Published online 17 March 2010 Nucleic Acids Research, 2010, Vol. 38, No. 11 e121
doi:10.1093/nar/gkq039

Evaluating oligonucleotide properties for DNA microarray probe design

Xiao-Qin Xia[1,*], Zhenyu Jia[2], Steffen Porwollik[3], Fred Long[3], Claudia Hoemme[4], Kai Ye[5], Carsten Müller-Tidow[4], Michael McClelland[2,3,*] and Yipeng Wang[2,3,*]

[1]Lechner-Haag Genomics Core, Vaccine Research Institute of San Diego, 10835 Road to the Cure, Suite 150, San Diego, CA 92121, [2]Department of Pathology and Laboratory Medicine, University of California, Irvine, CA 92697, [3]Department of Cancer Genetics, Vaccine Research Institute of San Diego, 10835 Road to the Cure, Suite 150, San Diego, CA 92121, USA, [4]Department of Medicine A, Hematology and Oncology and Interdisciplinary Center for Clinical Research (IZKF), University of Münster, Domagkstr. 3, 48129 Münster, Germany and [5]Molecular Epidemiology, Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands

Received March 19, 2009; Revised December 18, 2009; Accepted January 17, 2010

ABSTRACT

Most current microarray oligonucleotide probe design strategies are based on probe design factors (PDFs), which include probe hybridization free energy (PHFE), probe minimum folding energy (PMFE), dimer score, hairpin score, homology score and complexity score. The impact of these PDFs on probe performance was evaluated using four sets of microarray comparative genome hybridization (aCGH) data, which included two array manufacturing methods and the genomes of two species. Since most of the hybridizing DNA is equimolar in CGH data, such data are ideal for testing the general hybridization properties of almost all candidate oligonucleotides. In all our data sets, PDFs related to probe secondary structure (PMFE, hairpin score and dimer score) are the most significant factors linearly correlated with probe hybridization intensities. PHFE, homology and complexity score are correlating significantly with probe specificities, but in a non-linear fashion. We developed a new PDF, pseudo probe binding energy (PPBE), by iteratively fitting dinucleotide positional weights and dinucleotide stacking energies until the average residue sum of squares for the model was minimized. PPBE showed a better correlation with probe sensitivity and a better specificity than all other PDFs, although training data are required to construct a PPBE model prior to designing new oligonucleotide probes. The physical properties that are measured by PPBE are as yet unknown but include a platform-dependent component. A practical way to use these PDFs for probe design is to set cutoff thresholds to filter out bad quality probes. Programs and correlation parameters from this study are freely available to facilitate the design of DNA microarray oligonucleotide probes.

INTRODUCTION

Microarray technology surveys many thousands of genes to investigate gene expression (1), transcription factor binding profiles (2–5), DNA methylation profiles (4–6), DNA copy numbers (5) and genomic sequences (7).

Oligonucleotide probes provide higher hybridization specificity than longer PCR products (8–10). Falling costs of oligonucleotide synthesis, along with the development of new microarray manufacturing technologies, such as the NimbleGen maskless array synthesizer (11) and Agilent's ink-jet oligonucleotide synthesizer, make custom long (>50 bases) oligonucleotide arrays possible for many experimental applications. Optimal probe design algorithms are consequently desirable.

Hybridization on an array is characterized by several interconnected processes, including the affinity of a target for a probe, formation of stem–loop structures of a probe, formation of secondary structures (loops and helices) of a target, and probe-to-probe dimerization

*To whom correspondence should be addressed. Tel: +1 858 581 3960; Fax: +1 858 581 3970; Email: xqxia70@gmail.com
Correspondence may also be addressed to Michael McClelland. Tel: +1 858 581 3960; Fax: +1 858 581 3970;
Email: mcclelland.michael@gmail.com.
Correspondence may also be addressed to Yipeng Wang. Tel: +1 858 581 3960; Fax: +1 858 581 3970; Email: yipengw@gmail.com.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first Authors.

(12–16). There are a variety of factors governing these processes, including probe hybridization free energy (PHFE) (17), probe minimum folding energy (PMFE) (18), probe dimer and hairpin scores (19), as well as homology and complexity scores (20). Most of the current oligonucleotide probe design software packages estimate these properties (20–28).

To systematically and quantitatively study how these factors influence probe performance in microarrays, we collected a large amount of array CGH data and used these data to evaluate the utility of each PDF for probe selection. Using aCGH data, a novel PDF, pseudo probe binding energy (PPBE), was developed. PPBE is more accurate in predicting probe performance than all other factors and can thus be used for iterative improvement of the choice of oligonucleotides on the array. While the specific physical properties measured by PPBE remain unknown, they encompass platform-specific parameters.

METHODS

Microarray CGH data sets

Four comparative genome hybridization microarray data sets were used in the study (Table 1). Human genomic DNA (Data sets 1, 2 and 4) and *Salmonella* genomic DNA (Data set 3) samples were hybridized to their corresponding arrays. The array platforms include NimbleGen arrays (3′ end of the oligo is linked to the solid phase) and in-house spotted oligonucleotide arrays (5′ end of oligos is linked to the solid phase). The majority of probes on the arrays we used are 50 bases in length. However, there are also probes of different lengths, e.g. there are 9989 46-mer probes and 4721 55-mer probes on the array for data set 4. We found that the correlations of PDFs to probe sensitivities for these probes were very similar to those for the 50-mer probes (data not shown). In order to make data comparable across platforms, only data from 50-mer oligonucleotide probes were used. Hybridization intensity values were natural log transformed before fitting the linear models.

Data set 3 used pooled *Salmonella* genomic DNA *Xba*I restriction fragments, representing half of the genome in 3-fold excess, in one channel, and whole genomic DNA in the other. Data set 4 contain 205 replicates of human lung tissue genomic DNA hybridizations which were used as control channel in two-color hybridizations experiments.

PDF. The following DNA microarray PDFs were included in this study.

PHFE. PHFE was calculated based on the dinucleotide stacking energies.

$$PHFE = \varepsilon_{\text{head}} + \sum_{k=1}^{n-1} \varepsilon(b_k,\, b_{k+1}) + \varepsilon_{\text{tail}}$$

where n is the oligonucleotide length, $\varepsilon(b_k,b_{k+1})$ is the k-th position dinucleotide stacking energy, and $\varepsilon_{\text{head}}$ and $\varepsilon_{\text{tail}}$ are the terminal nucleotide stacking energies. The salt concentrations for the calculations were set to 1 M Na^+, 0 M Mg^{++}, and the temperature was set to 40, 50 or 60°C for the computation of PHFE. The dinucleotide stacking energies are computed according to SantaLucia (17) and shown in Supplementary Table 1.

PPBE. For a probe sequence $(b_1,b_2,\ldots,b_n)$ with n bases, the PPBE model is parameterized by dinucleotide stacking energies ε and position-dependent weights ω, $PPBE = \varepsilon_{\text{head}} + \sum_{k=1}^{n} \omega_k \varepsilon(b_k,b_{k+1}) + \varepsilon_{\text{tail}}$. The position-dependent weight ω is first estimated by fitting the linear model, employing dinucleotide stacking energies (as used in the PHFE model) as initial values. Then, with the same linear model fitting scheme, the pseudo dinucleotide stacking energies ε are approximated by treating previously estimated weights as known quantities. Such a process of 'reciprocal' estimation was iteratively carried out three times, at which point the average residue sum of squares (ARSS) for the PPBE model reached its minimum or near-minimum (see also the 'Linear modeling' section below, and Figure 1A).

PMFE. PMFE is the minimum folding energy of a single strand DNA sequence and represents the stability of the secondary structure of a given sequence. PMFE

Table 1. Array CGH data set used in this study

Data Set	Microarray platform	Sample	Manufacturer	Designer	Oligos	Bases	Role of data set in the analysis	Number of samples
1	NimbleGen HG18 whole genome CGH Array	Normal human male genomic DNA	NimbleGen Inc.	NimbleGen Inc.	137 280	50	Sensitivity	6
2	NimbleGen Human Promoter Array (custom design)	Human prostate cell line (PC3M, 267B1) genomic DNA	NimbleGen Inc.	authors	220 475	50	Sensitivity	4
3	NimbleGen *Salmonell*a Whole Genome Array (custom design)	*Salmonella* LT2 genomic DNA	NimbleGen Inc.	authors	288 238	50	Sensitivity, specificity	4
4	In-house Spotted Human Promoter Array (custom design)	Normal human lung tissue genomic DNA	authors	authors	11 653	50	Sensitivity, reproducibility	205

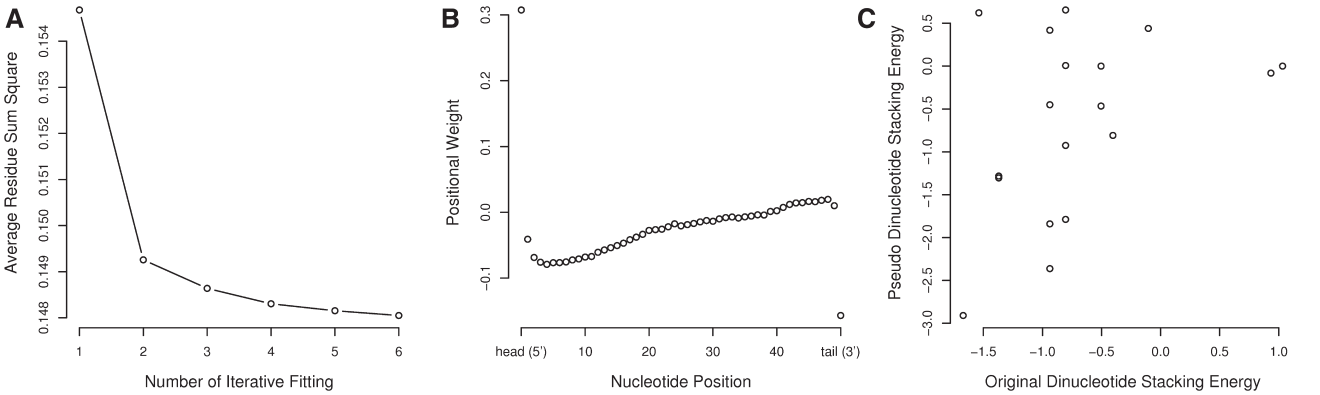

Figure 1. ARSS, positional weights and pseudo stacking energies of the PPBE model for data set 1. (**A**) Convergence of the PPBE model after three cycles of iterative fitting of both of positional weights and pseudo dinucleotide stacking energies (six cycles total); (**B**) Plot of positional weights; (**C**) Comparison of traditional dinucleotide stacking energies and pseudo dinucleotide stacking energies.

were computed by using the MFOLD program (18). The program *hybrid-ss-min* was downloaded from http://www.bioinfo.rpi.edu/applications/hybrid/download.php and executed on GNU/Linux. The parameters were set as DNA–DNA hybridization, 1 M Na^+, 0 M Mg^{++}, and the temperature was set to 40, 50 or 60 for the calculation of PMFE.

Probe dimer score, hairpin score. The calculation of the probe dimer score and the hairpin score was described as part of the AutoDimer program based on a sliding algorithm (19). For screening probe dimers, two probe sequences are incrementally overlapped, and the presence or absence of base pairing is evaluated and tabulated. A dimer score value was then determined by combining the number of Watson–Crick base pairs (+1) with mismatches (−1).

Hairpin secondary structures were screened by using the probe sequence to check for the presence of 4 and 5 base loops. A minimum of a 2-base stem was deemed to be necessary in a hairpin structure. Hairpin scores were sums of matched base pairs (+1) in hairpin stems where mismatches are not permitted.

Homology score. The homology score for each oligonucleotide estimates the degree of cross-hybridization, and is based on a BLAST search of the input sequence against a species-specific database. The calculation of the homology score was similar to the one used in the OligoWiz program (20).

$$\text{Homology Score} = \frac{100 \times L - \sum_{i=1}^{L} \max(B_{1i}, \ldots, B_{mi})}{100 \times L}$$

where L is the length of the oligonucleotide, m is the number of Blast hits considered in position i of the oligonucleotide and $B = \{B_{1i}, \cdots, B_{mi}\}$ is the bit score in position i.

Oligonucleotides with 100% identity to any considered BLAST hit along the full length received a score of 0. Percentages of identity <70% or shorter than 15 bp were removed, resulting in perfect homology scores of 1 for those oligos.

Complexity score. Complexity scores were calculated for estimating the degree of common sequence fragments in a given oligonucleotide, as described in the OligoWiz program (20). The information content can be calculated by the following equation:

$$I(w) = \frac{n(w)}{nt}\left(\log_2 \frac{n(w) \times 4^{l(w)}}{nt}\right)$$

where $n(w)$ is the number of occurrences of a pattern in the genome, $l(w)$ represents the pattern length, and nt is the total number of patterns found in DNA sequences present in the target pool, for example, the whole genome in an array comparative genomic hybridization. The following equation was used to calculate the complexity score for each oligonucleotide probe:

$$\text{Complexity Score} = 1 - \text{norm}\left(\sum_{L-l(w)+1}^{i=1} I(w_i)\right)$$

where L is the length of the oligonucleotide, w_i is the pattern in position i and norm is a function that normalizes the summed information to a value between 1 and 0 by dividing them by the maximum value. A complexity score of 0 indicates an oligonucleotide with very low complexity. Pattern lengths of 2, 5, 8 and 11 bases were tested in this study.

Oligonucleotide specificity and reproducibility

Data set 3, with known expected oligonucleotide signal ratios (3-fold changes) between the two channels, was used for estimating oligonucleotide probe specificity. The observed ratios were $\log_2$ base transformed for further analysis. Coefficient of variation (CV) was used for estimating probe reproducibility.

Linear modeling and model validation

R language (http://www.r-project.org) was used for linear modeling (29–31). In the four microarray data sets, simple linear models were used to evaluate each individual PDF, and multivariate models were used to estimate all PDFs together.

The ARSS, which reflects the model fitness, was defined as $r = (\sum_{i=1}^{n} (g_i - g_i^*)^2)/(n)$, where g_i was the observed *ln*-transformed intensity for probe *i*, g_i^* was the predicted *ln*-transformed intensity for probe *i*, and *n* was the number of probes. For model selection, the stepAIC function in the MASS package (http://www.r-project.org) was used to reduce the full model to the optimal one. This Akaike information criterion (AIC) is a measure of the quality of the fit of an estimated statistical model and balances the complexity of an estimated model with the accuracy with which the model fits the data (32).

The models were validated in two ways: within one data set and across different data sets. In both cases, the leave-many-out cross-validation (33) was used. Within-dataset validation uses half of the data from one data set to train the models and the other half for testing of the models. Cross-dataset validation uses different data sets, which may vary in array platforms and sample species, for training and testing.

RESULTS

Microarray CGH data sets

Array CGH data is a valuable source for studying microarray oligonucleotide probe performance because it can be assumed that most of the probes in these experiments hybridize to approximately equimolar target amounts, resulting in relatively uniform hybridization signals. Four large aCGH data sets on different array platforms, with a total of 657,646 oligos of 50 bases in length and 219 samples, were used in this study to evaluate PDFs and to develop new algorithms (Table 1).

Correlation of individual PDFs with probe hybridization intensities

The models examined are all presented in the 'Methods' section and will not be repeated here. All 10 PDFs, i.e. PHFE, PMFE, hairpin score, probe dimer score, homology score, complexity score (2 bases), complexity score (5 bases), complexity score (8 bases), complexity score (11 bases) and PPBE, showed highly significant correlation with probe hybridization intensities, as shown in Figure 2 (data set 1) and Supplementary Figure 1 (data sets 2, 3 and 4). The correlation coefficients (r), ARSS, intercepts and slopes for these linear regression models are listed in Table 2 and Supplementary Table 2.

The ARSS values of linear models based on individual PDFs were compared, as shown in Figure 3. Among these factors, PPBE generated the lowest ARSS, suggesting that this factor is superior to the traditional factors in correlating with probe hybridization intensity. PPBE was modeled by iteratively fitting dinucleotide stacking energies and positional weights, with the conventional dinucleotide stacking energies as initial values. The ARSS values from the PPBE model tend to stabilize after three cycles of iterative fitting of both positional weights and pseudo dinucleotide stacking energies (Figure 1 and Supplementary Figure 2). The positional weights and pseudo dinucleotide stacking energies generated from the different data sets are entirely different, reflecting the empirical nature of the model. The positional weights and pseudo stacking energies for PPBE models from different data sets are listed in Supplementary Tables 3 and 4, and the positional weights illustrate the effect of the distance of the dinucleotide to the solid phase. The positional weights of data sets 2 and 4, for example, showed inverse correlation to the distance to the probe's 5′ end, which may be due to the fact that these platforms differed in the ends of oligos that were linked to the solid phase (5′ versus 3′).

In most data sets, the best individual traditional factors were PMFE, dimer score and hairpin score. All these three PDFs showed that less stable probe secondary structure positively correlates with probe hybridization intensity, suggesting that the formation of secondary structure can severely hinder the probe hybridization capabilities.

PHFE's linear correlation with probe hybridization intensity was less significant, suggesting that hybridization behavior on microarrays might be different from that in solution. Moreover, quadratic rather than linear relationships were observed for data sets 1 and 3, and the mode (the peak points shown in Figure 2A and Supplementary Figure 1 and 2A) varies among these two data sets, suggesting that hybridization conditions were not the same for the two data sets. We tried to use quadratic equations to fit the data sets 1 and 3, but the ARSS values generated from these models were bigger than those obtained using simple linear models (data not shown). This is probably due to the fact that the majority of PHFE data points is clustered within a very narrow range, where the relationship between PHFE and intensities may be better described by a linear equation. In future studies, once there are sufficiently large data sets with a higher PHFE data spread across a wider range of values, more advanced models can be applied to scrutinize the relationship between PHFE and hybridization intensities in a non-linear fashion.

Blast score and complexity scores (2, 5, 8 and 11 bases) correlated least significantly with the probe hybridization intensity among the PDFs tested. No obvious differences were observed among the scores obtained for 2, 5, 8 and 11 bases when correlating them with probe hybridization intensity (Table 2).

Among all four data sets, PPBE, PMFE, dimer score and hairpin score showed positive correlations with probe hybridization intensity, and are therefore the more reliable indicators of probe sensitivity. The other PDFs displayed inconsistencies in correlation for different data sets. For example, PHFE is positively correlated with probe intensity in data sets 2 and 3, but is negatively correlated with probe intensity in data sets 1 and 4. More complex models might be developed for blast score and complexity scores (2, 5, 8 and 11 bases), but that is beyond the scope of this article.

As shown in Supplementary Table 2, enormous variations were observed among individual data sets for the trend coefficients (e.g. intercept and slope), possibly due to differences in array manufacture, sample and array processing and other factors.

The values of PHFE and PMFE are dependent on parameters such as hybridization temperature and

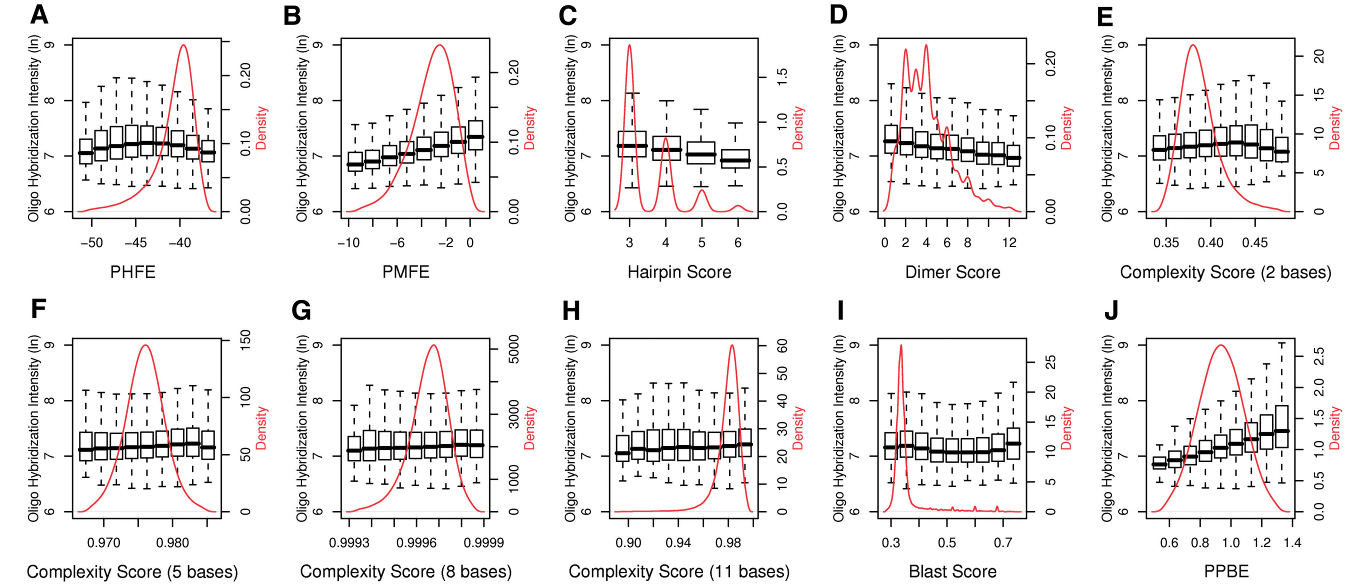

Figure 2. Box plots (black line) show the correlation of individual PDFs with observed oligonucleotide probe hybridization intensities for data set 1. The density curve (red line) is computed using kernel density estimates and shows the distribution of individual PDFs. The secondary *Y*-axis represent the density of different PDFs.

Table 2. Simple linear model average residue square sum (ARSS) and correlation coefficients (*r*) for the correlation of individual PDFs with probe hybridization intensities

	Data Set 1		Data Set 2		Data Set 3		Data Set 4	
	r	ARSS	*r*	ARSS	*r*	ARSS	*r*	ARSS
PHFE	0.11	0.168	0.03	0.504	0.03	0.460	0.13	1.668
PMFE	0.29	0.156	0.27	0.468	0.32	0.414	0.28	1.568
HairpinScore	0.21	0.162	0.22	0.479	0.20	0.442	0.21	1.621
DimerScore	0.19	0.164	0.23	0.478	0.17	0.448	0.15	1.660
ComplexityScore-2B	0.08	0.169	0.05	0.503	0.02	0.461	0.09	1.684
ComplexityScore-5B	0.04	0.170	0.11	0.498	0.01	0.461	0.02	1.698
ComplexityScore-8B	0.01	0.170	0.15	0.493	0.01	0.461	0.12	1.675
ComplexityScore-11B	0.01	0.170	0.10	0.498	0.02	0.461	0.10	1.683
BlastScore	0.02	0.170	0.11	0.498	0.01	0.461	0.18	1.641
PPBE	0.36	0.148	0.30	0.460	0.65	0.269	0.48	1.301

concentrations of sodium, most of which were unavailable to us. However, we computed PHFE and PMFE using various potential parameters, and changes in parameters did not cause significant differences in correlation assessments; the average difference of ARSS value are 0.0058 (0.010 for PHFE and 0.001 for PMFE) among different temperature settings. 60°C was used for the computation of PHFE, and 40°C was used for computation of PMFE for all data sets, because they slightly outperformed other temperatures.

Multivariate linear modeling

For each data set, a multivariate linear model with PPBE (W. PPBE model) was built based on all PDFs for predicting probe hybridization intensity and comparing the significance of the individual PDFs. This multivariate model showed significant improvement over all individual models based on each individual PDF (Figure 3, Supplementary Figure 3). The W. PPBE model parameters are shown in Supplementary Table 5.

Increasing the number of free parameters obviously improves the fit. On the other hand, overfitting is very likely to happen and reduces or destroys the ability of the model to generalize beyond the data it is built upon. The AIC is an operational way of trading off the complexity of an estimated model against how well the model fits the data (32). It not only rewards improvement of fit, but also includes a penalty that is an increasing function of the number of estimated parameters and thereby discourages overfitting. In this study, stepwise selection with AIC was used to search for the optimal model which only contains covariates (individual PDFs) related to the outcome (probe hybridization intensity). Stepwise model selection

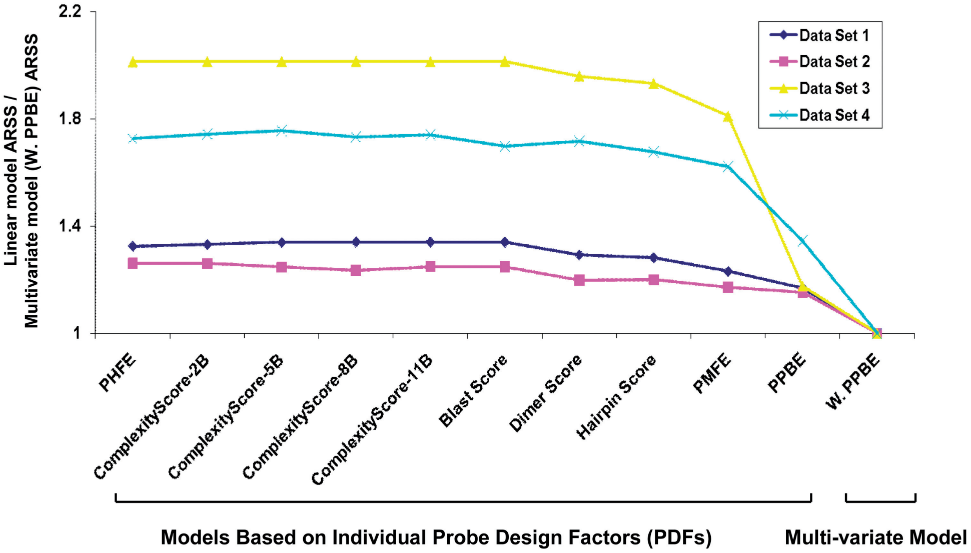

Figure 3. Relative ARSS of different models for different data sets.

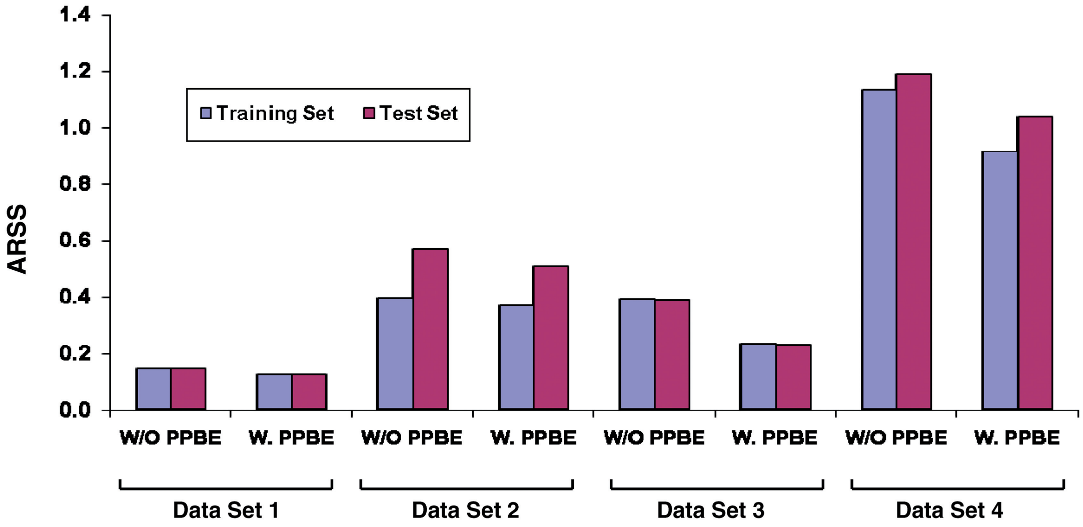

Figure 4. Comparisons of ARSS for within-dataset validations using the multivariate models W/O PPBE or W. PPBE.

analysis showed that all PDFs contributed to the prediction of probe hybridization intensity in all data sets with only one exception in which the complexity score (2 bases) was not significant in data set 1 (Supplementary Figure 4). The most significant factor is PPBE, followed by PMFE in all data sets. The order of significance of other PDFs varied among different data sets.

Generality of linear models

Two multivariate models, the W. PPBE model (includes all factors) and the W/O PPBE model (including all factors except PPBE), were developed using a training data set, and tested on independent data sets to determine if the models can be reliably used as a probe design tool.

Applying within-dataset validation, Figure 4 illustrates that the models developed from the training set can predict the performance of oligos in the test set almost as accurately as it can predict performance in the training set. The W. PPBE model outperformed the W/O PPBE in all cases, suggesting that PPBE is a reliable factor although it is generated by an empirical approach.

Cross-dataset validations (Supplementary Table 6) resulted in extremely high ARSS values in the test data sets when the W/O PPBE and W. PPBE models were applied, even when the array manufacture technique and sample species were identical between test and training set. The complex multivariate models developed from one data set can therefore not be directly and simply applied on other data sets. The adverse performance was not caused by PPBE, as there were no obvious differences between W/O PPBE and W. PPBE models. The substantial variations in correlation intercepts and slopes for each

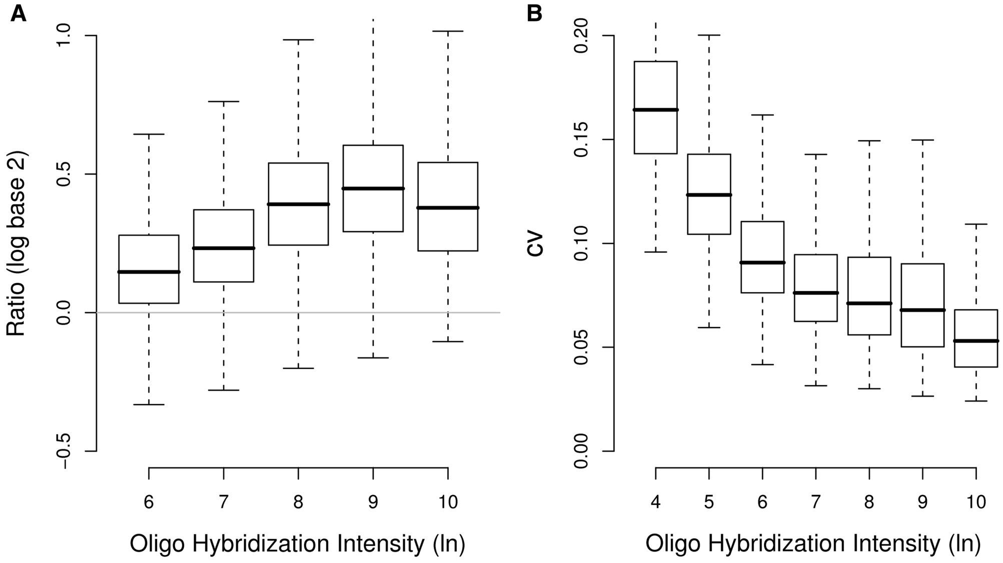

Figure 5. Correlation of probe hybridization intensity with probe specificity and reproducibility. (**A**) Correlation of probe hybridization intensity with probe specificity (observed $\log_2$ base transformed ratio) for data set 3. Gray line indicates no change; (**B**). Correlation of oligonucleotide probe hybridization intensity with probe reproducibility for data set 4, represented as coefficient of variation (CV).

individual PDF, as observed in Supplementary Table 2, severely hinder the cross-dataset probe intensity predictions using multivariate linear models.

Probe specificity

Probe specificity is a measurement of the capability of a probe to discriminate between its specific target sequences in a complex set of non-specific sequences. In a two-channel hybridization experiment, if one channel includes the target sequence and the other does not, then the probe with specificity for the target can be expected to yield a high ratio of hybridization signal intensity between the two channels, which is a measure of probe specificity in the mixture.

We estimated the oligonucleotide specificity using data set 3, where the targets in one channel included a 3-fold overrepresentation of approximately half of the *Salmonella* genome and 3-fold underrepresentation for the other half of the genome. Therefore, there are 3-fold differences in the target concentration between the two channels for all probes, and the expected hybridization ratio is 3 for specific hybridization. This was achieved by *Xba*I digestion of stationary phase *Salmonella enterica* sv Typhimurium LT2 genomic DNA, separation of the seven resulting fragments using pulsed field gel electrophoresis, capturing those fragments and pooling the six smaller fragments, while keeping the big fragment separate. Genomic DNA preparations from stationary phase LT2 were then supplemented either with the big fragment or with the pooled six smaller fragments, creating overrepresentations of the different halves of the genome.

Probes with stronger hybridization intensities displayed better specificity (Figure 5A). When each individual PDF and the predicted probe hybridization intensities were compared with the observed ratios, significant correlation was detected between the ratios and all the factors (Supplementary Figure 6), most significantly for PHFE, PMFE, PPBE and Complexity Score (8 bases). The Pearson correlation coefficients are listed in Supplementary Table 7. Note that PHFE is significantly and positively correlated with probe specificity. Probes with low PHFE values displayed both low specificity and relatively low sensitivity (as shown in Supplementary Figure 1–2).

As shown in Supplementary Figure 5, the relationships between $\log_2$ based ratios and some PDFs seem to be non-linear. For the sake of simplicity, only linear equations were considered in the current study.

Probe reproducibility

Data set 4, which includes 205 replicated hybridizations, was used to estimate probe reproducibility using CV. High probe reproducibility (corresponding to low CV values) is positively correlated with the observed probe hybridization intensities (Figure 5B). When examined individually, each PDF shows a significant but distinct level of association with CV (Supplementary Figure 6). PPBE and PHFE are the most significant factors. Correlation coefficients are listed in Supplementary Table 7.

Software

Programs for computing of PHFE, PMFE, probe dimer score and hairpin score, blast score and complexity score were written in Python. All programs, including parameters for computation, are freely available upon request.

DISCUSSION

Microarray probe hybridization signals are determined by the equilibrium of probe–target complex formation and probe–probe hybridization capability, and are also influenced by non-specific binding from the complex target. The PDFs we studied here covered these three aspects.

While Affymetrix Chips are designed for one-sample-for-one-array, it is very common to apply multiple samples onto the same array on customized platforms, including in-house spotted arrays and Nimblegen arrays. The natural log transformed intensity values from multiple arrays were averaged for each probe to minimize variation caused by sample processing and hybridization. We used genomic DNA samples because these hybridizations allow a comparison of probe performance under similar target concentrations.

Linear models were selected to model the relationships between individual PDFs and probe performance based on our observation that most scatter plots generated from multiple data sets consistently showed a linear relationship. The actual relationships may be far more complex. Nevertheless, from a practical point of view, linear models are easy to handle and generate more accurate predictions than more complex models based on model diagnosis with ARSS (34). The finding of these correlations is a useful first step in trying to understand the physical phenomena, which are clearly not subsumed in all the parameters currently in use. In future research, we plan to identify more advanced models (for example non-linear association models) that may reduce the ARSS we have achieved in the current study.

PMFE, dimer score and hairpin score are factors that estimate probe–probe hybridization capability. Of all the traditional PDFs (all factors except PPBE), PMFE correlated most significantly with probe hybridization intensity in all four data sets, followed by dimer score and hairpin score in most data sets. Although these three PDFs contain redundant information for estimation of the probe–probe hybridization capabilities, they cannot be simply replaced by each other as shown in the stepAIC analysis, which optimizes the complexity of the model versus the fit (32). All three PDFs therefore deliver unique information that needs to be considered for probe design.

Probe hybridization free energy (PHFE) is a long-established parameter for measuring probe–target hybridization capability in solution. In our study, PHFE was not as reliable in predicting probe hybridization intensity as other factors (PMFE, dimer score and hairpin), which may be largely due to the linkage of probes to a solid phase in microarray hybridizations. To compensate for the attachment of one end of the probe to the matrix, we introduced PPBE, which modifies the PHFE calculation by adding a positional weight parameter and iteratively fitting positional weights and dinucleotide stacking energies. PPBE showed much better capabilities of predicting probe hybridization than all other PDFs, and was a tremendous improvement over PHFE. The drawback of PPBE is that it is platform-dependent, and preliminary aCGH data are required for developing the PPBE model prior to application. The quality of the training data is critical for the construction of an accurate PPBE model. There are many factors that may result in bad quality arrays, e.g., bad sample quality. In order to solve these problems, we suggest that multiple CGH be performed using genomes without copy number variation to minimize the noise caused by sample processing.

Both PMFE and PHFE are sodium dependent. Generally, changes in free energy are linearly correlated to log-transformed sodium concentration (17), which has been confirmed by us on the Mfold web server (18) for PMFE and PHFE. That means that all oligonucleotide PMFE/PHFE values will change proportionally if the sodium concentration changes. Subsequently, these changes will be canceled out through adjustment of related coefficients in linear models. Therefore, changes in sodium concentration had no influence on the significance of linear modeling.

The PPBE model is empirical by nature, similar to the positional-dependent nearest neighbor model, which was designed for the Affymetrix array platform (34). Parameters of this model similarly need to be empirically estimated based on hybridization data, and significantly vary among different Affymetrix array platforms. At this stage, we do not understand the physical properties governing the parameters, but present a practical approach to optimize oligo design.

The position dependence of the weighting factors is a conspicuous feature in such models. In previous work, the sensitivity profiles of base C and base A change in a parabola-like fashion in a 25-base probe sequence, while the same profiles for G and T change monotonically (35–38). The overall position weighting factors change like the shape of a parabola, with peak and width varying across different GeneChip platforms (14,34,39). Our data reveal weight distribution patterns different from this previous work. Our data were obtained on two types of platforms: Nimblegen *in situ* synthesized oligonucleotide arrays and a spotted oligonucleotide array. For three Nimblegen platforms, the weights change linearly for the first 35~45 bases or so from the 3′ end and get weaker at the free end (Figure 1B, Supplementary Figure 2B and 2E). In contrast, a parabola-like curve is observed on the other platform (Supplementary Figure 2H). Although it is not the object of this article to explore a physical explanation for these differences, we point out some facts that may be important in further studies:

- We are using platforms of 50-mer probes, while the quoted previous work used 25-mer Affymetrix GeneChip platforms. Lengthening of the sequence on the platform inevitably reduces the importance of each single base or position, and weakens the position dependence.
- Unlike Affymetrix platforms and Nimblegen platforms, the probes of the spotted array in this study are linked to the array at the 5′ end, and there are no terminal oligonucleotide linkers between probes and the array surface. The impact of this difference is unknown, but it may reduce the freedom of a probe and even its effective length, leading to a pattern of position dependence similar to platforms of lower probe length, e.g. Affymetrix platforms.

For the fitting of the PPBE model, it is not critical whether weights or energies were fitted first. Either way, the final converged models reach similar ARSS values (average difference is <0.005). The final weights and pseudo-stacking energies are similar as well. We began to fit the models with the conventional dinucleotide stacking energies simply because the modes reached convergence faster. The dinucleotide stacking energies may express a relevant part of the physical properties underlying the model. However, further evidence is required to confirm this speculation.

Blast and complexity scores reflect occurrences of sequence segments similar to the probe, and are used for evaluating probe specificity. It would be simpler and easier to use cutoff thresholds for these PDFs to filter out bad quality probes. In this study, we applied four different patterns for the complexity score calculation, which are based on 2, 5, 8 and 11 base patterns. The complexity score (8 bases) showed better correlation with probe specificity than other complexity score patterns and blast score.

Langmuir isotherm oriented models were not included in our studies. Although Langmuir models were initially developed for adsorption of gases on glass surfaces (40), its variations have been widely applied in research for hybridization of oligonucleotides on DNA microarrays (13–16,41). In these models, the hybridization signal intensities were in essence divided into two parts: the hybridization of the probe with its perfect-matching target and the background noise. Although such models fit hybridization intensity values well for spiked-in genes and corresponding targets with controlled concentrations, they are of less help in screening probes for microarray design because they are based on the equilibrium constant or the change of standard Gibbs free energy ΔG°, which is a PDF of less sensitivity and specificity in comparison to PMFE and PPBE in our study. In contrast, platform-dependent empirical models based on pseudo free energies and position weights can make predictions very close to the observed hybridization intensities (34,39). This fact encouraged us to explore pure empirical models in microarray design.

In summary, we used aCGH as a model system to study the correlation between individual PDFs and probe performance during microarray hybridization. These individual correlations can be used as guidance for designing microarray probes for other complex experimental setups such as gene expression analysis. In gene expression microarray hybridization, non-specific binding, probe–targets complex formation and probe–probe binding capability will all be influenced by the varying concentrations of the targets. Systematic study of probe performance in such systems is beyond the scope of this study.

Nevertheless, if preliminary aCGH data is available, a complex multivariate linear model including the empirical factor PPBE can be developed and used for refining arrays. The model can predict a probe hybridization intensity value which will be an indicator of probe quality. Higher predicted intensity values will be equivalent to higher sensitivity, improved specificity and reproducibility. In practice, this strategy can be used for improving an existing array platform by replacing bad probes or by expanding the array by selecting probes predicted to perform well.

If aCGH data are unavailable for microarray platform design, we suggest using each individual PDF to filter or rank probes instead of using a complex model, because the coefficient parameters (intercept and slopes) vary significantly among different data sets/platforms. PMFE, hairpin score and probe dimer score can be used to rank probe qualities. PHFE, blast score and complexity score can be used to filter probes with low specificity. We have provided all correlation parameters generated from four data sets to be used as a guideline for filtering or ranking probes. All the programs for calculating individual PDFs are also available from the authors.

Supplementary Data

Supplementary Data are available at NAR Online.

FUNDING

The Prostate Cancer Foundation, the Mary Kay Ash Foundation, National Institute of Health grants (R01CA68822, U01CA114810, R01AI052237, R01AI073971, R01AI075093) to M.M. DOD Breast Cancer Research Program, (BC073899). Funding for open access charge: R01CA68822 and BC073899.

Conflict of interest statement. None declared.

REFERENCES

1. Ramsay,G. (1998) DNA chips: state-of-the art. *Nat. Biotechnol.*, **16**, 40–44.
2. Hayakawa,J., Mittal,S., Wang,Y., Korkmaz,K.S., Adamson,E., English,C., Ohmichi,M., Omichi,M., McClelland,M. and Mercola,D. (2004) Identification of promoters bound by c-jun/atf2 during rapid large-scale gene activation following genotoxic stress. *Mol. Cell.*, **16**, 521–535.
3. Hoemme,C., Peerzada,A., Behre,G., Wang,Y., McClelland,M., Nieselt,K., Zschunke,M., Disselhoff,C., Agrawal,S., Isken,F. *et al.* (2008) Chromatin modifications induced by pml-raralpha repress critical targets in leukemogenesis as analyzed by chip-chip. *Blood*, **111**, 2887–2895.
4. Wang,Y., Hayakawa,J., Long,F., Yu,Q., Cho,A.H., Rondeau,G., Welsh,J., Mittal,S., Belle,I.D., Adamson,E. *et al.* (2005) ‘‘promoter array’’ studies identify cohorts of genes directly regulated by methylation, copy number change, or transcription factor binding in human cancer cells. *Ann. N Y Acad. Sci.*, **1058**, 162–185.
5. Peeters,J.K. and derSpek,P.J.V. (2005) Growing applications and advancements in microarray technology and analysis tools. *Cell Biochem. Biophys.*, **43**, 149–166.
6. Wang,Y., Yu,Q., Cho,A.H., Rondeau,G., Welsh,J., Adamson,E., Mercola,D. and McClelland,M. (2005) Survey of differentially methylated promoters in prostate cancer cell lines. *Neoplasia*, **7**, 748–760.
7. Herring,C.D. and Palsson,B. (2007) An evaluation of comparative genome sequencing (cgs) by comparing two previously-sequenced bacterial genomes. *BMC Genomics*, **8**, 274.
8. Relgio,A., Schwager,C., Richter,A., Ansorge,W. and Valcrcel,J. (2002) Optimization of oligonucleotide-based dna microarrays. *Nucleic Acids Res.*, **30**, e51.

9. Chou,C.-C., Chen,C.-H., Lee,T.-T. and Peck,K. (2004) Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res.*, **32**, e99.
10. He,Z., Wu,L., Fields,M.W. and Zhou,J. (2005) Use of microarrays with different probe sizes for monitoring gene expression. *Appl. Environ. Microbiol.*, **71**, 5154–5162.
11. Singh-Gasson,S., Green,R.D., Yue,Y., Nelson,C., Blattner,F., Sussman,M.R. and Cerrina,F. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.*, **17**, 974–978.
12. Matveeva,O.V., Shabalina,S.A., Nemtsov,V.A., Tsodikov,A.D., Gesteland,R.F. and Atkins,J.F. (2003) Thermodynamic calculations and statistical correlations for oligo-probes design. *Nucleic Acids Res.*, **31**, 4211–4217.
13. Held,G.A., Grinstein,G. and Tu,Y. (2003) Modeling of dna microarray data by using physical properties of hybridization. *Proc. Natl. Acad Sci. USA*, **100**, 7575–7580.
14. Held,G.A., Grinstein,G. and Tu,Y. (2006) Relationship between gene expression and observed intensities in dna microarrays–a modeling study. *Nucleic Acids Res.*, **34**, e70.
15. Carlon,E. and Heim,T. (2006) Thermodynamics of RNA/DNA hybridization in high-density oligonucleotide microarrays. *Physica A Stat. Mech. Appl*, **362**, 433–449.
16. Fish,D.J., Horne,M.T., Brewood,G.P., Goodarzi,J.P., Alemayehu,S., Bhandiwad,A., Searles,R.P. and Benight,A.S. (2007) DNA multiplex hybridization on microarrays and thermodynamic stability in solution: a direct comparison. *Nucleic Acids Res.*, **35**, 7197–7208.
17. SantaLucia,J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearestneighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
18. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
19. Vallone,P.M. and Butler,J.M. (2004) Autodimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques*, **37**, 226–231.
20. Nielsen,H.B., Wernersson,R. and Knudsen,S. (2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. *Nucleic Acids Res.*, **31**, 3491–3496.
21. Wernersson,R. and Nielsen,H.B. (2005) Oligowiz 2.0–integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res.*, **33**, W611–W615.
22. Wernersson,R., Juncker,A.S. and Nielsen,H.B. (2007) Probe selection for dna microarrays using oligowiz. *Nat. Protocol.*, **2**, 2677–2691.
23. Rouillard,J.-M., Herbert,C.J. and Zuker,M. (2002) Oligoarray: genome-scale oligonucleotide design for microarrays. *Bioinformatics*, **18**, 486–487.
24. Rouillard,J.-M., Zuker,M. and Gulari,E. (2003) Oligoarray 2.0: design of oligonucleotide probes for dna microarrays using a thermodynamic approach. *Nucleic Acids Res.*, **31**, 3057–3062.
25. Mrowka,R., Schuchhardt,J. and Gille,C. (2002) Oligodb–interactive design of oligo dna for transcription profiling of human genes. *Bioinformatics*, **18**, 1686–1687.
26. Nordberg,E.K. (2005) Yoda: selecting signature oligonucleotides. *Bioinformatics*, **21**, 1365–1370.
27. Chen,H. and Sharp,B.M. (2002) Oliz, a suite of perl scripts that assist in the design of microarrays using 50mer oligonucleotides from the 3' untranslated region. *BMC Bioinformatics*, **3**, 27.
28. Reymond,N., Charles,H., Duret,L., Calevro,F., Beslon,G. and Fayard,J.-M. (2004) Roso: optimizing oligonucleotide probes for microarrays. *Bioinformatics*, **20**, 271–273.
29. Li,C. and Wong,W. (2003) The Analysis of Gene Expression Data: Methods and Software Chapter DNA-Chip Analyzer (dChip). Springer, New York, pp. 120–141.
30. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.
31. Jia,Z. and Xu,S. (2008) Bayesian mixture model analysis for detecting differentially expressed genes. *Int. J. Plant Genomics*, **2008**, 892927.
32. Akaike,H. (1974) A new look at the statistical model identification. *Automat. Control IEEE Trans.*, **9**, 716–723.
33. Geisser,S. (1975) The predictive sample reuse method with application. *J. Amer. Stat. Assoc.*, **70**, 320–328.
34. Zhang,L., Miles,M.F. and Aldape,K.D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.
35. Naef,F. and Magnasco,M.O. (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.*, **68(pt 1)**, 011906.
36. Mei,R., Hubbell,E., Bekiranov,S., Mittmann,M., Christians,F.C., Shen,M.-M., Lu,G., Fang,J., Liu,W.-M., Ryder,T. *et al.* (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **100**, 11237–11242.
37. Binder,H., Preibisch,S. and Kirsten,T. (2005) Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. *Langmuir*, **21**, 9287–9302.
38. Carlon,E., Heim,T., Wolterink,J.K. and Barkema,G.T. (2006) Comment on ''solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays''. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.*, **73(Pt 1)**, 063901; author reply 063902.
39. Zhang,L., Wu,C., Carta,R. and Zhao,H. (2007) Free energy of dna duplex formation on short oligonucleotide microarrays. *Nucleic Acids Res.*, **35**, e18.
40. Langmuir,I. (1918) The adsorption of gases on plane surfaces of glass, mica and platinum. *J. Am. Chem. Soc.*, **40**, 1361–1403.
41. Wick,L.M., Rouillard,J.M., Whittam,T.S., Gulari,E., Tiedje,J.M. and Hashsham,S.A. (2006) Onchip non-equilibrium dissociation curves and dissociation rate constants asmethods to assess specificity of oligonucleotide probes. *Nucleic Acids Res.*, **34**, e26.