Published online May 17, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 9 ***e66***
doi:10.1093/nar/gkl133

Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted

Alex Pozhitkov[1,3], Peter A. Noble[1], Tomislav Domazet-Lošo[2], Arne W. Nolte[3], Rainer Sonnenberg[3], Peer Staehler[4], Markus Beier[4] and Diethard Tautz[3,*]

[1]Civil and Environmental Engineering, University of Washington, Seattle, WA 98195, USA, [2]Ruder Bošković Institute, Division of Molecular Biology, Zagreb, HR-10002, Croatia, [3]Institute for Genetics, Cologne, D-50674, Germany and [4]Febit Biotech GMBH, Im Neuenheimer Feld 515, D-69120 Heidelberg, Germany

Received November 23, 2005; Revised December 20, 2005; Accepted March 14, 2006

ABSTRACT

Hybridization of rRNAs to microarrays is a promising approach for prokaryotic and eukaryotic species identification. Typically, the amount of bound target is measured by fluorescent intensity and it is assumed that the signal intensity is directly related to the target concentration. Using thirteen different eukaryotic LSU rRNA target sequences and 7693 short perfect match oligonucleotide probes, we have assessed current approaches for predicting signal intensities by comparing Gibbs free energy (ΔG°) calculations to experimental results. Our evaluation revealed a poor statistical relationship between predicted and actual intensities. Although signal intensities for a given target varied up to 70-fold, none of the predictors were able to fully explain this variation. Also, no combination of different free energy terms, as assessed by principal component and neural network analyses, provided a reliable predictor of hybridization efficiency. We also examined the effects of single-base pair mismatch (MM) (all possible types and positions) on signal intensities of duplexes. We found that the MM effects differ from those that were predicted from solution-based hybridizations. These results recommend against the application of probe design software tools that use thermodynamic parameters to assess probe quality for species identification. Our results imply that the thermodynamic properties of oligonucleotide hybridization are by far not yet understood.

INTRODUCTION

High throughput technologies, such as DNA microarrays, have significant potential for identifying organisms in many areas of biomedical science, including health care, biological defense and environmental monitoring. Several microarray platforms are currently used: dot blots on synthetic membranes or planar arrays (1,2) and gel-pad microarrays on glass slide (3–5). In addition, several platforms are under development: microbead microarrays (6,7) and electronic (8,9) and cantilever arrays (10). All platforms share the common attribute that a sensor detects a signal from target sequences hybridized to immobilized oligonucleotide probes. The intensity of this signal provides a measure of the amount of bound nucleic acid from a sample.

Ribosomal RNAs (rRNA) are particularly suitable for species identification procedures, because they occur universally, contain conserved as well as divergent regions, and are highly abundant in cells. Identification of microorganisms relies heavily on rRNA hybridization schemes (11,12), while applications for small eukaryotic soil or water organisms are currently emerging (14,15,18). The promise of these latter applications is that PCR amplification steps may not be required for detection, since multicellular organisms contain a sufficient amount of rRNA to allow direct detection of single individuals on a microarray platform (11–13,18).

In comparison to standard microarray applications for detecting specific mRNAs, there are extended requirements for the specific and reliable detection of organisms. First, since it is necessary to potentially distinguish closely related species, which differ only at a few nucleotide positions, one can only use relatively short oligonucleotides as probes, to ensure specificity. Second, because of the same reason, one has often only a limited set of options for choosing specific probes. And

[*]To whom correspondence should be addressed. Tel: 0049 221 470 2465; Fax: 0049 221 470 5975; Email: tautz@uni-koeln.de

finally, it is of particular importance that the specific probes yield a high signal to noise ratio, i.e. can discriminate accurately between perfectly matching and slightly mismatching targets.

Accordingly, it is necessary to have a reliable predictor for the hybridization performance of a specific probe. Tiling experiments with probes along specific mRNAs have shown that there can be huge differences in hybridization efficiency of probes (19,20). Furthermore, it has become clear that the simple notion that short oligonucleotides with a mismatch (MM) should hybridize less efficiently than perfect match (PM) probes is not always applicable. It has been shown that the hybridization intensity of MM probes can depend on the nucleotide type (i.e. A, C, G or T) and position of the MM relative to the termini (4,16) and that some MM probes yield higher signal intensities to the target than those of corresponding PM probes (17).

The focus of this study was to assess the utility of *in silico* predictions of probe-target duplex stabilities using DNA microarrays for detecting rRNA sequences in the context of possible applications for species identification. In particular, we investigate how well one can predict the hybridization performance of particular probes in the context of secondary structure predictions for the rRNA. In addition, we study the effects of single-base pair mismatches of all possible types and positions on probe-target hybridizations.

Our specific objectives were (i) to generate a set of probes forming a PM with target rRNA sequences, (ii) to measure the signal intensity of each probe on a microarray and to correlate fluorescent intensity values to theoretically-calculated duplex stability measures and (iii) to systematically assess the influence of single-base pair mismatches on signal intensity values of known target sequences.

We report lack of a simple relationship between hybridizations of probe-target duplexes as inferred from signal intensity values and *in silico* predictions based on Gibbs free energies. On the other hand, we can show that type and position of the MM significantly affects signal intensities of target sequences. Most interestingly, the order of stabilities of MM pairs in microarrays are different from that observed in solution, with pyrimidine–pyrimidine MM pairs being more stable than purine–purine pairs. However, even for these results the variances were high and cannot be explained for each individual oligonucleotide. Hence, it is currently not possible to predict *in silico* the performance of particular probes in microarray experiments. Accordingly we conclude that microarray designs for organism identification via rRNA hybridization will require meticulous testing of all possible oligonucleotide combinations.

MATERIALS AND METHODS

Experimental material

The ribosomal rRNA targets were derived from two different projects. For the first project, we used D3–D5 expansion segment fragments from the LSU of organisms that are present in the meiobenthos (15,18). These experiments were done in conjunction with Febit GmbH (Heidelberg), which includes also the systematic study of PM versus MM comparisons. In a second project, we have used D1–D2 expansion segment fragments from nine nematode species, for which we constructed PM tiling arrays in conjunction with NimbleGen Systems Inc. (Madison).

Table 1. Sequences used and numbers of perfect match (PM) and MM probes by sequence

Sequence	Organism	Accession no	Number of probes	
			PM	MM
1	Algae	DQ086764	39	2340
2	Chironomid	DQ086592	47	2820
3	Harpacticoid	DQ086556	42	2520
4	Ostracod	DQ086565	46	2760
Cb	*Caenorhabditis briggsae*[a]	—	824	—
Ce	*Caenorhabditis elegans*[a]	—	803	—
Cr	*Caenorhabditis remanei*[a]	—	824	—
Po	*Panagrolaimus* spec[a]	—	942	—
Pm	*Plectus minimus*[a]	—	965	—
Rb	*Rhabditis belari*[a]	—	779	—
Rd	*Rhabditis dolichura*[a]	—	795	—
Rt	*Rhabditis terricola*[a]	—	784	—
Tr	*Therimax rhabditidae*[a]	—	803	—
		Total	7693	10 440

[a] Species designation was done by E. Schierenberg (University of Cologne).

Target preparation

For the first set of experiments, cloned rDNA fragments (18) from four organisms were used (Table 1). The sequences were cloned into a pZErO-2 vector (Invitrogen Inc.). Depending on the orientation of the insert, the plasmids were cut with either SpeI or XbaI restriction enzymes and *in vitro* transcribed with SP6 or T7 RNA polymerase, respectively. The transcription and labeling mix contained 18 µl of a master-mix (10 mM ATP, CTP, GTP 8 µl each; 10 mM UTP 6 µl; 1 mM ChromaTide Alexa Fluor 546-14-UTP 20 µl; 10× Transcription buffer 16 µl; 40u/µl RNasin 8 µl); 2 µl of SP6 or T7 polymerase 20 u/µl; and 20 µl of the linearized plasmid at 50 ng/µl.

For the second set of experiments, ribosomal rRNA templates from nine nematode species were derived from a project in which the D1–D2 region of the LSU rRNA was sequenced (Table 1). The sequences were amplified using universal primers (28sFw-tailT3 5′-AATTAACCCTCACTAAAGGGAGCGGAGGAAAAGAAACTA-3′; 28sRew 5′-TACTAGAAGGTTCGATTAGTC-3′) of which the forward primer carries a tail with a T3-RNA Polymerase initiation site at its 5′ end. PCR products obtained with these primers were directly used for *in vitro* transcription. The transcription was performed with the MEGAscript Kit (Ambion) according to the instructions of the supplier. The master-mix was supplemented with 1.875 mM biotin-conjugated UTP (PerkinElmer) and 1.875 mM biotin-conjugated CTP (PerkinElmer) to label all transcripts.

Hybridization

Each of the rRNAs were diluted in hybridization solution (5× SSC, 0.2 mg/ml BSA, 12 mM ribonuclease inhibitor—Ribonucleoside Vanadyl Complex; New England Biolabs) to a final volume of 100 µl (3.75 ng/µl RNA) and heated to 80°C for 1 min. The following hybridization and washing protocol was used: (i) the microarrays were preheated to 70°C, (ii) the hybridization solution was added to each microarray and the microarrays were incubated at 80°C for 1 min, (iii) a low-stringency hybridization was performed by incubating the microarrays at 45°C for 24 h, (iv) the microarrays were then washed with a low-stringency buffer (5× SSC at 20°C, 3-fold volume exchange), (v) the first image of the microarrays was recorded, (vi) the microarrays were washed with a high-stringency buffer (0.1× SSC at 20°C, 3-fold volume exchange) and (vii) a second image of the microarrays was recorded.

Hybridization on the NimbleGen platform was performed according to the protocol routinely used at NimbleGen. Briefly, each biotin-labeled rRNA target was separately hybridized to the specific compartment on the 12-well NimbleGen array (a single array with 12 compartments physically isolated from each other), such that no interference between targets was allowed. Hybridization conditions were similar to that of Febit microarray, namely 45°C, 1 M Na^+. After 16–20 h hybridization, the microarray was washed with non-stringent and stringent buffers and images were recorded.

Probe design

A set of oligonucleotide probes was generated using a C++ program specifically written for this study. The set consisted of PM 20mer probes that were complementary to the rRNA targets (see Target preparation section). Randomly selected 20 nt long portions of the target were considered as potential hybridization sites. In addition to the PM probes, single-MM variants were designed. The entire array of these variants made up a complete set to investigate the effects of every position of the 20mer and every type of the MM on signal intensity values. All probes were replicated four times to provide a measure of intra-microarray reproducibility. In total, 42 456 oligonucleotide probes were synthesized by the GENIOM One® instrument (Febit GmbH, Heidelberg, Germany) on the microarray as described previously (19).

The probes for the NimbleGen experiments were designed as a tiling set (1 nt shift) of perfectly matching 25 nt oligonucleotides to the rRNA sequences of the nine nematodes. In total, 7519 oligonucleotides were synthesized on the surface of the 12-well NimbleGen array (a single array with 12 compartments physically isolated from each other), each well containing the full set of oligonucleotides.

Oligonucleotide arrays

A light-activated *in situ* oligonucleotide synthesis was performed within the GENIOM instrument on the activated 3D reaction carrier, which contained a glass-silicon-glass sandwich, using a digital micromirror device (Texas Instruments). Four individually accessible microchannels (referred to as arrays) were etched into the silicon layer of the DNA processor and connected to the microfluidic system of the GENIOM instrument acting as a custom DNA synthesizer. Oligonucleotides were synthesized using standard DNA synthesis reagents and RayDite 3′-phosphoramidites, carrying a 5′-photolabile protective group (Proligo LLC; Boulder, CO, USA). Prior synthesis, the array surface was activated and enough distance between oligonucleotides was secured with a spacer to facilitate probe-target interaction and avoid probe–probe interference.

Thermodynamic calculations

The following thermodynamic parameters were calculated using different software tools: free energies of probe-target binding ($\Delta G°_b$) and probe–probe dimerization ($\Delta G°_d$) at 45°C were calculated using an Excel macro written by Matveeva *et al.* (21); free energy of self-looping probes ($\Delta G°_p$) at 45°C was determined by Mfold program (22). In addition, free energy of the local denaturation of the target rRNA ($\Delta G°_t$), and the overall free energy of probe-target binding ($\Delta G°_{Ob}$) resulting from the consideration of all competing processes (i.e. $\Delta G°_b$, $\Delta G°_d$, $\Delta G°_p$, see Discussion), were calculated using RNAstructure v. 4.2 [(23), set with a fixed temperature of 37°C and a probe concentration of 1 μM, and 1 M Na^+]. All tools used the Nearest-Neighbor model.

Secondary structure prediction

The secondary structure of rRNA was determined by two alternative methods. First, the sequences were aligned to the best BLAST match from the European Ribosomal RNA Database (24), which contains an alignment of numerous LSU rRNA sequences with annotated secondary structure. Second, the rRNA targets were allowed to attain their lowest energy state. The free energies of the alternative folding were calculated using RNA folding software (RNAStructure).

Secondary structure of the targets used for hybridization with NimbleGen arrays was predicted only by energy minimization algorithm due to the lack of information about experimentally determined secondary structure of the D1–D2 expansion segment.

Data management and statistical analysis

The data were stored in a relational database created in Microsoft Access, which is available at http://faculty.washington.edu/pozhit/default.htm. The data were extracted through queries and analyzed in Microsoft Excel and SAS (Cary, N.C.). Principal component analysis (PCA) was employed to examine the distribution of the variables relative to signal intensity variables and to construct ordination plots. Pearson produce-moment correlation was used to determine the degree of association between variables. Linear regressions were used to estimate the relationship of one variable to another (25). The datasets were prepared for the ANOVA in the way that signal intensities of all duplexes where averaged using the median values of the four replicates. Median was used, as a measure of central tendency, which is less sensitive to outliers to account for possible hybridization artefacts. Median values of every probe containing a MM were normalized using the median of corresponding perfectly matched probe. These normalized values where then analyzed by three-way ANOVA using MM position, MM type and type of neighboring nucleotides (NN) as fixed factors. NN where defined as nucleotides located on the probe strand one position left and right from a MM position. Partial Eta squared (η^2) was used as a measure of the degree of association between normalized signal intensity and analyzed factors. The Hochberg's GT2 test was used for post-hoc analysis of contrast and pair-wise comparisons between means.

An artificial neural network (ANN) package (Neuroet, 26) was used to investigate the nonlinear relationships among input variables (i.e. $\Delta G°$ values) and outputs (i.e. signal

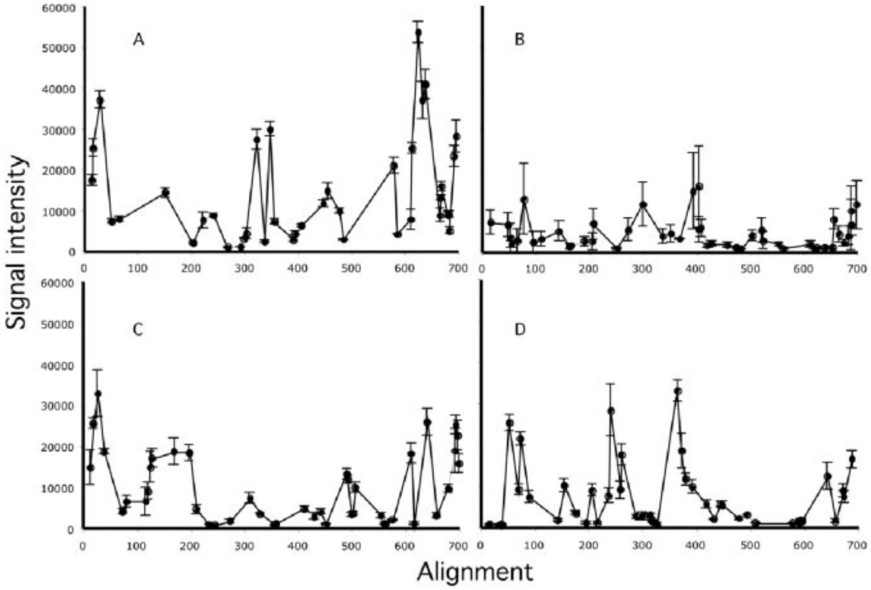

Figure 1. Signal intensity profiles of PM probes as a function of their position along target rRNAs. Error bars reflect the variance between the four replicates. The *x*-axis represents the position determined from the alignment based on the secondary structure predictions using the LSU database. (**A**) sequence 1; (**B**) sequence 2; (**C**) sequence 3 and (**D**) sequence 4.

intensity values). Unless otherwise specified, the following settings were used for training NNs: input and output scaling was set to standard linear (0,1); the logistic transfer function was used for hidden neurons and pure linear transfer function was used for output neurons; 80% of the data were used for training, 10% was used for testing and 10% was used for validating the NN; and, Levenberg–Marquardt error minimization was used to train the NN. The architectures of all NNs were optimized prior to conducting analyses by adjusting the number of hidden neurons (1 to 8) and identifying the architecture that provided the best predictive model. Comparison of different predictive models was conducted by computing their median Akaike's Information Criterion corrected (AICc) value (27) and determining the probability that one model was better than another. The model yielding the lowest AICc score contained the optimal number of hidden neurons.

RESULTS

In our first experiment, we constructed a set of PM probes for four different LSU rRNA fragments from meiobenthos organisms and synthesized every possible MM combination for all PM probes (Table 1). The hybridization profiles of PM probes to their respective target revealed large differences (up to 70-fold) in signal intensities by alignment position (Figure 1), similarly to what has been observed previously with mRNA targets (19,20). Matveeva *et al.* (19) had suggested that thermodynamic properties of probe folding and probe hybridization could partly explain these differences in hybridization efficiency. Luebke *et al.* (20) suggested that the predicted free energy of hybridization minus the predicted free energy for intramolecular folding of the probe provides a partial explanation, while no consistent correlation was found with the secondary structure of the mRNA targets.

Table 2. Comparison of free energies of LSU-RNA folding (kcal/mol)

Sequence	Alignment prediction	Minimum energy
1	−122.6	−216.7
2	−117.9	−232.5
3	−126.3	−248.6
4	−119.5	−250.5

Given that rRNA is known to form by far more extensive secondary structures than mRNA, we reasoned that if there would be any calculable effect of secondary structure on hybridization efficiency, it should be most pronounced for rRNA targets. Thus, in addition to calculating the parameters suggested by Matveeva *et al.* (19) and Luebke *et al.* (20), we considered also the free energy of the secondary structure of the rRNA.

Relationship of Gibbs free energy terms to signal intensity values of PM duplexes

To ensure that all possible known parameters are assessed, we calculated various Gibbs free energy terms singly or in combination using three different programs, which all consider nearest-neighbor models (see Materials and Methods). This includes the predicted free energy of hybridization (probe-target binding—ΔG°_b), probe hybridization (probe–probe dimerization—ΔG°_d), free energy for intramolecular folding of the probes (self-looping of probes—ΔG°_p), free energy of the local denaturation of the target rRNA (ΔG°_t) and the overall free energy of probe-target binding (ΔG°_{Ob}) resulting from the consideration of all competing processes (i.e. ΔG°_b, ΔG°_d, ΔG°_p, see Discussion). For considering secondary structure elements in rRNA, one can either use the secondary structure predictions inferred from alignments and experimental validation in ribosomes [taken from 'The European Ribosomal RNA

Table 3. R^2-values based on linear models for the regression between various change in Gibbs free energy terms and signal intensity values as a function of washing conditions and sequence

Sequence	Wash	n	Oligoanal (45°C)		Mfold (45°C)	RNA structure (37°C, 1 µM probe)						
			$\Delta G°_b$	$\Delta G°_d$	$\Delta G°_p$				Aligned		Minimum energy	
						$\Delta G°_b$	$\Delta G°_d$	$\Delta G°_p$	$\Delta G°_t$	$\Delta G°_{Ob}$	$\Delta G°_t$	$\Delta G°_{Ob}$
1	NS	156	0.06	—	0.09	—	0.05	0.09	0.18	0.29	—	0.06
	S		0.07	—	0.08	—	0.05	0.09	0.19	0.30	—	0.06
2	NS	188	0.05	0.08	—	0.07	0.08	—	—	0.05	—	0.11
	S		0.03	0.08	—	0.07	0.06	—	—	0.06	—	0.11
3	NS	168	0.05	0.28	0.15	0.26	0.05	0.12	—	0.11	—	0.09
	S		0.05	0.29	0.14	0.27	0.05	0.12	—	0.13	—	0.09
4	NS	184	0.12	—	0.32	—	0.12	0.25	—	—	—	—
	S		0.12	—	0.31	—	0.12	0.25	—	—	—	—

NS, non-stringent.
S, stringent.
'—' not significant at $\alpha = 0.05$.

Database', (24)], or the predictions derived from a folding algorithm that minimizes Gibbs free energy of the structure [RNAstructure, (23)]. A comparison of the free energies calculated for secondary structure predicted by alignment and that of predicted by minimum energy revealed that the alignment-defined secondary structure produced folds that were significantly different from their energy minimum (Table 2). This finding is consistent with the notion that the lowest energy state is not necessarily attained by mature rRNA and suggests that rRNA reaches a conformation that is between these extremes (i.e. those based on alignment predictions and those based on the energy minimum). However, the two versions of calculation that we use here are the only ones available based on the current knowledge.

Linear and nonlinear regression (polynomial, up to three terms) models were used to assess the relationship between the various $\Delta G°$ terms and signal intensity values of probe-target duplexes. In general, the models poorly explained the relationship between $\Delta G°$ terms and signal intensity values, regardless of microarray platform used (Febit or NimbleGen—see below), software package, washing conditions, target sequence or whether or not secondary structure of the RNA was considered when $\Delta G°$ was calculated (Table 3). Polynomial models did not fit the data (data not shown) and therefore were not further considered. One example of a weak linear correlation is shown for the relationship for $\Delta G°_{Ob}$ and signal intensity values for sequence 1 (Figure 2). In this case, up to 30% of the variability in the data were explained, while all other correlations for this sequence and the other sequences were worse (Table 3).

Because the first experiment included only relatively few PM probes, we sought to corroborate these findings with a second experiment, involving 7519 additional PM probes from nematodes (Table 1). In this experiment, the R^2-values for certain $\Delta G°$ terms on the nematode sequences explained as much as 74% of the variability of the data (Table 4). However, this was an exception rather than the rule since many R^2-values (~30%, 19 out of 63) were not statistically significant. Note that in the case of *Rhabditis terricola*, ΔG^0_{Ob} had no relation with signal intensity. It is particularly surprising since in theory, the ΔG^0_{Ob} should account for more variability than all other terms, but this is not the case, supporting the notion that predicted thermodynamic parameters do not accurately predict signal intensity values of duplexes with rRNAs in this experiment as well.

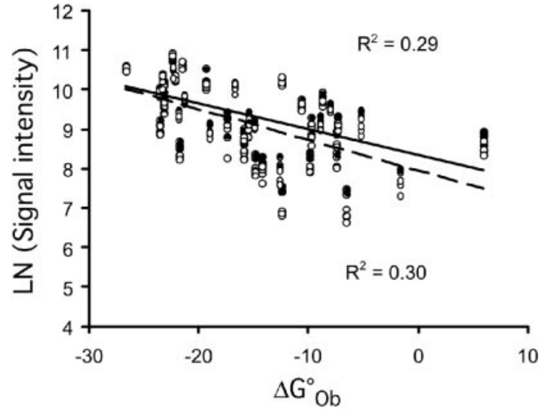

Figure 2. Relationship between ΔG^0_{Ob} and signal intensity for PM probes hybridized to target sequence 1. Free energy calculations were constrained by secondary structure predictions obtained from alignment. Closed circles, non-stringent wash; open circles, stringent wash. Solid trend line, non-stringent wash; dashed trend line, stringent wash.

These results are somewhat in contrast to the results from Matveeva *et al.* (19) and Luebke *et al.* (20), who found consistently weak correlations for the free energy terms they tested. However, the magnitudes of their correlations are within the range of the subset of experiments, where we also found some correlations. In balance, we can conclude from these results that signal intensity values for rRNA hybridizations are only poorly predicted by *in silico* software packages.

Since individual free energy parameters are such poor predictors, Luebke *et al.* (20) proposed a linear combination of two parameters, namely the predicted free energy of hybridization ($\Delta G°_b$) minus the predicted free energy for intramolecular folding of the probe ($\Delta G°_p$), as a reasonably good predictor of hybridization intensity. However, this is only one of all possible combinations of the parameters. To systematically evaluate all possible linear combinations of individual parameters, we employed a PCA, which can find even hidden relationships.

The initial PCA analysis involved constructing 2D ordination plots of $\Delta G°$ terms and GC values and color-coding each

Table 4. R^2-values based on linear models for the regression between various change in Gibbs free energy terms and signal intensity values as a function of sequence, for hybridization with the nematode sequences

Sequence[a]	n	Oligoanal (45°C)		RNAstructure (37°C, 1 μM probe)				
		$\Delta G^\circ{}_b$	$\Delta G^\circ{}_d$	$\Delta G^\circ{}_b$	$\Delta G^\circ{}_d$	$\Delta G^\circ{}_p$	$\Delta G^\circ{}_t$	$\Delta G^\circ{}_{Ob}$
Cb	824	—[b]	0.39	0.20	—	0.13	0.38	0.39
Ce	803	—	0.24	0.07	—	—	0.29	0.20
Cr	824	—	0.34	0.17	—	0.09	0.34	0.27
Pm	965	—	0.19	0.19	0.12	0.19	0.40	0.40
Po	942	0.08	0.49	0.40	0.66	0.05	—	—
Rb	779	0.67	—	0.11	—	0.06	0.15	0.21
Rd	795	0.74	—	0.08	0.06	—	0.09	0.07
Rt	784	0.54	—	0.10	0.16	—	0.06	—
Tr	803	—	0.16	0.14	0.05	—	0.16	0.07

Secondary structure of the target was determined by energy minimization using RNAstructure.

[a] Refer to Table 1 for sequences.

[b] Not significant at $\alpha = 0.05$.

Table 5. Correlation coefficients of variables relative to PCA axes by sequence (based on RNAStructure)

Probe characteristics	Sequence											
	1			2			3			4		
	PC axes (n = 156)			PC axes (n = 188)			PC axes (n = 168)			PC axes (n = 184)		
	1	2	3	1	2	3	1	2	3	1	2	3
GC	−0.60	−0.75	—	−0.50	−0.77	0.30	−0.62	−0.62	—	−0.54	0.56	−0.58
$\Delta G^\circ{}_b$	0.52	0.77	0.19	0.43	0.71	−0.49	0.66	0.65	−0.26	0.54	−0.75	0.32
$\Delta G^\circ{}_d$	0.57	0.31	−0.40	0.30	0.51	0.70	0.39	0.39	0.65	0.41	0.37	0.43
$\Delta G^\circ{}_p$	0.55	0.40	—	—	0.75	0.38	0.37	—	0.77	—	0.49	0.52
$\Delta G^\circ{}_t$ (aligned)	0.78	−0.25	0.47	−0.75	0.30	−0.46	−0.62	0.53	—	0.83	—	−0.44
$\Delta G^\circ{}_{Ob}$ (aligned)	−0.59	0.68	−0.36	0.92	—	—	0.86	—	—	−0.67	−0.33	0.58
$\Delta G^\circ{}_t$ (min energy)	0.79	—	−0.47	−0.59	0.70	—	−0.33	0.84	—	0.94	—	—
$\Delta G^\circ{}_{Ob}$ (min energy)	−0.52	0.61	0.58	0.80	−0.38	−0.24	0.68	−0.52	—	−0.84	−0.33	—
Ln signal intensity (stringent)[a]	0.50	−0.37	0.36	−0.37	—	0.64	−0.36	—	0.80	—	0.59	0.58
Proportion of eigenvalues:	0.39	0.24	0.15	0.34	0.29	0.19	0.33	0.26	0.20	0.39	0.20	0.20

'—' Not significant at $\alpha = 0.05$.

[a] Similar results were obtained for non-stringent conditions and not shown for brevity.

point on the plots by its corresponding signal intensity value. Examination of the four ordination plots revealed no obvious relationship between any of the variables and signal intensity values (data not shown). To more thoroughly investigate the relationship between ΔG° terms and signal intensity values, signal intensity values were included as a variable in PCA. PCA results of the data from different target sequences revealed that 78–82% of the total matrix variance was explained by three principal axes, with PC1 explaining 33–39%, PC2 explaining 20–29%, and PC3 explaining 15–20% of the total matrix variance (Table 5). However, Pearson correlation coefficients of the variables relative to the PC axes revealed inconsistent results for the data from different target sequences. For example, in the case of sequences 1 and 4, PC1 was most strongly positively correlated to $\Delta G^\circ{}_t$, while sequences 2 and 3 PC1 was negatively correlated to $\Delta G^\circ{}_t$. For sequences 2 and 3, $\Delta G^\circ{}_{Ob}$ was most strongly correlated to PC1, while this was negatively correlated for sequences 1 and 4. Similar results were also obtained for the other PC axes, indicating differences in the ordination of variables for data from different target sequences, which was also evident in the two dimension plots (data not shown).

The same analysis was carried out for the second experiment on the NimbleGen arrays. Similarly, examination of the nine ordination plots revealed no obvious relationship between any of the variables and signal intensity values (data not shown). In order to more thoroughly investigate the relationship between ΔG° terms and signal intensity values, the signal intensity values were included as a variable in PCA. PCA results of the data from different target sequences revealed that 83–91% of the total matrix variance was explained by three principal axes, with PC1 explaining 36–58%, PC2 explaining 17–29%, and PC3 explaining 14–24% of the total matrix variance (Supplementary Tables S1–S3). However, Pearson correlation coefficients of the variables relative to the PC axes revealed inconsistent results for the data from different target sequences.

To assess hidden nonlinear relationships, ANN analysis was used to investigate the relationship between ΔG° terms and signal intensity values, because neural networks have been shown to handle noisy, nonlinear data better than conventional linear approaches, such as PCA (28). For these analyses, the optimal number of hidden neurons was found to be 4, when ΔG° terms are used as inputs and signal intensity values are used as outputs. A model of the relationship between ΔG° terms and signal intensity values was generated by training an ANN using the data from one target sequence and cross validating the generated model by using data from another target sequence.

The correlation coefficients between actual and predicted signal intensity values of the models are shown in Table 6. A correlation close to 1 or −1 indicates that a model accurately

Table 6. Cross validation (CV) of ANN results. ΔG° terms from RNAstructure were used as inputs and signal intensity values were used as outputs

ANN trained by target sequence	*n*	Correlation coefficient by target sequence used for CV			
		1	2	3	4
1	156	0.94	0.34	—	0.29
2	188	0.25	0.82	—	—
3	167	0.30	0.26	0.98	0.25
4	184	0.37	—	—	0.98

In all cases, the ANN were trained using four hidden neurons, 10% of the data were used for testing, and another 10% was used for validation of the model. '—' Not significant at $\alpha = 0.05$.

Table 7. Three-way ANOVA of normalized signal intensities obtained in the stringent hybridization experiment

Component of variance	*df*	F	P	Partial η^2
MM position (MMP)	17	45.86701	1.20E−146	0.096
MM type (MMT)	11	41.68068	1.69E−88	0.059
NN	4	33.89908	4.52E−28	0.018
MMP × MMT	187	1.38596	0.000454	0.034
MMP × NN	34	2.864812	5.75E−08	0.013
MMT × NN	44	2.036145	6.43E−05	0.012
MMP × MMT × NN	374	0.831392	ns	
Error	7322			

predicts signal intensity values when provided with ΔG° terms, while a correlation value close to zero implies no correlation. We also included the correlation coefficient for each ANN trained with the same data since it represents the 'best' possible correlation for each model. Note that the 'best' possible correlations were based on the analysis of all predicted and actual signal intensity values in the data from one sequence. The reason why the 'best' correlations were not exactly 1 or −1 was because only 80% of the data were used to train the ANN model. The remaining (20%) of the data were used for local testing and validation of the model.

Poor correlations of ANN predictions to actual values could be attributed to over- or under-training of the ANNs. For example, an over-trained ANN learns to memorize the training data, and consequently generates high correlations between predicted and actual values for data it was trained on, but poor or no correlations for test data that was not used for training. We carefully trained each ANN model to generalize predictions by optimizing the architecture of the model prior to training, and by stopping training when there was no change in the error over a specified period of time or, after a specified number of iterations [see ref. (26)]. This approach ensured that each ANN model produced outputs that accurately predict signal intensity values for ΔG° terms not used for training. We conclude that the reason the ANN models are unable to accurately predict signal intensity values when provided with data from sequences not used for training, was because there is a poor relationship between ΔG° terms and signal intensity values. These findings corroborate the PCA results and suggest that no combinations of the ΔG° terms are major determinants for predicting signal intensity values.

An assessment of the effects of mismatches on signal intensity values

Three-way ANOVA was used to assess the effects of MM position, MM type, and the type of NN that flank a MM, on normalized signal intensity values (see Materials and Methods). The model revealed that all three factors had low, albeit significant effects on the normalized signal intensity values (Table 7). Most of the variance of normalized signal intensity was explained by MM position (9.6%), followed by MM type (5.9%), whereas NN type had the lowest effect on the observed variance among the factors (1.8%) measured by partial η^2. In addition, there were interactions among all combinations of two factors (Table 7). The strongest interaction was observed for MM positions and MM types (3.4%), while interactions between MM position and NN type (1.3%) and between MM type and NN type (1.2%) were comparable. We were not able to detect significant effects of simultaneous interactions among all three factors (Table 7).

Moving the position of the MM away from the 5′ or 3′ termini to the center of the probe significantly decreased signal intensities (Figure 3). ANOVA post-hoc contrasts between means showed that duplexes with MM between positions 6 and 15 formed a homogenous group ($\alpha = 0.05$) with the most pronounced effects on duplex stability. This finding indicates that the most optimal discrimination of MM from PM duplexes is provided with the MM in the middle of the duplex. However, we emphasize that this was an average result, and note that in some individual cases, MM probes with central mismatches (positions 9–11) were observed to have signal intensities that were equal or up to 1.6 times higher than that of corresponding PM probes.

A heat map on the effects of the MM type by position is shown on Figure 4. Clearly, there are differences in average signal intensity by MM type and position. Post-hoc ANOVA contrasts were able to discriminate five homogenous groups (Figure 5), two groups with clearly separated extremes: (i) GA and GG mismatches (which destabilize duplexes the most) and (ii) TC, TU and TG mismatches (which destabilize duplexes the least). Differences in signal intensity values as a function of position are clearly visible for these two groups in Figure 4. These findings indicate that distinguishing PM duplexes from those containing a single-MM was highly dependent on the type of MM pairs.

To more fully understand simple patterns of MMs as a function of type and position, we pooled MM types to three categories: purine–purine, pyrimidine–pyrimidine and purine–pyrimidine MM pairs. Figure 6 shows that signal intensities of duplexes with pyrimidine–pyrimidine MM pairs were more similar to PM duplexes than purine–pyrimidine or purine–purine MM pairs. An interaction was evident at the termini of probes where differences in the normalized signal intensities among MM pairs were more pronounced towards the 3′ end of the probe. Differences in intensity values at the 5′ and 3′ end might be due to the orientation of the probe on the microarray since the 3′ end was closest to the microarray surface.

Figure 7 illustrates the effects of the type of NN that flank a MM on normalized signal intensity values. We analyzed separately the cases when a MM is located at the termini of sequence from the cases when it is located elsewhere. The reason for this is that a MM at the termini could have only one neighboring nucleotide, while in all other positions it has two neighbors. We assessed the effect of NN by categorizing

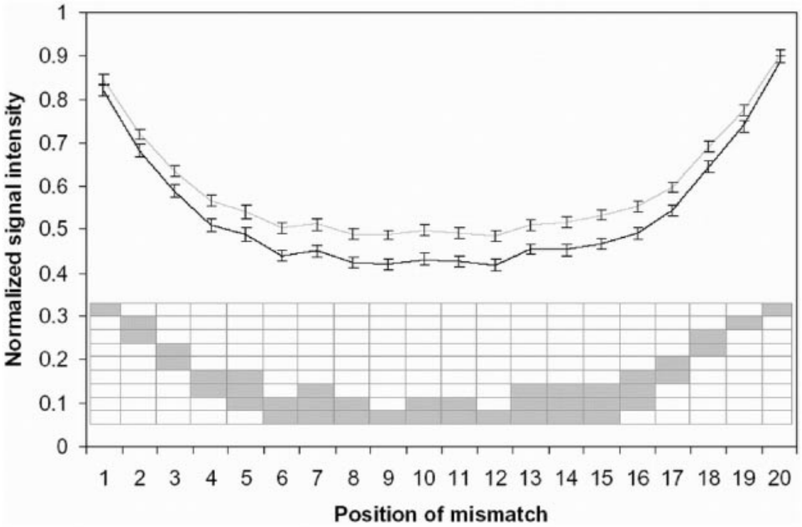

Figure 3. Average signal intensity values of MM duplexes at positions 1 to 20 normalized to that of the PM duplex (based on around 400 values per position). Error bars represent ±1 standard error of mean. Intensities from low-stringency (gray line) and high-stringency (black line) experiments are shown. Differences between low-stringency and high-stringency experiments are significant for overall dataset and for the individual MM positions at least on the $\alpha = 0.05$ level by paired *t*-test. Shaded cells in each row represent a homogenous set of means revealed by GT2 post-hoc analysis at $\alpha = 0.05$ level.

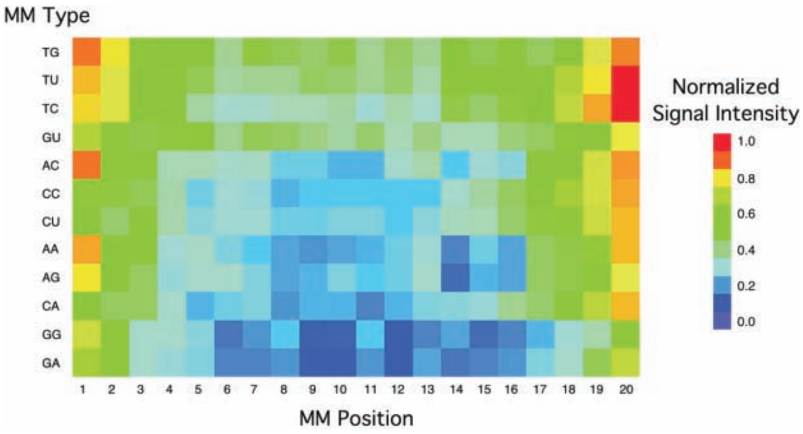

Figure 4. Heat map of MM type by position as a function of average signal intensity, normalized to the signal intensity of the PM duplex. Each box represents at least 120 replicates.

probes with terminal MMs into two states: those with a purine and those with a pyrimidine neighbor. Elsewhere we grouped MMs having nucleotides flanking a MM into three categories: purines only, pyrimidines only, and purine–pyrimidine combinations. In addition to the asymmetric impact of MM type at the end of the probe described earlier, we detected asymmetry at the ends of the probe concerning NN type. Figure 7 shows that purine neighbor at the 5′ end stabilized the duplex more than pyrimidine (GT2 post-hoc test, $P = 0.001$). Surprisingly, at the 3′ end the opposite trend is true—although it is not statistically significant. When non-terminus mismatches were considered, the most stabilizing effect on the duplex occurred with purine flanking neighbors. Pyrimidine flanking neighbors yielded the lowest duplex stability. Purine–pyrimidine neighbors were in the middle of these two extremes (Figure 7B, all differences are significant at $P = 0.001$ by GT2 post-hoc test). Interactions between NN type and MM position and type are significant as previously stated. However, due to the minor effects on the variance and peculiar patterns of interaction, we excluded it from further discussion.

DISCUSSION

The thermodynamic properties of nucleic acid duplex formation and dissociation in solution have been well established (29). For example, the behavior of a probe and a target sequence in solution can be predicted by using a nearest-neighbor model (30). However, duplex formation using surface-immobilized DNA oligonucleotides is less well understood, presumably due to the complex factors affecting the kinetics and thermodynamics of target capture. Some factors affecting duplex formation on DNA microarrays include: probe density, microarray surface composition and the stabilities of oligonucleotide-target duplexes, intra- and intermolecular self-structures and RNA secondary structures

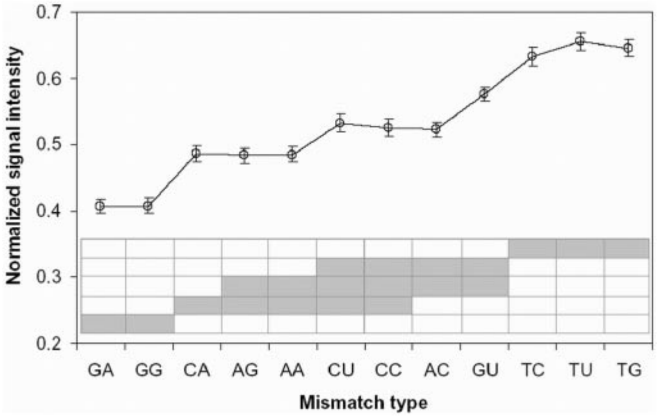

Figure 5. Average signal intensity values of MM duplexes categorized by MM type. Shaded cells in each row represent a homogenous set of means revealed by GT2 post-hoc analysis at $\alpha = 0.05$ level. Error bars represent ±1 standard error of mean. Note that each member of mirrored MM pairs (GU and TG, TC and CU, GA and AG, CA and AC) belongs to the different homogenous group. All differences within mirrored pair of mismatches are significant at least at $\alpha = 0.01$ level in GT2 pair-wise comparisons.

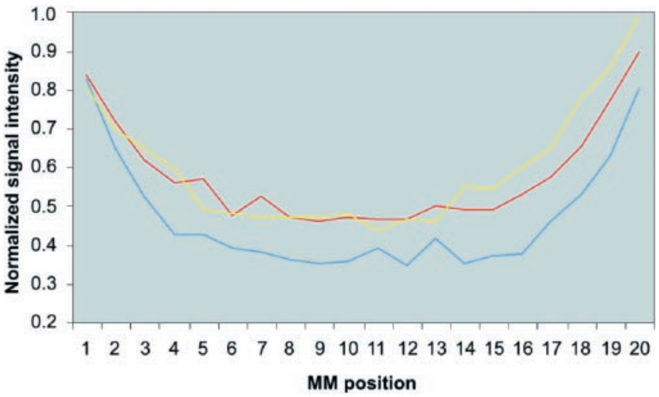

Figure 6. Average signal intensity values of MM duplexes at positions 1 to 20 normalized to signal intensity of the PM duplex by MM type. Pyrimidine:pyrimidine MMs, yellow; Purine:pyrimidine, red; Purine:purine MMs, blue.

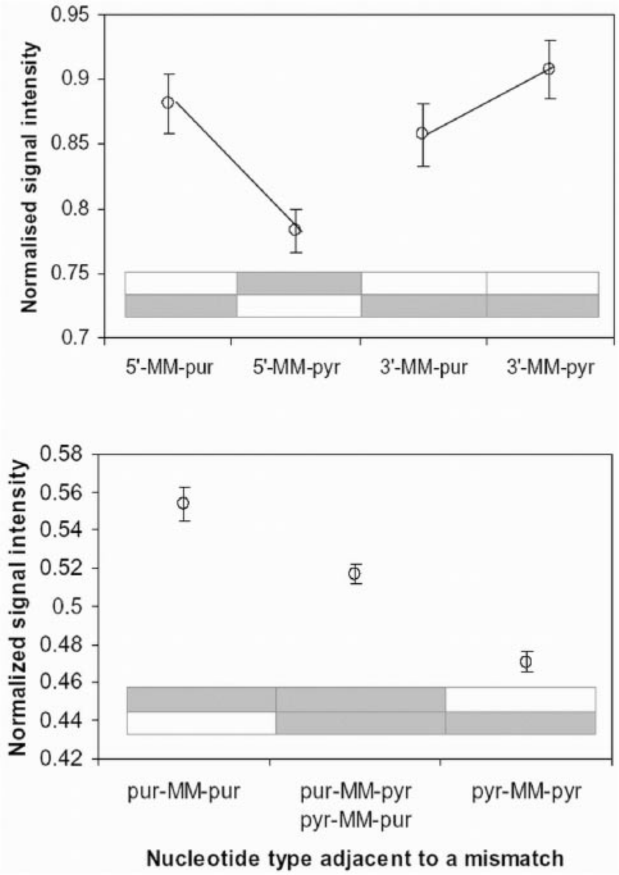

Figure 7. Average signal intensity values of MM duplexes categorized by nucleotide type flanking a MM in the probe sequence. Shaded cells in each row represent a homogenous set of means revealed by GT2 post-hoc analysis at $\alpha = 0.05$ level. Error bars represent ± standard error of mean. Upper panel: Effect of nucleotide types adjacent to a MM at the end of probe. Note that at 5′ end neighboring purine residues stabilize duplex more then pyrimidine residues ($P = 0.001$), while pattern at the 3′ is opposite and there is no significant difference in GT2 pair-wise comparison. Lower panel: Effect of nucleotide types flanking a MM. Purine residues are stabilizing duplex more than purine–pyrimidine combinations or pyrimidine alone. All differences are significant at $\alpha = 0.001$ level in GT2 pair-wise comparisons.

(21,31,32). We reasoned that examination of the thermodynamic stabilities of probe-target duplexes using existing models might provide valuable information on the relationships between predicted stabilities of targets hybridized to immobilized probes and their corresponding signal intensity values on DNA microarrays. We also reasoned that the position and type of MM, and the nature of neighboring bases to the MM should also affect signal intensity values.

Relationship between thermodynamic predictions and signal intensity values

As proposed by Matveeva *et al.* (21), hybridizations of a target to probes on a planar microarray are affected by several overlapping processes which include: (i) the affinity of a target to bind to a probe (ΔG°_b), (ii) the formation of stem–loop structures of a probe (ΔG°_p), (iii) the formation of secondary structure (loops and helices) of a target (ΔG°_t) and (iv) probe to probe dimerization (ΔG°_d) (Figure 9). In addition, the overall Gibbs free energy of binding (ΔG°_{Ob}) can be calculated by considering the combined effects of all four terms (i.e. ΔG°_b, ΔG°_p, ΔG°_d and ΔG°_t) on hybridization predictions (23).

ΔG°_t values could have been considered of special relevance for rRNA, because of the known potential to form extensive secondary structures. The values were calculated by considering the secondary structure of the targets as determined from the LSU rRNA database. Two different approaches were used to calculate ΔG°_t since we did not know if aligned (constrained) or not aligned (free form) secondary structure significantly affected free energies determination. The aligned folding preserves the annotated single strands while the not aligned folding allows the molecule to reach a conformation that corresponds to the calculated global energy minimum.

In our analysis, all Gibbs free energy terms were poorly correlated and linear and nonlinear regressions had low R^2-values, to signal intensity values of PM probe-target duplexes. Moreover, there does not appear to be a consistent pattern in Gibbs free energy terms by target sequence.

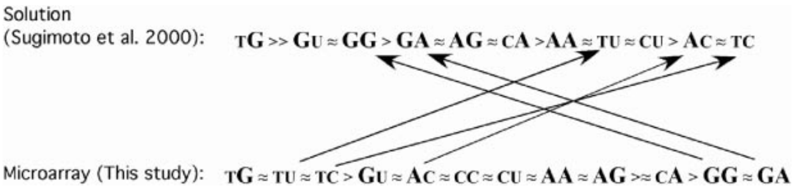
Solution
(Sugimoto et al. 2000):
TG >> GU ≈ GG > GA ≈ AG ≈ CA >AA ≈ TU ≈ CU > AC ≈ TC
Microarray (This study):
TG ≈ TU ≈ TC > GU ≈ AC ≈ CC ≈ CU ≈ AA ≈ AG >≈ CA > GG ≈ GA

Figure 8. The order of stability of RNA/DNA duplexes with a single-base pair MM pairs in solution Sugimoto *et al.* (39) and on the microarray. For each MM-pair: probe DNA is on the left and target RNA is on the right. The size of the letter distinguishes purines (large) from pyrimidines (small). Lines depict major differences in the order of stability.

Furthermore, while PCA and ANN analyses were able to establish significant correlations between Gibbs free energy terms and signal intensity values when each sequence was separately analyzed, cross validation using different target sequences revealed inconsistent results. These findings indicate that Gibbs free energy terms and signal intensity values are target dependent and suggest that other factors, such as surface density of the probes (31) and/or brush effects (32), might have greater effects on signal intensity values than previously anticipated.

Thermodynamic stabilities of target RNA hybridized to immobilized oligonucleotide probes have been investigated in the following studies: (i) Naef and Magnasco (17) and Mei *et al.* (33) both described an *ad hoc* model that examined the affinity of a probe to a target based on the sum of position-dependent base-specific contributions, (ii) Zhang *et al.* (34) described an *ad hoc* model that considered position-dependent nearest-neighbor effects, (iii) Held *et al.* (35) examined the effects of free energies of RNA/DNA duplex formation and (iv) Wu and Irizarry (36) developed a model that considered both stochastic and deterministic aspects of probe-target hybridizations. The unifying features of these studies are: (i) they are all based on the analysis of multiple probes targeting mRNA transcripts (i.e. expression data), (ii) with exception of Held *et al.* (35), they only considered single-base pair mismatches that occurred in the middle of the duplex (position 13 of 25mers), (iii) they assumed that binding of various RNA targets was independent and noncompetitive. Unfortunately, none of the studies satisfactorily predicted signal intensity values on oligonucleotide microarrays since there were significant disagreements between actual and predicted values.

The effect of single-base pair mismatches on duplex signal intensity values

In solution, single-base-mismatches in oligonucleotide probes can stabilize or destabilize a duplex depending on the identity of the MM, its position in the helix and its neighboring base pairs (37). Although it has been established that there are differences in experimental results conducted in solution versus those using microarrays (17), we investigated MM type and position, and neighboring base pairs on signal intensity values because the effects of these variables on planar microarrays are not well understood.

We found that the position of the MM affected duplex stability (as inferred by signal intensity values). This finding is consistent with previous studies (16) showing that terminal mismatches are less destabilizing than internal ones. We also found asymmetry in the pattern of signal intensity values by position. Specifically, normalized signal intensities among MM pairs were more pronounced towards the 3′ end of the probe. This phenomenon was presumably due to orientation of the probe on the microarray since the 3′ end was tethered to the microarray surface. Since electrostatic effects of the microarray surface are distance dependent, mismatches closest to the 3′ end might be responsible for the observed effect (38)—although further studies are needed to verify this.

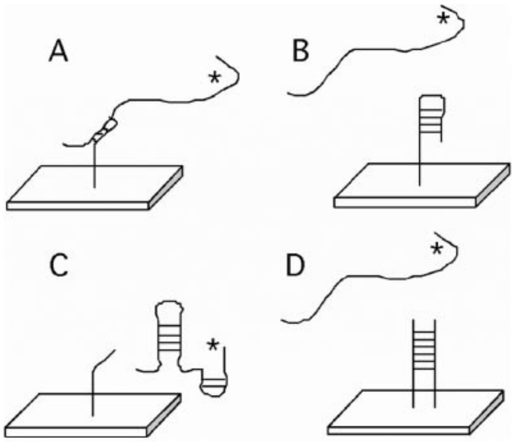

Figure 9. Depiction of four competitive processes on signal intensity values. Each panel shows a labeled (*) target and an immobilized probe on a microarray. (**A**) hybridization of a target to a probe; (**B**) probe self-folding; (**C**) folding of the target and (**D**) dimerization of adjacent probes.

Studies conducted in solution have shown that different MM types cause diverse effects on duplex stability (39). We found that the order of stabilities of MM pairs in solution were different from that observed in microarrays (Figure 8). In general, the microarray results revealed that pyrimidine–pyrimidine MM pairs were more stable (left side of Figure 8) than purine–purine MM pairs (right side of Figure 8). This result was anticipated since purines are composed of large double-ringed nucleotides that distort the geometry of the double helix—incurring a large steric and stacking cost. Hence, MM pairs containing purine destabilize the duplex and have lower signal intensity values than its corresponding PM duplex. Pyrimidine–pyrimidine mismatches, on the other hand, are composed of small single rings that do not distort the geometry of the double helix, resulting in higher stabilities and signal intensity values than MM pairs containing one or two purines. Possible reasons for the discrepancy in the order of stabilities in solution versus those in microarrays include: the number of samples examined [Sugimoto *et al.* (39) versus this study, 52 versus 10 440 MM pairs, respectively], the size of the oligonucleotide probes on the microarrays (9mers versus 20mers, respectively), and neighboring bases employed (C-MM-G, G-MM-C, C-MM-C, G-MM-G versus every possible neighboring combination, respectively).

Interestingly, we also found some asymmetries in signal intensity values of mismatches that contain the same pair of bases (e.g. GA and AG; Figures 4 and 5) but differ only in the sense that MM nucleotide is either on the probe or target strand of the duplex. Sugimoto *et al*. (39) also found this asymmetry for mismatches occurring in short oligonucleotides in solution. This effect can currently not be explained.

The bases neighboring a probe MM can also significantly affect signal intensity values. Bases neighboring a MM at the 5′-terminus had contrasting affects on signal intensity values to those at the 3′-terminus. For example, at the 5′-terminus, purine neighbors had higher signal intensity values than pyrimidine neighbors, while at the 3′-terminus, purine neighbors had the opposite effects on signal intensity values (Figure 7). These differences may be due to steric effects of MMs at the 3′-terminus, which are close to the microarray surface. In contrast to bases neighboring a MM at the terminus, bases neighboring an internal MM yielded a consistent trend: mismatches flanked by purine neighbors had a more stabilizing effect on duplexes than other combinations. These findings are consistent with Sugimoto *et al*. (39), which showed that both the MM type and the neighboring bases of the probe influenced duplex stability.

CONCLUSION

In summary, there is little evidence to support the notion that thermodynamic parameters accurately predict signal intensity values of duplexes with rRNAs on oligonucleotide (20–25 nt) DNA microarrays. As a consequence, we recommend that thermodynamic criteria (e.g. 21, 40) not be used for designing oligonucleotide probes for species identification—instead, an empirical verification of each probe is advised to obtain the best signal intensities. Thorough empirical calibration of microarrays has recently been shown to be useful in a related field [methylation pattern analysis via microarray-based genotyping, (41)] to select best probes within one or two optimization and selection cycles. With respect to MM effects, we find that the position and type of single-base pair MM and composition of neighboring bases affected the stability of duplexes on DNA microarrays—but in different ways from what is known from experiments conducted in solution. Key differences are: (i) positional affects of MMs were asymmetric, presumably due to steric affects of mismatches close to the surface of the microarray; (ii) pyrimidine–pyrimidine MM pairs were more stable than purine–purine MM pairs and (iii) duplexes with mismatches flanked by purine neighbors were more stable than other combinations of neighbors. However, we point out that even these effects, although consistent, have only a partial predictive value.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Kristina Hillesland, Nicolas Pinel and Paul Berube for their critical comments on earlier versions of the manuscript. The authors also thank Einhard Schierenberg for providing the nematode strains for sequencing. This work was supported by a grant from the Ministerium für Schule Wissenschaft und Forschung des Landes Nordrhein-Westfalen as well as grants 1U01DE014955-01 from NIH/NIDCR and R-82945801 from EPA-CEER-GOM to P.A.N. Funding to pay the Open Access publication charges for this article was provided by Deutsche Forschungsgemeinschaft.

Conflict of interest statement. None declared.

REFERENCES

1. DeSantis,T.Z., Stone,C.E., Murray,S.R., Moberg,J.P. and 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.
2. Wilson,W.J., Strout,C.L., DeSantis,T.Z., Stilwell,J.L., Carrano,A.V. and Andersen,G.L. (2002) Sequence-specific identification of 18 pathogenic microorganisms using microarray technology. *Mol. Cell. Probes*, **16**, 119–127.
3. Pozhitkov,A., Smidt,H., Könneke,M., Chernov,B., Yershov,G. and Noble,P.A. (2005) Evaluation of gel-pad oligonucleotide microarray technology using artificial neural networks. *Appl. Environ. Microbiol.*, **71**, 8663–8676.
4. Urakawa,H., Noble,P., A, El Fantroussi,S., Kelly,J.J. and Stahl,D.A. (2002) Single-base-pair discrimination of terminal mismatches by using oligonucleotide microarrays and neural network analyses. *Appl. Environ. Microbiol.*, **68**, 235–244.
5. Liu,W.T., Mirzabekov,A.D. and Stahl,D.A. (2001) Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. *Environ. Microbiol.*, **3**, 619–629.
6. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S.J., McCurdy,S., Foy,M., Ewan,M. *et al*. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
7. Brenner,S., Williams,S.R., Vermaas,E.H., Storck,T., Moon,K., McCollum,C., Mao,J.I., Luo,S.J., Kirchner,J.J., Eletr,S. *et al*. (2000) *In vitro* cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl Acad. Sci. USA*, **97**, 1665–1670.
8. Barlaan,E.A., Sugimori,M., Furukawa,S. and Takeuchi,K. (2005) Electronic microarray analysis of 16S rDNA amplicons for bacterial detection. *J. Biotechnol.*, **115**, 11–21.
9. Zimmermann,K., Eiter,T. and Scheiflinger,F. (2003) Consecutive analysis of bacterial PCR samples on a single electronic microarray. *J. Microbiol. Methods.*, **55**, 471–474.
10. McKendry,R., Zhang,J.Y., Arntz,Y., Strunz,T., Hegner,M., Lang,H.P., Baller,M.K., Certa,U., Meyer,E., Guntherodt,H.J. *et al*. (2002) Multiple label-free biodetection and quantitative DNA-binding assays on a nanomechanical cantilever array. *Proc. Natl Acad. Sci. USA*, **99**, 9783–9788.
11. Peplies,J., Lau,S.C.K., Pernthaler,J., Amann,R. and Glockner,F.O. (2004) Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. *Environ. Microbiol.*, **6**, 638–645.
12. Small,J., Call,D.R., Brockman,F.J., Straub,T.M. and Chandler,D.P. (2001) Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **67**, 4708–4716.
13. Chandler,D.P., Newton,G.J., Small,J.A. and Daly,D.S. (2003) Sequence versus structure for the direct detection of 16S rRNA on planar oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **69**, 2950–2958.
14. Blaxter,M., Elsworth,B. and Daub,J. (2004) DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.*, **271**, S189–S192.
15. Markmann,M. (2000) Entwicklung und Anwendung einer 28S rDNA Sequenzdatenbank zur Ausschlusselung der Artenveilsalt limnischer Meiobenthosfauna im Himblick auf den Einsatz moderner Chiptechnologie. PhD Thesis, Ludwig Maximilians University, Munich.

16. Urakawa,H., El Fantroussi,S., Smidt,H., Smoot,J.C., Tribou,E.H., Kelly,J.J., Noble,P.A. and Stahl,D.A. (2003) Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **69**, 2848–2856.
17. 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**. Art. No. 011906 Part 1.
18. Markmann,M. and Tautz,D. (2005) Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Phil. Trans. R. Soc. B*, **360**, 1917–1926.
19. Baum,M., Bielau,S., Rittner,N., Schmid,K., Eggelbusch,K., Dahms,M., Schlauersbach,A., Tahedl,H., Beier,M., Guimil,R. *et al.* (2003) Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. *Nucleic Acids Res.*, **31**, e151.
20. Luebke,K.J., Balog,R.P. and Garner,H.R. (2003) Prioritized selection of oligodeoxyribonucleotide probes for efficient hybridization to RNA transcripts. *Nucleic Acids Res.*, **31**, 750–758.
21. 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 oligoprobes design. *Nucleic Acids Res.*, **31**, 4211–4217.
22. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
23. Mathews,D.H., Burkard,M.E., Freier,S.M., Wyatt,J.R. and Turner,D.H. (1999) Predicting oligonucleotide affinity to nucleic acid targets. *RNA-Publ. RNA Soc.*, **5**, 1458–1469.
24. Wuyts,J., Perriere,G. and de Peer,Y.V. (2004) The European ribosomal RNA database. *Nucleic Acids Res.*, **32**, D101–D103.
25. Sokal,R.R. and Rohlf,F.J. (1981) *Biometry 2nd edn.*. W.H. Freeman and Co., NY.
26. Noble,P.A. and Tribou,E. (2006) Neuroet: an easy-to-use artificial neural network for ecological and biological modeling. *Ecological Modelling* (in press).
27. Motulsky,H. and Christopoulos,A. (2004) *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting.* Oxford University Press, NY.
28. Noble,P.A., Almeida,J.S. and Lovell,C.R. (2000) Application of neural computing methods for interpreting phospholipid fatty acid profiles of natural microbial communities Appl. *Environ. Microbiol.*, **66**, 694–699.
29. Marky,L.A. and Breslauer,K.J. (1987) Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. *Biopolymers*, **26**, 1601–1620.
30. SantaLucia,J., Allawi,H.T. and Seneviratne,A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.
31. Peterson,A.W., Heaton,R.J. and Georgiadis,R.M. (2001) The effect of surface probe density on DNA hybridization. *Nucleic Acids Res.*, **29**, 5163–5168.
32. Halperin,A., Buhot,A. and Zhulina,E.B. (2005) Brush effects on DNA chips: thermodynamics, kinetics, and design guidelines. *Biophys. J.*, **89**, 796–811.
33. 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.
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. 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.
36. Wu,Z. and Irizarry,R.A. (2004) Stochastic models inspired by hybridization theory for short oligonucleotide microarrays. In *Proceedings of RECOMB 2004*. San Diego, CA.
37. Kierzek,R., Burkard,M.E. and Turner,D.H. (1999) Thermodynamics of single mismatches in RNA duplexes. *Biochemistry*, **38**, 14214–14223.
38. Vainrub,A. and Pettitt,B.M. (2002) Coulomb blockage of hybridization in two-dimensional DNA arrays. *Phys. Rev. E.*, **66**. Art. No. 041905 Part 1.
39. Sugimoto,N., Nakano,M. and Nakano,S. (2000) Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes. *Biochemistry*, **39**, 11270–11281.
40. Tanaka,F., Kameda,A., Yamamoto,M. and Ohuchi,A. (2005) Design of nucleic acid sequences for DNA computing based on a thermodynamic approach. *Nucleic Acids Res.*, **33**, 903–911.
41. Mund,C., Beier,V., Bewerunge,P., Dahms,M., Lyko,F. and Hoheisel,J.D. (2005) Array-based analysis of genomic DNA methylation patterns of the tumour suppressor gene p16(INK4A) promoter in colon carcinoma cell lines. *Nucleic Acids Res.*, **33**, e73.