INTERNATIONAL JOURNAL OF
SYSTEMS AND SYNTHETIC BIOLOGY

1(2) December 2010, pp. 171-183

The Effect of Probe Length and GC% on Microarray Signal Intensity: Characterizing the Functional Relationship

Xuhua Xia[1,2]

[1]Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, Canada K1N 6N5

[2]Ottawa Institute of Systems Biology, 451 Smyth Road, Ottawa, Canada K1H 8M5

ABSTRACT

The quality of a microarray experiment is measured by sensitivity and specificity which depend on hybridization efficiency and non-specific cross-hybridization. The length and GC% of probe sequences are known to strongly affect hybridization and cross-hybridization. However, the joint effect of both the length and GC% of the probe sequences on microarray signal intensity has not been systematically assessed. Here I use a set of yeast microarray data with the GC% of probe sequences varying from 12.5% to 68.75% and with the probe length varying from 27 to 40nt to simultaneously assess both the effect of probe length and GC% on DNA hybridization. Both probe length and GC% have significant impact on signal intensity (SI) and a model derived from the data shows how changes in probe GC% can be compensated by the probe length and why such compensation did not work in some previous studies. SI increases sigmoidally with the probe GC% based on a data set where the probe length is constant. Our characterization of the effect of the probe length and GC% on SI suggests new ways to design microarrays and to normalize microarray data to reduce error variation.

Keywords: microarray, model fitting, probe length, probe GC%, DNA hybridization

INTRODUCTION

All living systems feature three major components: the genome, the transcripts and the proteins, as products of three essential biological processes: genome replication, transcription and translation. While the genome is essentially identical in all living cells of a multicellular organism, the transcripts and the proteins are dynamic features that change over time. To understand how living systems work, it is important to characterize the dynamic nature of transcripts and proteins as consequences of gene regulation. Microarray technology [1, 2] remains one of the most economic high-throughput methods for characterizing transcripts in living cells.

The quality of a microarray experiment is measured by sensitivity and specificity [3, 4]. Sensitivity is the fraction of probes with signal intensity (SI) above background when the

*Corresponding author: *E-mail: xxia@uottawa.ca*

target is present. Ideally, specificity should be 100%, i.e., SI is always above background when the target is present. Specificity is the fraction of probes that have background SI when there is no target. Ideally, specificity should also be 100%, i.e., all probes should have background SI when no target is present. How to maximize sensitivity and specificity in microarray experiment is a hot technical issue in microarray research.

SI is affected by the length [4-7], GC% [3, 8, 9] and concentration [4, 5] of probe and target sequences, and these factors tend to act synergistically rather than independently [4]. Microarray design mainly involves finding the best combination of the probe length, the probe GC% and probe concentration that would maximize sensitivity and specificity in hybridization with target samples. The effect of probe concentration is minimal compared to that of probe length and probe GC% [5]. I focus only on probe length and probe GC% in this paper.

That GC content can affect DNA hybridization on solid support has been recognized a long time ago [10], as the melting temperature increases with GC% with a slope of about 0.41. However, the effect of GC content of microarray probe sequences on signal intensity has been noted only recently [3, 8, 9]. The probe GC% has been demonstrated empirically to affect SI in microarray experiment [8]. The effect remains strong when factors such as nucleotide and dinucleotide identity have been controlled for [9].

A previous empirical study suggests that increased probe GC% tends to increase SI and consequently sensitivity, but may decrease specificity [3] due to cross hybridization. However, the functional relationship between GC% and SI has not been quantified. Also, it is not clear whether the effect of the probe GC% on SI may be confounded by the probe length which also affects SI.

The probe length is a dominant factor contributing to SI [4-7]. SI increased exponentially with the probe length, especially for lowly expressed genes [5]. Unfortunately, the GC% of the probe is not included in the analysis. The probes of 30mers have an average GC% slightly lower than the probes of 70mers in the study. In particular, three 30mers are particularly GC-poor, with GC% equal to 13.33% for two 30mers and 16.66 for one 30mer. In contrast, the probe with the lowest GC% in the 70mer group is 27.14%. So the relationship between SI and the probe length might be confounded by the discrepancy in the probe GC% between the two groups.

An early study [6] showed that increasing probe lengths decreased specificity. This is expected on theoretical ground. For example, suppose that a probe of length L_p is mixed with N random target sequences of length L_t and that hybridization between the probe and the target requires an exact match of L consecutive bases. Also assume that probes and targets have equal nucleotide frequencies. These conditions would lead to the expected number of the random targets that could hybridize with the probe being [11, pp. 4-10]

$$E = N(L_p - L + 1)\,(L_t - L + 1)\bullet 0.25^L \quad (1)$$

If $N = 10^9$, $L_p = 25$, $L_t = 1000$ and $L = 20$, then $E \approx 5$. However, if L_p is increased to 100, then $E \approx 72$. Note that E is the expected number of random targets that can hybridize to the

probe. In short, increasing L_p has two effects. First, increasing L_p will increase the noise by a linear rate of ($L_p - L + 1$). Second, a probe with a large L_p is a wider net for the target sequence and will consequently increase SI and sensitivity. A probe length of 60 appears to be a good compromise between sensitivity and specificity [12].

While major array manufacturers typically would make an effort to choose probe sequences with roughly the same GC%, there are cases where this approach cannot be taken, such as in the cases of high-density tiling arrays [13-15]. In addition, probe sequences in two-channel cDNA probe arrays designed for specific purposes [16], such as those targeting exon-intron junctions or exon-exon junctions [17-19], are often constrained by the target sequences and cannot be optimized to have similar GC%. It is therefore important to assess the effect of GC% and sequence length of probes on SI of microarrays.

Here I take advantage of a recently designed cDNA microarray for characterizing dynamics of intron splicing in the yeast, *Saccharomyces cerevisiae* [18, 19] to quantify the effect of probe GC% and probe length on DNA hybridization in microarray experiments. Three probes were designed for each intron-containing gene, one targeting the exon to characterize the total mRNA transcript, one targeting the intron to characterize the unspliced transcript and one targeting the exon-exon junction to characterize the spliced product. The probe sequences targeting the exon and intron are all 32 nt long and the program ArrayoligoSelector (*http://arrayoligosel.sourceforge.net*) was used to minimize the difference in GC% by choosing the target GC content of 35%. Because yeast intron sequences are both GC-poor and short relative to exon regions, it is practically difficult to design intron probes with GC content higher than 35%. So a target GC content of 35% was chosen and exon probes were chosen to have GC% matching that of intron probes. However, many probe sequences still vary widely in GC%, from 18.75% to 62.50%. This data set, including exon-targeting and intron targeting probes, will be referred to hereafter as Non-Junction Data Set. The data set will be used for characterizing the relationship between SI and probe GC%.

For probe sequences targeting the exon-exon junction, there is little freedom for minimizing the differences in GC% because the probe sequence is fixed by the 3'-end of the upstream exon and the 5'-end of the downstream exon. For this reason, the GC% of these probes vary widely from12.5% to 68.75%. This GC% range implies a melting temperature (T_m) difference of almost 13°C, given that T_m increases with GC% with a slope of 0.41 [10]. In order to optimize the probes so that the melting temperature will be roughly the same, probes with low GC% are designed to be longer than those with high GC% (Figure 1). Also, the binding energy for the 3'-end of the upstream exon and that for the 5'-end of the downstream exon were designed to be roughly the same [19]. This data set is to assess (1) whether lengthening the probe sequences is sufficient to offset the effect of GC% on DNA hybridization on microarrays and (2) whether the probe length and probe GC% affect signal intensity independently or synergistically (i.e., the effect of the probe length on signal intensity depends on the probe GC and vice versa). This data set will be referred hereafter as the Junction Data Set, and will be used to characterize the relationship between SI as dependent variable and the probe length and GC% as the two independent variables.

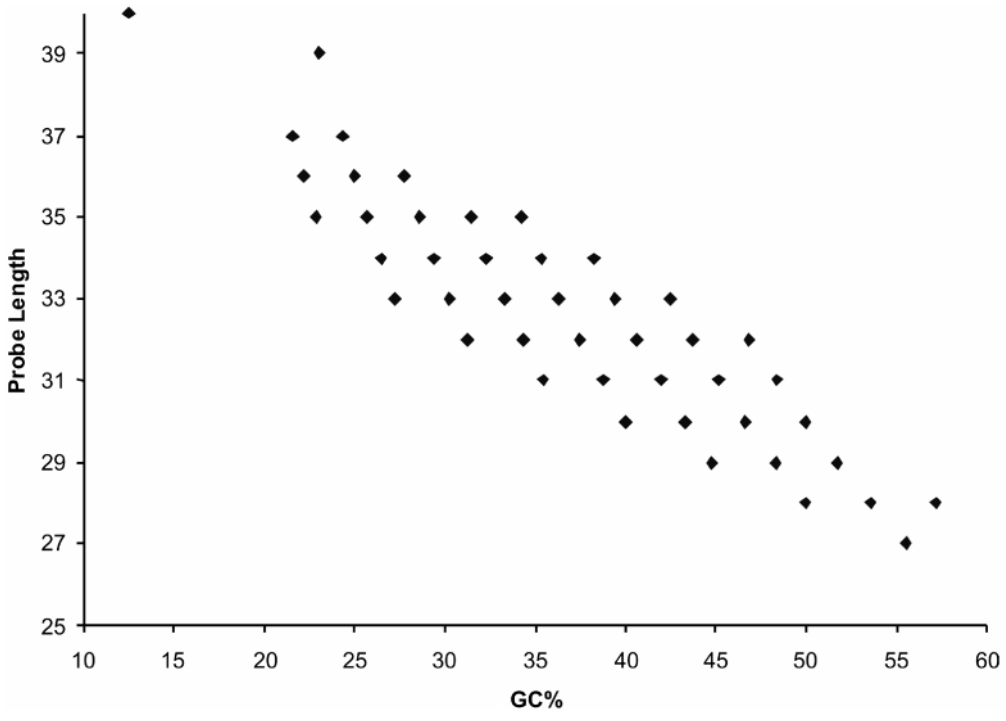

Figure 1: Relationship between GC% and Sequence Length for Microarray Probes Targeting the Exon-exon Junctions

MATERIALS AND METHODS

Microarray data for yeast introns (GEO record GSE7419, supplementary file GSE7419_RAW.tar of 259.8 Mb) were downloaded from *http://www.ncbi.nlm.nih.gov/geo* and uncompressed, resulting in one GenePix Array List (.GAL) file and 297 GenePix result (.GPR) files. Each microarray has a total of 16 blocks arranged in a 4´4 configuration. Each block contains 24×24 probe cells. So the total number of probe cells is 9216 (=4×4×24×24). The probes includes those for characterizing 232 yeast introns as well as others used for controls and normalization purposes [18, 19]. Three different probes were designed for each of the 232 yeast introns, with one targeting the exon to characterize the total mRNA transcript, one targeting the intron to characterize the unspliced transcript and one targeting the exon-exon junction to characterize the spliced product. The sequence length of probes targeting the exon-exon junctions differs with GC% (Figure 1). All other probes are 32 nt long. The probe sequences are in the platform GPL5052, retrieved at *http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL5052.*

Characterizing Relationship between Signal Intensity (SI) and GC%

In the original experiments [18, 19], the two-channel microarray data result from competitive hybridization between the wild type yeast and each of several mutants. Thus, SI values from both channels are indices of intron abundance, with the average correlation between the red and the green signal intensity being 0.92 for the 297 microarrays.

The median foreground and background is used throughout the analysis as they are more robust against microarray surface irregularities than mean intensities. The background is corrected as follows. First, a two-dimensional surface fitting of the background intensity is performed (Figure 2), using two-dimensional loess [20-22] fitting with the smoothing parameter equal to 0.01 and five times of iterative reweighting to minimize the effect of outliers to achieve a robust fit. The SAS procedure LOESS was used to perform the robust local regression. The fitted background values are then subtracted from the foreground signal intensity to generate the background-corrected signal intensity. Unless otherwise specified, signal intensity (SI) refers to background-corrected SI.

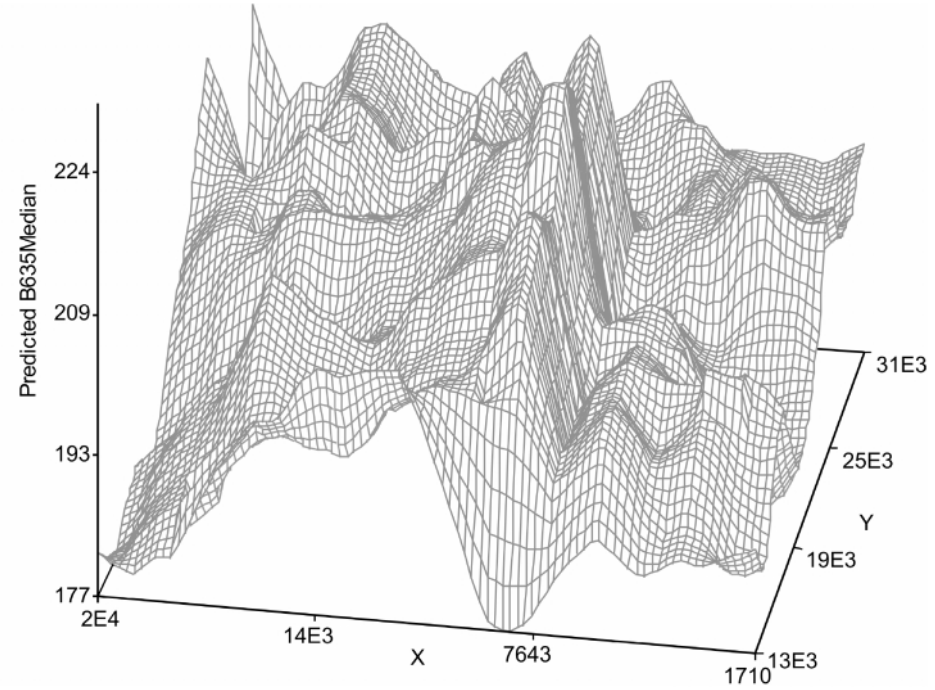

Figure 2: Two-dimensional Robust Local Fitting of Background Intensity from the Red Channel (Wave Length = 635 nm) of the Microarray Data file GSM179172.gpr, with the Smoothing Parameter Equal to 0.01 and the Number of Iterative Reweighting Equal to 5

Two separate data sets were used, one containing only microarray data for probes targeting exon-exon junctions (Junction Data set), and the other containing data for all the other probes (Non-junction Data set). The Non-junction Data are used to characterize the general relationship between SI and CG%, whereas the Junction Data are for characterizing the joint effect of the probe length and the probe GC%, as well as for checking whether the effect of GC% on SI is offset by the associated change in the probe length.

The probe sequences were read into DAMBE [23, 24] and GC% computed. DAMBE was also used to extract SI for intron-targeting, exon-targeting and exon-exon junction-targeting

probes. A nonlinear regression was fitted with the NLIN procedure in SAS [25] and further refined by the AMOEBA algorithm for optimization [26] implemented in DAMBE.

RESULTS

Signal Intensity (SI) Increases with Probe Length and GC%

For microarray data in the Junction Data Set (i.e., probes of variable lengths targeting the exon-exon junctions), probes with high GC% are shorter and probes with low GC% longer (Figure 1), with the purpose of all probes having roughly the same melting temperature. I used this data set to evaluate the joint effect of probe length and GC% on SI. The limitation of the data is discussed later.

Mean SI changes with both the probe length (L_p) and the probe GC% (Figure 3, which plots the mean SI for different combinations of L_p and GC%). Two interesting patterns are worth highlighting. First, SI increases with probe GC%, and those probe sequences with less than 30% of GC essentially do not hybridize with the targets. Second, SI may appear to decrease with probe length (L_p), i.e., those probes with L_p greater than 35 essentially do not hybridize whereas those short probes appear to hybridize well (Figure 3). However, this interpretation is wrong because L_p is correlated with probe GC%. For probes with GC% sufficiently high to hybridize, SI tend to increase with L_p. This is better characterized by modeling the effect of both L_p and probe GC% (GC) on SI by fitting the follow linear model:

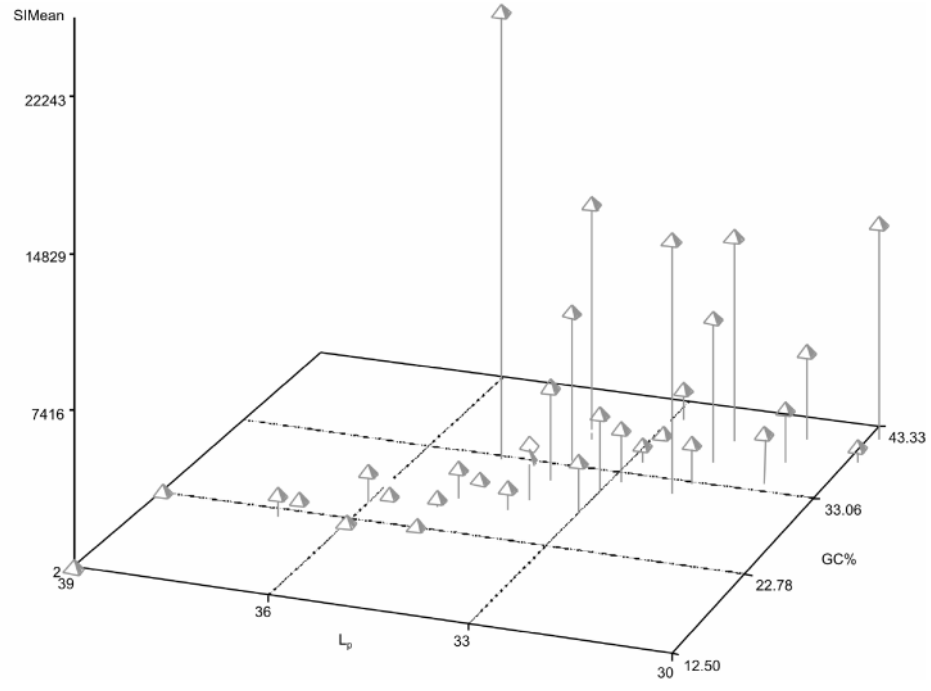

Figure 3: Visualizing the Effect of Probe Length (L_p) and GC% on Signal Intensity (SI)

$$sI = a_0 + a_1 Lp + a_2 GC + a_3 L_p \bullet GC \quad (2)$$

where the last term is for testing the interactions between L_p and GC, i.e., whether the effect of L_p and GC% on SI is multiplicative instead of additive. Note that, although the relationship between SI and the two independent variables (L_p and GC) may be nonlinear, the linear model should be sufficient for this particular data set because ΔGC for each given L_p and ΔL_p for each given GC% are both small (Figures 1 and 3). The model accounts for only 6.04% of the total variation in SI but is highly significant (F = 32.3195, numerator DF = 3, denominator DF = 1508, p < 0.00001). The slope is positive for both L_p and GC%, i.e., SI increases with both L_p and GC%. The interaction is not significant (p = 0.9173). The final fitted model without the interaction term is

$$SI = -48892.2 + 997.3053 L_p + 591.451 GC \quad (3)$$

The slopes for both L_p and GC% are highly significant, with p equal to 0.0037 for L_p and < 0.00001 for GC%. To derive an empirical formula for compensating the effect of decreasing probe GC on SI by increasing L_p, we keep SI constant and solve for L_p:

$$L_p = (0.001002701981\ SI + 49.02430580) - (0.5930490894\ GC = A - 0.5930490894\ GC$$

where A is a constant. Eq. implies that, for probe GC% to decrease from 55% to 15%, L_p needs to be increased by 24. In other words, if L_p = 27 for probe GC% = 55%, then L_p should be equal to 27 + 24 = 51 to compensate for the effect of probe GC% on SI. In the studies by Pleiss *et al.* [18, 19], L_p is lengthened by only about 12 nucleotides on average when the probe GC% changes from 55% to 15% (Figure 1). This compensation by L_p is insufficient and may have contributed to the lack of hybridization when probe GC% is lower than 30% (Figure 3).

One may argue that the lack of hybridization for probes with GC% smaller than 30% is not due to the low probe GC% but because of the lack of target sequences, i.e., those low-GC probes may happen to correspond to lowly expressed genes with few transcripts present in the sample. This can explain a few data points. For example, the yeast gene YCR097W has the probe GC% equal to 12.5%, and its SI is low not only for the probe targeting the exon-exon junction (average SI equal to 544.667) of this gene, but also low for the probes targeting the exon and the intron (average SI equal to 632.167 and 268.833, respectively, for the exon and intron). All these SI values are not significantly different from the background, suggesting that the gene is either not transcribed or lowly transcribed. The low SI for the gene is therefore better attributed to the lack of targets than to the low probe GC%. The low SI for gene YDR129C (with GC% = 21.6%) can be similarly explained without invoking low probe GC%.

This explanation of low SI as a consequence of little transcriptional activity, however, cannot explain the observed low average SI values for a number of other probes with a low GC%. For example, the exon-targeting probes for genes YOR182C (*RPS30B*), YDL061C (*RPS29B*) and YHR021C (*RPS27B*) have extremely high average SI values (55164.17, 59634.17, and 57423.17, respectively), i.e., they are highly expressed genes. This is not surprising because all three are ribosomal proteins in the small (40S) ribosomal subunit and known to be highly expressed. Furthermore, the intron-targeting probes for these two genes have low average SI values (351.67 for YOR182C, 970.17 for YDL061C, and 507 for

YHR021C), suggesting that these highly expressed mRNAs are efficiently spliced with few mRNA in unspliced form. This is again not surprising because highly transcribed ribosomal protein-coding genes have the necessity, and are known, to have high splicing efficiency. These observations jointly suggest that most of the mRNA species for these three highly expressed genes should be in the spliced form featuring the exon-exon junction, i.e., we should expect to see high SI values for the probes targeting the exon-exon junction for these three genes. Surprisingly, the SI values for the probes targeting the exon-exon junction for these three genes are very small, being 395.333 for YOR182C, 472.000 for YDL061C, and 233.67 for YHR021C which are hardly above the background. This strongly suggests the possibility that the low average SI values for the exon-exon junction of these three genes is due to poor hybridization caused by low GC% in the probe (21.6% for YOR182C, 22.2% for YDL061C and 23.1% for YHR021C). Given L_p = 27 for the probe GC% = 55%, we should have L_p = 47 for GC% = 21.6, L_p = 46 for GC% = 22.2%, and L_p = 46 for GC% = 23.1% according to Eq.7. However, the maximum L_p in the studies [18, 19] is only 39 nt which is insufficient to compensate for the low probe GC%.

With the significant effect of probe GC% and L_p on SI, it seems obvious that SI should be adjusted by CG% and L_p. The general approach should be to fit SI as a function of GC% and L_p, i.e., SI = F(GC, L_p) and then take the residual as the new SI. However, one should not use the linear model in Eq. because it is appropriate only with small variation in GC% and Lp, i.e., when a short segment of a curve can be sufficiently approximated by a straight line. Instead, one should use either the parametric 2-D model fitting or the local robust regression such as loess [20-22].

Non-linear Relationship between SI and Probe GC% with Fixed L_p

For microarray data in the Non-Junction Data Set (i.e., with probes targeting the exons and introns but excluding the probes of variable lengths targeting the exon-exon junctions), L_p is always 32 nucleotides, but probe GC% varies widely, from 18.75% to 62.50%. This data set is used here for characterizing the relationship between SI and probe GC%.

Mean SI, computed from probes with the same GC%, increases with GC% of probe sequences (Figure 4). The relationship is statistically highly significant ($p < 0.0001$), by fitting either a linear, a non-linear model to the data, or by a nonparametric Spearman correlation.

I consider two nonlinear models for fitting the sigmoid relationship in Figure 4. The first, designated NLIN4 (for a non-linear model with four parameters α, β, δ, and γ) in Eq. (5), is perhaps the most general one. I also consider a reduced non-linear model with three parameters (α, β, and δ), designated NLIN3 in Eq. (5). These two models are compared against the linear model with only two parameters (intercept and slope).

$$NLIN4: SI = \frac{\alpha}{\gamma + e^{-\beta(GC-\delta)}}$$
$$NLIN3: SI = \frac{\alpha}{1 + e^{-\beta(GC-\delta)}} \quad (5)$$

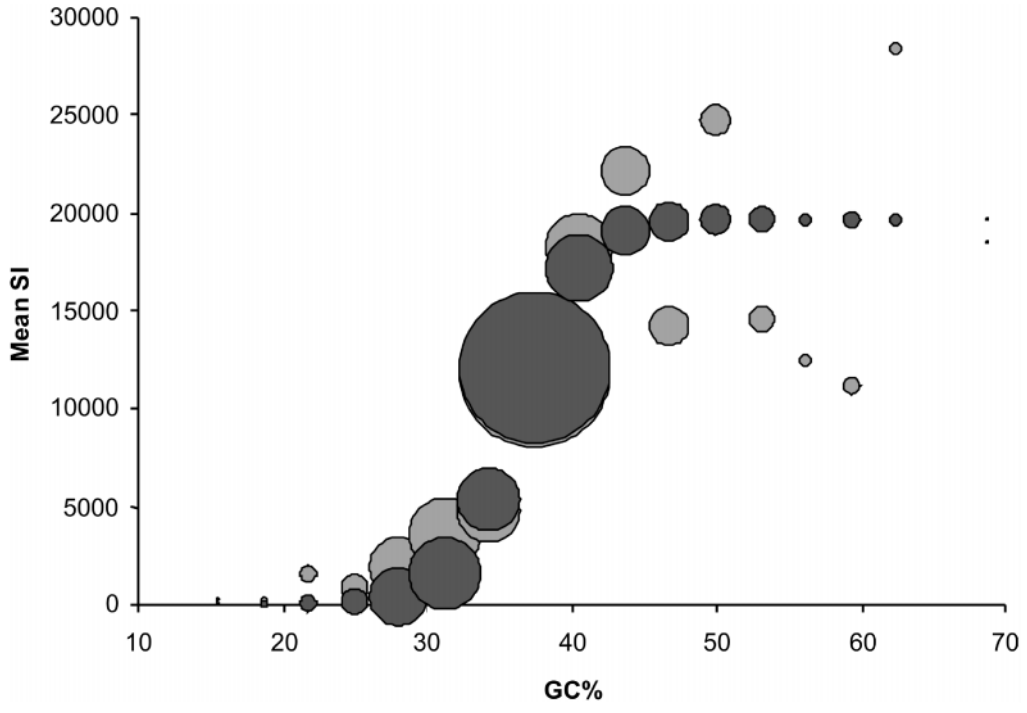

Figure 4: SI Increases with GC% for Microarray Data from Probes Targeting Exon-exon Junctions in the GSM179172.gpr File. The Blue and Red Bubbles are, Respectively, the Observed Data and the Expected Value. The Bubble Size Indicates the Number of Probes with the Same GC%

I used the model selection criteria AICc and AICu [27] to choose the best model. Due to the relationship between least-squares (LS) estimation and maximum likelihood (ML) theory [27, p. 110], we have

$$\ln[L(p,\sigma^2 \mid data)] = -\frac{n\ln(\sigma^2)}{2} = -\frac{n\ln\left(\frac{RSS}{n}\right)}{2} \tag{6}$$

where the left of the equation is the log-likelihood (hereafter referred to as lnL), σ^2 is the variance of the residuals, RSS is the residual sum of squares, n is the number of data points, and RSS/n is the likelihood estimate of σ^2. I will designate RSS_{linear} as RSS for the linear regression and RSS_{NLIN4} and RSS_{NLIN3} for the two non-linear models in Eq. (5). Fitting the 7704 pairs (SI, GC%) of data results in $RSS_{NLIN4} = RSS_{NLIN3} = 26996393152.7967$ (with $\alpha = 25887.0355$, $\gamma = 1.3170$, $\beta = 0.4689$ and $\delta = 37.1103$ for NLIN4 and $\alpha = 19655.9683$, $\beta = 0.4689$, and $\delta = 36.5231$ for NLIN3) and $RSS_{linear} = 89616295700$ from which we obtain $lnL_{NLIN4} = lnL_{NLIN3} = -58047.6144$ and $lnL_{linear} = -62669.3746$. The difference in lnL, i.e., the likelihood ratio, is 4621.76. We note that the linear model with two parameters (intercept and slope) and the nonlinear models as specified in Eq. are not nested so that a conventional likelihood ratio

test cannot be used. However, we can use information theoretic indices [27, 28] to choose which of the model is more appropriate for the data. Two information-theoretic indices that perform better than either the conventional AIC [29, 30] or BIC [31] are AICc and AICu defined as [32, pp. 22-32]:

$$AICc = \ln\left(\frac{RSS}{n}\right) + \frac{n+p}{n-p-2}$$
$$AICu = \ln\left(\frac{RSS}{n-p}\right) + \frac{n+p}{n-p-2} \quad (7)$$

where RSS is RSS_{linear} for the linear model and RSS_{NLIN4} and RSS_{NLIN3} for the nonlinear models NLIN4 and NLIN3, respectively, n is the number of observations (i.e., 7704) and p is the number of parameters (2 for the linear model, 4 for NLIN4 and 3 for NLIN3). The smaller the indices, the better the model is. The difference between AICc and AICu is the estimation of variance. AICc uses the maximum likelihood estimate RSS/n, whereas AICu uses the unbiased estimate RSS/(n – p). As n (= 7704) is far greater than p, the difference between AICc and AICu is minimal.

Both AICc and AICu (Table 1) identified NLIN3 as the best model. This is true not only just for the data in file GSM17972.gpr, but also for other 297 files in the set of microarray experiments. This suggests that the sigmoidal relationship is general and can be used to correct the effect of probe GC% in microarray experiments. I should add that the non-linear relationship characterized by the equation is purely descriptive and does not imply any physico-chemical process underlying the hybridization process.

Table 1
Computational Details for Information-theoretic Indices AICc and AICu

	Linear	*NLIN4*	*NLIN3*
$n^{(1)}$	7704	7704	7704
n_p (2)	2	4	3
RSS(3)	89616295700	26996393153	26996393153
RSS/n	11632437.137	3504204.719	3504204.719
RSS/(n-n_p)	11635457.764	3506025.085	3505569.816
(n+n_p)/(n-n_p-2)	1.001	1.001	1.001
AICc	17.270	16.071	16.071
AICu	17.270	16.071	16.071

(1) n – number of probes.
(2) n_p – number of parameters in the model.
(3) Residual sum of squares (squared deviations of observed SI from predicted SI).

DISCUSSION

Our characterization of the relationship between SI and the two independent variables (L_p and probe GC%) has several implications. First, for microarray experiments with limited variation in L_p and GC%, a linear relationship can be used either to find L_p to compensate for the effect

of varying GC% on SI or to correct the systematic bias introduced to SI by varying L_p and GC%. Second, for microarray data with a large variation in the probe GC%, a sigmoidal relationship should be fitted to eliminate the systematic bias introduced to SI by differences in the probe GC%. This nonlinear relationship, characterized by NLIN3 in Eq. can be used to guide the design of the correction protocol. Third, the traditional approach of equalizing free energy to compensate the effect of the difference in the probe GC% by changing the probe length appears not a good approach, because the probes should be lengthened substantially more to compensate for the GC% effect. These results are particularly relevant to high-density tiling arrays [13-15] where one has little freedom in optimize probe GC%.

Our results help to understand the lack of hybridization for probes with GC% lower than 30% in previous studies [18, 19]. This is significant for microarray experiments on GC-poor genomes. For example, the genome of *Mycoplasma genitalium* and *M. pulmonis* have genomic GC% lower than 30% [33]. For such genomes, the 25mers in a conventional Affymetrix probe array will almost certainly fail to achieve consistent hybridization.

In presenting the results, I have assumed that cross hybridization is minimal. A previous study suggests that such an assumption is reasonable when probe GC% is not higher than 55% [3]. There are a few probes with GC% higher than 55% in the data used in this paper, but they are highly unique and have little sequence similarity to other sequences. Another assumption is the linear relationship in Eq. (3) which may not have a solid theoretical basis. However, because ΔGC% is small for any given L_p, the linear characterization in Eq. (3) is sufficient for the data, and adding any non-linear terms does not improve the fit when judged either by adjusted R^2, Akaike information content or likelihood ratio tests for model selection [27].

There are three shortcomings in this current study, both being associated with the limitation of the data. First, the original microarray experiment was not designed specifically for evaluating the effect of the probe length and the probe GC% on SI. For this reason we do not have long probes with high GC% or short probes with low GC%. This precluded the simultaneous characterization of potential nonlinear effect of both the probe length and the probe GC% on SI. Also, a two-channel microarray is typically designed without aiming for identical amount of probes in each microarray spot. This implies substantial inherent noise in the data.

The second limitation of the paper is that it does not consider the position of G and C in the probe. The melting and annealing of DNA depend not only on GC%, but also on the distribution of G and C along the sequence. Suppose that we have two DNA sequences (designated DNA1 and DNA2, respectively) of the same length and the same nucleotide composition, but DNA1 has GC base pairs located at the two terminals whereas DNA2 has GC base pairs located in the middle of the sequence. Because melting occurs much more likely at the two terminals than in the middle [34], especially when the terminals are AT-rich , we should expect DNA1 to be more stable than DNA2. However, the data sets I analyzed contain few probe sequences with such extreme nucleotide distributions, and a specifically designed experiment would be needed to evaluate the effect of the spatial distribution of GC base pairs on DNA hybridization.

In short, DNA hybridization is strongly affected by the probe length and the probe GC%, and the relationship between the signal intensity and the probe GC% can be well characterized

by a three-parameter sigmoidal function. Lengthening the probe sequence can at least partially compensate the effect of low GC% on hybridization, but such compensation may not be achieved by equaling probe melting temperature or free energy in binding.

ACKNOWLEDGEMENTS

This study is supported by NSERC's Discovery, and Strategic Grants. I thank J. A. Pleiss and G. Whitworth for their lengthy and patient explanation to me on the design of their microarray chip, the details of their experiment, and the interpretation of their microarray data. E. Prankeviciene provided helpful comments, and two anonymous referees checked for errors and provided helpful suggestions.

REFERENCES

[1] M. Schena, "Genome Analysis with Gene Expression Microarrays," *Bioessays*, 18, 427-31, 1996.

[2] M. Schena, *Microarray Analysis*. New York: Wiley-Liss, 2003.

[3] K. Kucho, H. Yoneda, M. Harada, and M. Ishiura, "Determinants of Sensitivity and Specificity in Spotted DNA Microarrays with Unmodified Oligonucleotides," *Genes Genet Syst*, 79, 189-97, 2004.

[4] A. Jayaraman, C. K. Hall, and J. Genzer, "Computer Simulation Study of Probe-target Hybridization in Model DNA Microarrays: Effect of Probe Surface Density and Target Concentration," *J. Chem. Phys*., 127, 144-912, 2007.

[5] L. Ramdas, D. Cogdell, J. Jia, E. Taylor, V. Dunmire, L. Hu, S. Hamilton, and W. Zhang, "Improving Signal Intensities for Genes with Low-expression on Oligonucleotide Microarrays," *BMC Genomics*, 5, 35, 2004.

[6] A. Relogio, C. Schwager, A. Richter, W. Ansorge, and J. Valcarcel, "Optimization of Oligonucleotide-based DNA Microarrays," *Nucl. Acids Res.*, 30, e51- 2002.

[7] C. C. Chou, C. H. Chen, T. T. Lee, and K. Peck, "Optimization of Probe Length and the Number of Probes per gene for Optimal Microarray Analysis of Gene Expression," *Nucl. Acids Res.*, 32, e99- 2004.

[8] H. C. Yang, Y. J. Liang, M. C. Huang, L. H. Li, C. H. Lin, J. Y. Wu, Y. T. Chen, and C. S. J. Fann, "A Genome-wide Study of Preferential Amplification/hybridization in Microarray-based Pooled DNA Experiments," *Nucl. Acids Res.*, gkl446 2006.

[9] Y. Chen, C. C. Chou, X. Lu, E. Slate, K. Peck, W. Xu, E. Voit, and J. Almeida, "A Multivariate Prediction Model for Microarray Cross-hybridization," *BMC Bioinformatics*, 7, 101 2006.

[10] J. Meinkoth and G. Wahl, "Hybridization of Nucleic Acids Immobilized on Solid Supports," *Anal Biochem*, 138, 267-84 1984.

[11] X. Xia, *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics.* New York: Springer US, 2007.

[12] D. L. Leiske, A. Karimpour-Fard, P. S. Hume, B. D. Fairbanks, and R. T. Gill, "A Comparison of Alternative 60-mer probe Designs in an in-situ Synthesized Oligonucleotide Microarray," *BMC Genomics*, 7, 72 2006.

[13] J. Yazaki, B. D. Gregory, and J. R. Ecker, "Mapping the Genome Landscape using Tiling Array Technology," *Curr Opin Plant Biol*., 10, 534-42 2007.

[14] X. S. Liu, "Getting Started in Tiling Microarray Analysis," *PLoS Comput Biol*., 3, 1842-4 2007.

[15] T. C. Mockler, S. Chan, A. Sundaresan, H. Chen, S. E. Jacobsen, and J. R. Ecker, "Applications of DNA Tiling Arrays for Whole-genome Analysis," *Genomics*, 85, 1-15 2005.

[16] K. Srinivasan, L. Shiue, J. D. Hayes, R. Centers, S. Fitzwater, R. Loewen, L. R. Edmondson, J. Bryant, M. Smith, C. Rommelfanger, V. Welch, T. A. Clark, C. W. Sugnet, K. J. Howe, Y. Mandel-Gutfreund, and J. M. Ares, "Detection and Measurement of Alternative Splicing using Splicing-sensitive Microarrays," *Methods*, 37, 345 2005.

[17] T. A. Clark, C. W. Sugnet, and M. Ares, Jr., "Genomewide Analysis of mRNA Processing in Yeast using Splicing-specific Microarrays," *Science*, 296, 907-10 2002.

[18] J. A. Pleiss, G. B. Whitworth, M. Bergkessel, and C. Guthrie, "Rapid, Transcript-specific Changes in Splicing in Response to Environmental Stress," *Mol Cell*, 27, 928-37 2007.

[19] J. A. Pleiss, G. B. Whitworth, M. Bergkessel, and C. Guthrie, "Transcript Specificity in Yeast pre-mRNA Splicing Revealed by Mutations in Core Spliceosomal Components," *PLoS Biol*, 5, e90 2007.

[20] W. S. Cleveland and S. J. Devlin, "Locally-Weighted Fitting: An Approach to Fitting Analysis by Local Fitting.," *Journal of the American Statistical Association*, 83, 596-610 1988.

[21] W. S. Cleveland and E. Grosse, "Computational Methods for Local Fitting.," *Statistics and Computing*, 1, 47-62 1991.

[22] C. Loader, *Local Regression and Likelihood*. New York: Springer, 1999.

[23] X. Xia and Z. Xie, "DAMBE: Software Package for Data Analysis in Molecular Biology and Evolution," *Journal of Heredity*, 92, 371-373 2001.

[24] X. Xia, *Data Analysis in Molecular Biology and Evolution.* Boston: Kluwer Academic Publishers, 2001.

[25] SAS Institute Inc., *SAS/STAT User's guide. Version 6, Volume 2.*, Vol. 2, 4th ed. Cary, NC: SAS Institute Inc., 1989.

[26] W. H. Press, S. A. Teukolsky, W. T. Tetterling, and B. P. Flannery, *Numerical Recipes in C: the Art of Scientifi Computing.*, 2nd ed. Cambridge: Cambridge University Press, 1992.

[27] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer, 2002.

[28] X. Xia, "Information-theoretic Indices and an Approximate Significance Test for Testing the Molecular Clock Hypothesis with Genetic Distances," *Molecular Phylogenetics and Evolution*, 52, 665-676 2009.

[29] H. Akaike, "Information Theory and an Extension of Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki, Eds. Budapest: Akademiai Kiado, 1973, pp. 267-281.

[30] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE. Trans. Autom Contr. AC*, 19, 716-723 1974.

[31] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464 1978.

[32] A. D. R. McQuarrie and C. L. Tsai, *Regression and Time Series Model Selection.*: World Scientific, 1998.

[33] X. Xia, "DNA Methylation and Mycoplasma Genomes," *Journal of Molecular Evolution*, 57, S21-S28 2003.

[34] K. Y. Wong and B. M. Pettitt, "The Pathway of Oligomeric DNA Melting Investigated by Molecular Dynamics Simulations," *Biophysical Journal*, 95, 5618 2008.