (from R^2=0.66 to 0.59) (figure not shown; see Sup10.xls). Also, using the 5th order polynomial function and setting the value of *b* to the perfect match intensity did not improve the correlation between actual and predicted target concentrations (R^2=0.65 to 0.49; figure not shown; see Sup14.xls). Hence, setting *b* to the perfect match intensity of the recorded fingerprint did not yield any decisive results.

3.5. Reasons for the lack of correlation between actual and predicted target concentrations

To further resolve potential sources of variability, we examined the relationship between the Ln of the raw signal intensities and the predicted intensities (i.e., Y_{tot}) of the probes for each Latin-square mixture, using the actual concentrations of the targets in the mixtures, the background intensity for each mixture, and the signal intensities (~inferred binding energies) for perfect match probes. Y_{tot} refers to the predicted value of signal intensity cumulatively contributed to by all targets (see Sup15.xls to Sup18.xls for examples). Note that this experiment does not involve any optimization steps (i.e., Solver was not used). Because we provide all the known values for Eq. (5), the relationship between the Ln of the raw signal intensity and Y_{tot} should be highly correlated. It should be noted that the values for , *b* and *h* were included through the calculation of the cross-hybridization calculator in this experiment, and as a consequence, they might also contribute to the error in the results.

For both Latin-square data sets, we were not able to establish any linear relationship that was statistically significant for Ln signal intensities and the predicted probe intensities (i.e., Y_{tot}) (all R^2 b 0.25; *n*=16 mixtures) (Fig. 5a; see Sup15.xls and Sup16.xls). Due to the fact that neither optimization algorithm (BLUE or MS Solver) was used, this experiment ruled out the possibility that the failure to obtain significant correlations was because of the algorithm. Therefore, the most likely reason for lack of a linear relationship is the natural variability in the microarray data (see grey dots in Fig. 2a), which is not accounted for by the averaged cross-hybridization values (see white circles in Fig. 2a).

In all cases, accounting for the specific binding energies of perfect match probes using the fingerprint data (i.e., modified Eq. (5)), did improve in the relationship between Ln signal intensities and the predicted probe intensities (i.e., Y_{tot}) for each Latin-square mixture (Fig. 5b; see Sup17.xls and Sup18.xls). R^2 for the first and second Latin-square experiments ranged from 0.58 to 0.82 (*n*=8 mixtures), and 0.59 to 0.79 (*n*=8 mixtures), respectively. However, this modification was not sufficient for accurate quantification of targets in complex target

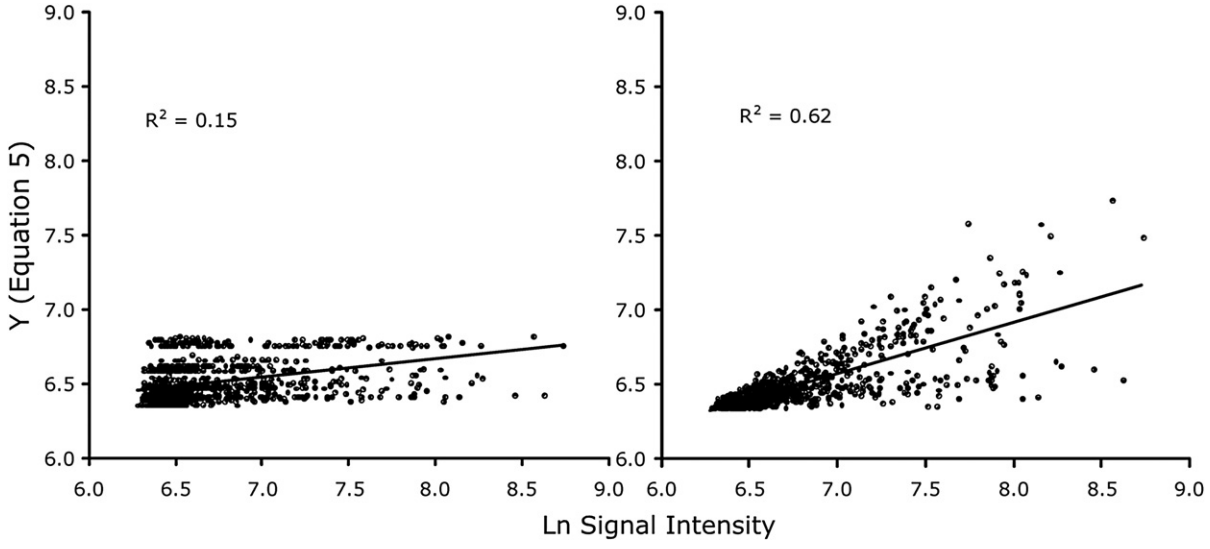

Fig. 5. Comparison of Y_{tot} from Eq. (5) and Ln signal intensity given the actual concentrations ($_k$) of each of the six targets in mixture 1 of the first Latin-square experiment (64, 32, 16, 8, 6, and 1 pM for targets 1 to 6, respectively), *b*=4861 a.u., *v*=576 a.u., and =0.0. (a) Control based on Eq. (5). (b) Based on Eq. (5) and the value of *b* was set to the signal intensity of perfect match probe-target duplexes obtained from the fingerprint data. Similar results were obtained for all Latin-square mixtures.

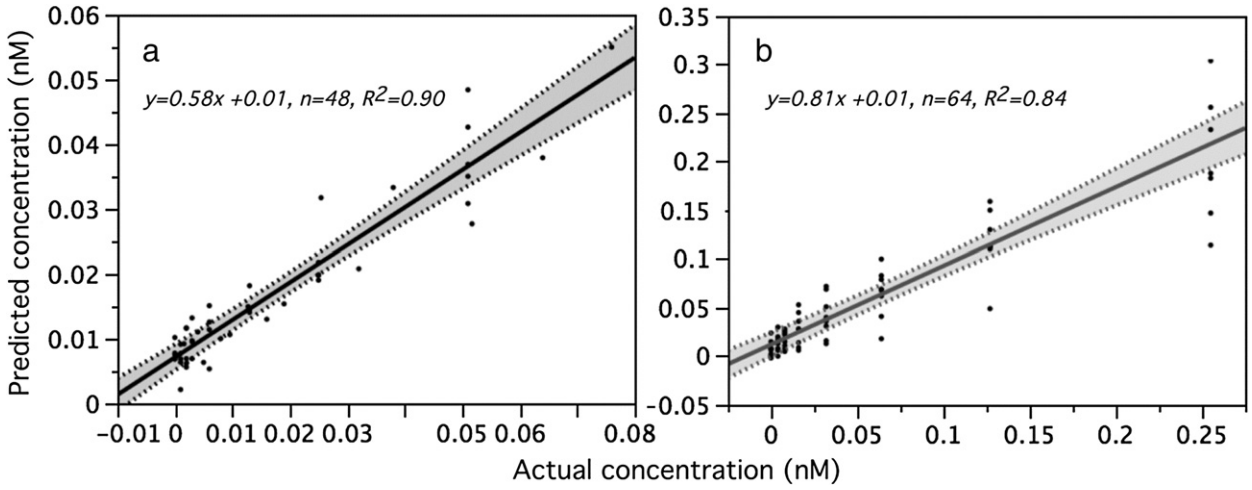

Fig. 6. Relationship between actual and predicted concentrations for the two independent Latin-square experiments using the fingerprint approach. (a) Results from the first Latin-square experiment (6 targets×8 mixtures). (b) Results from the second Latin-square experiment (8 targets×8 mixtures). The grey lines represent the 99% confidence limits of the regression line. Adapted from Pozhitkov et al. (2008).

mixtures. Clearly, further refinement of the HCQ approach is needed, which goes beyond the objectives of this study.

Given: (i) the problems with Eq. (6), (ii) the paucity of information on how was calculated by the authors of the HCQ study, (iii) the fact that the values of , *d* and *h* change under different conditions and sometimes produce non-physical results, (iv) that averaging the cross-hybridization by sequence identity grossly underestimated the real variability in the microarray data, and (iv) that none of our modifications to the approach adequately improved predictions of target concentrations, we are left with the realization that HCQ is insufficient for determining true target concentrations of complex target mixtures. Alternatively, essential information was left out of the description of HCQ approach making its replication not possible. For example, some variables (e.g., T_c, T′), equations, and settings were not explicitly stated or adequately explained in their article (as discussed in Section 3.1). In addition, the variable *b* was used for two different meanings. There was no other way to crosscheck our interpretation of their results because the HCQ study was not in compliance with MIAME standards (Brazma et al., 2001).

3.6. Previous analysis of the same Latin-square experiments

The Latin-square data sets used in this study have been previously used to evaluate the "fingerprint approach" (Pozhitkov et al., 2008b). This approach goes beyond the "single" optimal probe or probe set—single target relationship that is employed in other array approaches (e.g., the HCQ study). Rather, it uses hybridization patterns (fingerprints) to quantify specific nucleic acid targets in complex target mixtures (Pozhitkov et al., 2007c, 2008b). The entire set of probes and their corresponding signal intensities are viewed as a whole and considered to be a "fingerprint." The fingerprint of each target is recorded into a "library of fingerprints." A pattern of an unknown mixture is then numerically solved using matrix algebra, in terms of relative contributions (i.e., concentrations) of each pattern from the library of fingerprints. Our previous study showed that the regression between actual concentration of targets and those determined by the fingerprint approach were highly positively correlated with high R^2 values (e.g., R^2=0.90 for the first Latin-square experiment and R^2=0.84 for the second Latin-square experiment) (Fig. 6). These correlations would likely be further improved by: calibrating the array scanner used in that study so that microarray results are reported in terms of fluorophore density (which at the time was not physically possible) and increasing the probe replication (from duplicate to quadruplicate) using a higher-density array platform, which is part of our ongoing research.

4. Conclusion

Overall, the performance of the HCQ approach in this study suggests that it was not possible to extend quantitative microarray analysis to complex systems consisting of multiple organisms as alluded in the original publication (Marcelino et al., 2006). Although there are multiple causes that likely contribute to this conclusion (e.g., many aspects of the approach had to be determined to the best of our ability because they were not thoroughly explained in the original manuscript), the most reasonable explanation for the lack of success is that averaging the cross-hybridization by sequence identity does not adequately account for the natural variability in the microarray data (see grey dots in Fig. 2a). Since Marcelino et al. was not in compliance with MIAME standards (Brazma et al., 2001), we were not able to use the original data to reproduce their claims. Although our results suggest that the fingerprint approach (Pozhitkov et al., 2008b) was superior to that of HCQ, especially in terms of potential for quantitative microarray analysis of complex systems consisting of multiple microorganisms, further validation of the fingerprint approach is needed, which is part of our ongoing research.

Acknowledgements

We thank Thomas Beikler for his critical comments on the manuscript. This research was supported in part by a grant from the US National Oceanic and Atmospheric Administration (NAO3NOS4260216), and by Monika A. Leutenegger, the Royalty Research Fund, and the Provost Bridge Funding Program to P.A.N.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.mimet.2008.10.011.

References

Binder, H., Preibisch, S., 2005. Specific and nonspecific hybridization of oligonucleotide probes on microarrays. Biophys. J. 89, 337–352.
Bishop, J., Chagovetz, A.M., Blair, S., 2008. Kinetics of multiplex hybridization: mechanisms and implications. Biophys. J. 94, 1726–1734.
Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., et al., 2001. Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. Nature Genetics 29, 365–371.
DeSantis, T.Z., Stone, C.E., Murray, S.R., Moberg, J.P., Andersen, G.L., 2005. Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. FEMS Microbiol. Lett. 245, 271–278.

He, Z., Wu, L., Li, X., Fields, M.W., Zhou, J., 2005. Empirical establishment of oligonucleotide probe design criteria. Appl. Environ. Microbiol. 71, 3753–3760.
Letowski, J., Brousseau, R., Masson, L., 2004. Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. J. Microbiol. Methods 57, 269–278.
Li, X., He, Z., Zhou, J., 2005. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Res. 33, 6114–6123.
Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., Lockhart, D.J., 1999. High-density synthetic oligonucleotide arrays. Nat. Genet. 21, 20–24.
Marcelino, L.A., Backman, V., Donaldson, A., Steadman, C., Thompson, J.R., Preheim, S.P., Lien, C., Lim, E., Veneziano, D., Polz, M.F., 2006. Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. Proc. Natl. Acad. Sci. 103, 13629–13634.
Matveeva, O.V., Shabalina, S.A., Nemtsov, V.A., Tsodikov, A.D., Gesteland, R.F., Atkins, J.F., 2003. Thermodynamic calculations and statistical correlations for oligo-probes design. Nucleic Acids Res. 31, 4211–4217.
Mei, R., Hubbell, E., Bekiranov, S., Mittmann, M., Christians, F.C., Shen, M.M., Lu, G., Fang, J., Liu, W.M., Ryder, T., Kaplan, P., Kulp, D., Webster, T.A., 2003. Probe selection for high-density oligonucleotide arrays. Proc. Natl. Acad. Sci. 100, 11237–11242.
Pozhitkov, A., Stemshorn, K., Tautz, D., 2005a. An algorithm for the determination and quantification of components of nucleic acid mixtures based on single sequencing reactions. BMC Bioinformatics 6, 281.
Pozhitkov, A., Chernov, B., Yershov, G., Noble, P.A., 2005b. Evaluation of gel-pad oligo-nucleotide microarray technology using artificial neural networks. Appl. Environ. Microbiol. 71, 8663–8676.
Pozhitkov, A., Noble, P.A., Domazet-Loso, T., Staehler, P., Beier, M., Tautz, D., 2006. Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. Nucleic Acids Res. 34, e66.
Pozhitkov, A.E., Stedtfeld, R.D., Hashsham, S.A., Noble, P.A., 2007a. Revision of the nonequilibrium dissociation and stringent washing approaches for microbial identification studies using oligonucleotide DNA arrays. Nucleic Acids Res. 35, e70.
Pozhitkov, A., Tautz, D., Noble, P.A., 2007b. Oligonucleotide arrays: widely applied–poorly understood. Brief Funct. Genom. Prot. 6, 141–148.
Pozhitkov, A., Bailey, K.D., Noble, P.A., 2007c. Development of a statistically robust quantification method for microorganisms in mixtures using oligonucleotide microarrays. J. Microbiol. Methods 70, 292–300.
Pozhitkov, A.E., Rule, R.A., Stedtfeld, R.D., Hashsham, S.A., Noble, P.A., 2008a. Concentration-dependency of nonequilibrium thermal dissociation curves in complex target samples. J. Microbiol. Methods 74, 82–88.
Pozhitkov, A.E., Nies, G., Kleinhenz, B., Tautz, D., Noble, P.A., 2008b. Simultaneous quantification of multiple nucleic acids target in mixtures using high density microarrays and nonspecific hybridization as a source of information. J. Microbiol. Methods 75, 92–102.
Relogio, A., Schwager, C., Richter, A., Ansorge, W., Valcarcel, J., 2002. Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Res. 30, e51.
Rouillard, J.M., Herbert, C.J., Zuker, M., 2002. Oligoarray: genome-scale oligonucleotide design for microarrays. Bioinformatics 18, 486–487.
Suzuki, S., Ono, N., Furusawa, C., Kashiwagi, A., Yomo, T., 2007. Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays. BMC Genom. 8, 373.
Zhang, Y., Hammer, D.A., Graves, D.J., 2005. Competitive hybridization kinetics reveals unexpected behavior patterns. Biophys. J. 89, 2950–2959.