Arnold Vainrub
B. Montgomery Pettitt
Department of Chemistry, University of Houston, Houston, TX 77204-5003

Theoretical Aspects of Genomic Variation Screening Using DNA Microarrays

Received 9 October 2003; accepted 1 November 2003
Published online 8 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bip.20008

Abstract: *We present a theoretical model for typical microarray-based single nucleotide polymorphism (SNP) assay of small genomic DNA amount. We derived the adsorption isotherm expressing the on-array hybridization efficiency in terms of genomic target sequence and concentration, oligonucleotide probe sequence and surface density, hybridization buffer, and temperature. This isotherm correctly describes the surface probe density effects, the sensitivity peak, and the melting temperature depression, and is in accord with published experiments. We discuss optimization of parallel SNP genotyping. Our estimates show that SNP detection at a single temperature in aqueous hybridization buffer is restricted by DNA regions that differ by less than 20% in GC content. We predict that the variety of genotyped SNPs could be substantially extended using an assay design with high probe density and a large fraction of probes hybridized.* © 2004 Wiley Periodicals, Inc. Biopolymers 73: 614–620, 2004

Keywords: *microarray; SNPs; hybridization; isotherm*

INTRODUCTION

The detection of genomic nucleic acids with sensitivity and specificity when using blends with similar nucleotide sequences is in high demand in DNA microarray assays. In particular, a single base replacement should be recognized and reliably detected in single nucleotide polymorphism (SNP) and point mutation microarray studies. This discrimination between a perfect match (PM) and a single base mismatch (SMM) sequence is routinely done by hybridization to oligonucleotide probes when only a few sequences are present in assayed solution. However, microarrays examine very complex nucleic acid mixtures. The sample is often prepared by fragmentation of polymerase chain reaction (PCR)-amplified cellular mRNA or genomic DNA and contains many millions of different sequences present in the genome. Therefore, under microarray experimental conditions, for each specific probe a large variety of targets with partial sequence complementarities occur that can cause a background cross-hybridization signal and make detection and discrimination of PM and SMM transcripts much more complicated. This is especially true when dealing with genes with a low intrinsic expression. Although the mismatched DNA duplex is less stable thermodynamically, a stronger hybridization signal from SMM probes is often observed in the analysis of genomic samples.

Oligonucleotide microarrays are widely used for SNPs screening.[1–15] Recently impressive improvements in technology and bioinformatics tools have been reported.[6,8,13] For example, in a high-throughput study using Affymetrix Variation Detection Arrays™

Correspondence to: B. Montgomery Pettitt e-mail: pettitt@uh.edu; vainrub@uh.edu

Contract grant sponsors: NIH; R. A. Welch Foundation; TiiMES/NASA; NCC-1-02038; Texas Learning and Computation Center.

~80% of all the haploid and diploid sites were readable.[6] Achieving high genotyping accuracy allows confirming already known SNPs and facilitates discovery of new SNPs. However, the study[6] failed to genotype 20% of sites, and the need for further advancement of the technology is widely recognized. Theoretical considerations of on-array hybridization thermodynamics could be useful to provide scientific background to the still mainly empirical art of microarray assays.

As a physical model, we use the Coulomb blockage hybridization isotherm[16,17] that accounts for the dominant effects of the interface electrostatic interactions.[18a] In the present study, we use this theoretical framework to calculate the sensitivity, specificity, and throughput in on-array based screening of genome variations and discuss potential improvements.

ON-ARRAY HYBRIDIZATION ISOTHERM

We use the on-array hybridization isotherm that we derived previously.[16,17] In this section we briefly describe the physical meaning of the on-array isotherm and introduce the notation. In the next section, we derive and discuss a particular form of the isotherm for a typical microarray assay including the (usually small) available amount of genomic nucleic acid.

Nucleic acids are strong polyanionic electrolytes; their hybridization in solution phase is strongly affected by the electrostatic repulsion and screening. As known, addition of cationic counterions (usually Na^+ from added NaCl) shields the repulsion and results in a strong increase of the double-helix thermal stability.[19] This interaction is modified for on-array hybridization, where the target strand may be repelled by, or attracted to, the surface depending on material, and repels not only from a single hybridization partner on the surface, but from other arrayed probes as well. If we assume that this additional repulsion scales with the charge of the hybridized part (its length is Z_P) of the target and the charge accumulated in probe layer (including charge of hybridized targets), we can approximate the Gibbs free energy difference as

$$wZ_P(Z_P n_P + Z_T n_D) \tag{1}$$

where Z_P and Z_T are the probe and target length (number of bases), n_P and n_D are the surface densities of immobilized probes and hybridized duplexes (in molecules/m^2), and w (in J m^2 $mole^{-1}$) is an interaction strength constant that includes the effects of screening and thus is expected to diminish as the added salt concentration grows. Notice the effect of counter ion condensation on DNA that reduces its effective charge[21] could be also included in w. Thus we believe that Eqn. (1) could be a useful phenomenological mean field approximation for both linear Debye-Huckel and non-linear screening regimes at low and high salt concentrations, respectively for this system. We used Eq. (1) to derive an on-array hybridization isotherm[16,17] that relates the equilibrium hybridization efficiency θ ($0 \le \theta \le 1$) with the assayed nucleic acid target concentration C_0 (expressed in mol/L)

$$C_0 = \frac{n_D S}{N_A V} + \frac{n_D}{n_P - n_D} \exp\left(\frac{\Delta H_0 - T\Delta S_0}{RT}\right) \times \exp\left[\frac{wZ_P(Z_P n_P + Z_T n_D)}{RT}\right] \tag{2}$$

where ΔH_0 and ΔS_0 are the reference state enthalpy in J mol^{-1} and entropy in J mol^{-1} K^{-1}, respectively, for hybridization in isotropic solution, T is the temperature in Kelvin, R is the gas constant, Z_P and Z_T are the length of probe and target oligonucleotides in number of bases, n_P is the surface density of probes in m^{-2}, and the parameter $w = 4 \times 10^{-16}$ Jm^2 mol^{-1} for our conditions with 1 M NaCl. Often, in microarray assays the concentration C_0 and volume V of the hybridization solution are small. Thus, the target depletes to the concentration, $C_0 - n_D S/N_A V$, during the course of hybridization. Thus, the first term on the right accounts for a depletion of targets in solution due to their hybridization to the array. The first exponent stands for the free energy of duplex formation. The second exponential represents the target-array repulsion that causes the Coulomb (electrostatic) blockage of hybridization at high surface probe density n_P. As previously shown,[16,17] this equation fits experiment rather well and we refer to additional firming the isotherm model in Eq. (2). Notice, both species duplexes (DNA-DNA, RNA-DNA, etc.) as well as sequence composition effects are simple to account for by taking the proper ΔH and ΔS for our isotropic solution reference state.

ON-ARRAY HYBRIDIZATION OF SMALL GENOMIC NUCLEIC ACID SAMPLES

Usually, the available quantity of genomic nucleic acid is ~1 μg of mRNA. Therefore, assuming (as in human genome) the occurrence of 32 000 different genes with an average mRNA length of 2800 bases,

we find that $\sim 2 \times 10^7$ gene copies can be hybridized to an ideal microarray. This number is much smaller than the 5×10^8 probes in a typical microarray spot with 100×100-μm^2 size and $n_P = 5 \times 10^{12}$ strands/cm^2. Hence $n_D \ll n_P$ and a strong depletion of assayed nucleic acid can occur in microarray assays. In addition to hybridization to a matched probe, targets can hybridize with several of the roughly 10^3–10^5 probe spots with complementary fragments or undergo nonspecific adsorption. Assuming 10% of the targets to hybridize with a matched probe, in the above mentioned example we get $n_D = 2 \times 10^{10}$ cm^{-2} and $wZ_PZ_Tn_D/RT = 0.08$ for $Z_P = 25$ and $Z_T = 100$ bases. Using the inequalities

$$\frac{n_D}{n_P} \ll 1, \qquad \frac{wZ_PZ_Tn_D}{RT} \ll 1 \tag{3}$$

we get

$$\frac{n_D}{n_P - n_D} \exp\left(\frac{wZ_PZ_Tn_D}{RT}\right) \simeq \frac{n_D}{n_P}\left(1 + \frac{n_D}{n_P}\right)\left(1 + \frac{wZ_PZ_Tn_D}{RT}\right) \simeq \frac{n_D}{n_P} \tag{4}$$

Substitution into Eq. (2) gives a simplified isotherm equation:

$$C_0 = \frac{n_D S}{N_A V} + \frac{n_D}{n_P} \exp\left(\frac{\Delta H_0 - T\Delta S_0}{RT}\right)\exp\left(\frac{wZ_P^2 n_P}{RT}\right) \tag{5}$$

Interestingly, in contrast to the transcendental algebraic Eq. (4), the hybridization signal n_D can be found from Eq. (5) in closed form convenient for analytical analysis:

$$n_D = C_0\left[\frac{S}{N_A V} + \frac{1}{n_P} \times \exp\left(\frac{\Delta H_0 - T\Delta S_0}{RT}\right)\exp\left(\frac{wZ_P^2 n_P}{RT}\right)\right]^{-1} \tag{6}$$

The fraction of targets that are hybridized,

$$\alpha = \frac{n_D}{n_T} = \frac{n_D S}{C_0 N_A V} \tag{7}$$

is a useful characteristic of the hybridization efficiency ($0 \le \alpha \le 1$) and for further analysis we rewrite Eq. (6) as

$$\alpha = \left[1 + \frac{N_A V}{S n_P} \exp\left(\frac{\Delta H_0 - T\Delta S_0}{RT}\right)\exp\left(\frac{wZ_P^2 n_P}{RT}\right)\right]^{-1} \tag{8}$$

From this equation, we can calculate the melting temperature that corresponds to hybridization of 50% of the targets ($\alpha = \frac{1}{2}$)

$$T_M = \frac{\Delta H_0 + wZ_P^2 n_P}{\Delta S_0 + R \ln(S n_P / N_A V)} \tag{9}$$

and the width of the melting transition, which is given by

$$\delta T = \left.\frac{dT}{d\alpha}\right|_{\alpha = 1/2} = \frac{4R(\Delta H_0 + wZ_P^2 n_P)}{[\Delta S_0 + R \ln(S n_P / N_A V)]^2} \tag{10}$$

Thus, we have derived a convenient set of analytical forms for the fraction of targets hybridized, Eqn. (8), for the melting temperature, Eqn. (9), and the melting width, Eqn. (10), to compare with experiment.

COMPARISON WITH THE LANGMUIR ISOTHERM

Since the Langmuir isotherm is widely used to describe on-array hybridization[24–26] and to compute transcript concentration levels,[27,28] we briefly discuss and compare with our approach. The Langmuir isotherm is a particular (ideal) case of Eqs. (2), (5) and (8) when $w = 0$

$$\alpha_L = \left[1 + \frac{N_A V}{S n_P} \exp\left(\frac{\Delta H_0 - T\Delta S_0}{RT}\right)\right]^{-1} \tag{11}$$

First, we notice that any experiment *at a single surface density*, n_P, cannot distinguish between the Langmuir equation and Eq. (8) because the equations coincide if ΔH_0 in the Langmuir equation is replaced by

$$\Delta H_{0L} = \Delta H_0 + wZ_P^2 n_P \tag{12}$$

However, a clear difference appears if n_P is varied. As an example, we consider the 25-mer probe 5′-GTCCGATAAGCCTGTGTCCAATAAC-3′ with a perfectly matched (complementary) target 3′-CAGGCTATTCGGACACAGGTTATTG-5′, a probe spot area $S = 0.01$ mm^2, and a hybridization solution

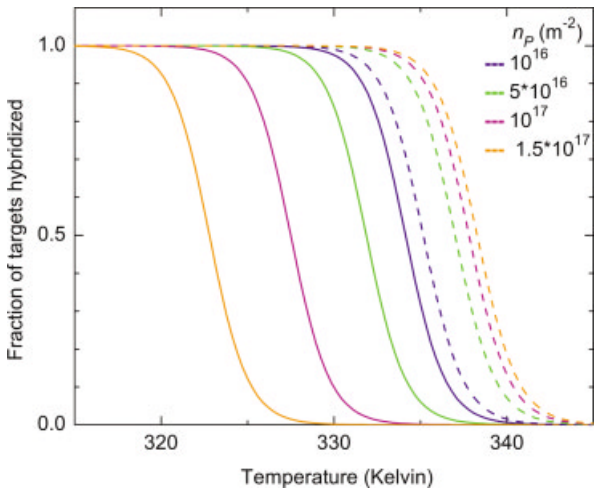

FIGURE 1 Comparison of the isotherm Eq. (8) (solid) and Langmuir isotherm (dashes) at different surface probe densities n_P as indicated.

volume $V = 30$ μl. For the specified duplex in aqueous solution with 1 M NaCl, the thermodynamic stability parameters are $\Delta H_0 = -819.3$ kJ mol^{-1} and $\Delta S_0 = -2.229$ kJ mol^{-1} K^{-1}, as given by a nearest-neighbor model.[29] Figure 1 shows the duplex melting curves $\alpha(T)$ and $\alpha_L(T)$ calculated for different values of n_P. The melting temperature T_M and melting transition width δT in Figure 1 obey Eqs. (9) and (10), respectively. Figure 1 demonstrates a clear difference in n_P dependence; $\alpha(T)$ shifts to lower temperatures, whereas in contrast $\alpha_L(T)$ moves toward higher temperatures with increasing n_P. Only the $\alpha(T)$ behavior is in accord with hybridization experiments for arrays on glass[30–32] and gold[7–12] the substrates that show depression of the melting temperature at high n_P.

Figure 2 compares $\alpha(n_P)$ and $\alpha_L(n_P)$ at different hybridization temperatures. At low temperatures, the hybridization is almost complete in both cases ($\alpha = 1$) and differences appear only at high surface density, where $\alpha(n_P)$ declines. However, at elevated temperatures, the difference is striking; in the Langmuir case, $\alpha_L(n_P)$ increases with n_P and saturates at $\alpha = 1$, whereas the curves from Eq. (8) pass through a maximum and then decay exponentially. Physically, the maximum appears as a result of increasing Coulomb blockage of hybridization at higher n_P is explained in Ref. 16. To find the position of the maximum, we calculate the derivative with respect to n_P from Eq. (8)

$$\frac{d\alpha}{dn_P} = \frac{\alpha^2 N_A V}{S n_P}\left(\frac{1}{n_P} - \frac{wZ_P^2}{RT}\right)\exp\left(\frac{wZ_P^2 n_P}{RT}\right) \quad (13)$$

or substituting α from Eq. (8)

$$\frac{d\alpha}{dn_P} = \frac{N_A V}{S n_P}\left(\frac{1}{n_P} - \frac{wZ_P^2}{RT}\right)\exp\left(\frac{wZ_P^2 n_P}{RT}\right)$$

$$\times\left[1 + \frac{N_A V}{S n_P}\exp\left(\frac{\Delta H_0 - T\Delta S_0}{RT}\right)\exp\left(\frac{wZ_P^2 n_P}{RT}\right)\right]^{-1} \quad (14)$$

We find the zero of the derivative at

$$n_{Pm} = \frac{RT}{wZ_P^2} \quad (15)$$

which defines the location of the maximum.[17] Experimentally, the maximum and the decrease in the hybridization targets efficiency with increasing n_P has been observed for DNA mounted on aminated polypropylene,[2] glass,[1,21–23,33] and gold[34–36] surfaces. The results are in accord with the behavior predicted from Eq. (8), as shown in Figure 2, but are not possible using the ideal Langmuir isotherm model. Interestingly, the hybridization decay is already noticeable at $n_P \approx 10^{12}$ cm^{-2},[22,23,35,36] corresponding to a mean interprobe distance on the surface of roughly 10 nm. This is large compared with the diameter of a DNA double helix, roughly 2-nm diameter. This suggests rather that electrostatic repulsion, and not steric hindrance, is dominant in this range.

To compare with theory quantitatively, we should take into account that all the experiments mentioned above were performed with an excess of targets $n_T > n_P$. In this case, condition $n_D \ll n_P$ of Eq. (2) is not satisfied. Therefore, instead of Eq. (4), the more general Eq. (2) should be used. We showed that Eq. (2) also predicts the maximum or saturation of n_D with

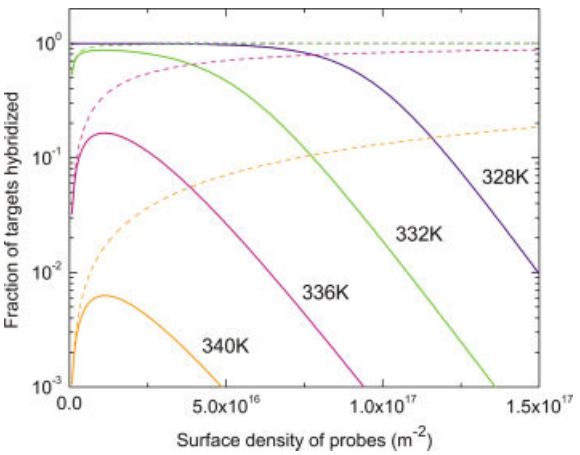

FIGURE 2 Fraction of targets that are hybridized according with Eq. (8) (solid) and Langmuir isotherm (dashes) as a function of the surface probe density at various temperatures.

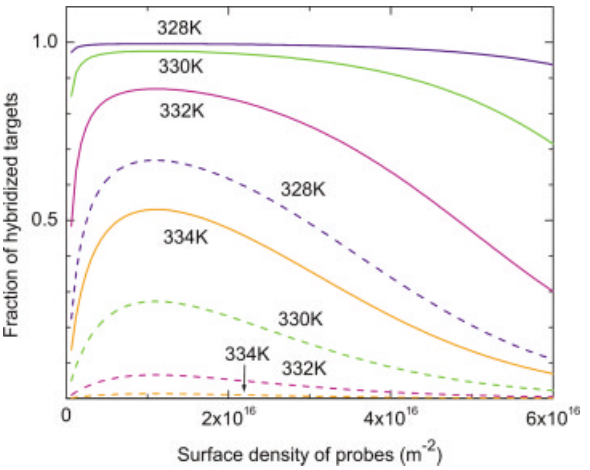

FIGURE 3 Fraction of targets hybridized for perfect match (solid) and single mismatch (dashes) targets as a function of the surface probe density at various temperatures.

increasing n_P. However, compared with n_{Pm} given by Eq. (15) the maximum shifts to higher probe density ($n_{Pm} + n_D$). The analysis shows good quantitative accord with experiments and will be published elsewhere.[37]

MATCH–MISMATCH DISCRIMINATION

To find optimal conditions for the detection of SNPs and point mutations in genotyping, we consider as an example, hybridization of the probe specified above with perfectly matched and single base mismatched targets. For the mismatched target 3′-CAGGCTATTCGGCCACAGGTTATTG-5′, we choose replacement of the central base A by C resulting in reduced duplex stability with $\Delta H_0 = -760.8$ kJ mol^{-1} and $\Delta S_0 = -2.09$ kJ mol^{-1} K^{-1} in solution phase.[29] Figure 3 displays the fraction hybridized versus probe surface density calculated from Eq. (8), for matched (α^{PM}) and mismatched (α^{MM}) targets for a set of temperatures. As expected, both strong signal and clear discrimination occur near the melting temperature T_m. Taking into account the typically strong background signal in genomic sample hybridizations, usually the difference

$$\Delta\alpha = \alpha^{PM} - \alpha^{MM} \tag{16}$$

is more directly measured, rather than the ratio (α^{PM}/α^{MM}). Figure 4 shows $\Delta\alpha(n_P)$ for a range of temperatures. It clearly shows that to optimize the match-mismatch discrimination the hybridization temperature should be precisely tuned depending on the probe density n_P. The temperature should be inside or close to the range between SMM and PM T_M's, and it lowers as n_P increases. Although as seen in Fig. 4 the discrimination slightly improves with increasing n_P, using smaller n_P and elevated temperature could be preferable because of suppressed cross-hybridization, reduced background, and better melting of the secondary structures.

PARALLEL DETECTION OF SNPs WITH DIFFERENT MELTING TEMPERATURE

The average SNP frequency is approximately one in a thousand DNA bases. More than two million human genome SNPs are already mapped and accessible in databases (e.g., see http://snp.cshl.org). Therefore, a large number of SNPs can be scored for each individual in a studied population. High-density arrays with hundreds of thousands of probe spots are ideal for this purpose. However, the variety of SNPs detectable in parallel is restricted by the demand that their melting temperatures be similar. To consider this issue, Figure 5 shows the match–mismatch discrimination as a function of temperature for probe and target oligonucleotides defined above. The discrimination curve shifts to lower temperature as n_P increases with no substantial change in width. Therefore, the range of screened SNPs with different melting temperatures is quite independent of the surface probe density. For example, if $\Delta\alpha = 0.2$ is chosen as a discrimination threshold, the corresponding width in Figure 5 is 9 K and SNPs with a difference in melting temperature of

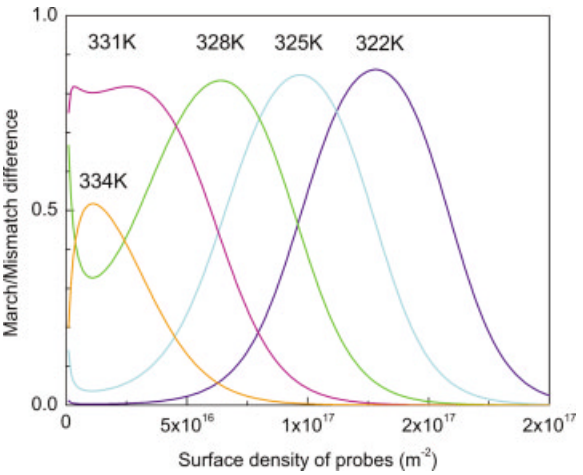

FIGURE 4 Match/mismatch discrimination versus surface density of probes at different temperatures as specified.

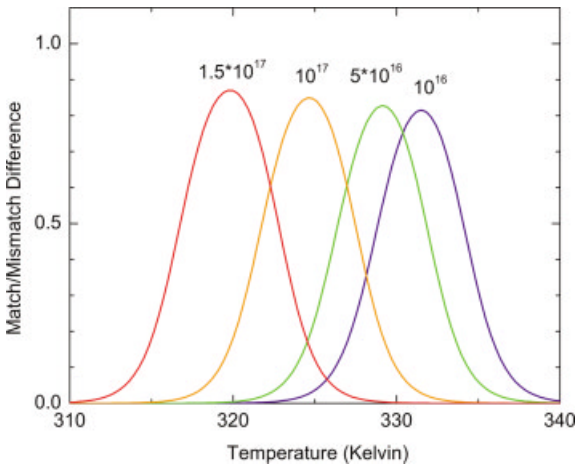

FIGURE 5 Match/mismatch discrimination as a function of temperature at different surface probe densities n_P (m^{-2}) as specified.

up to 9 K can be detected. This means that probes could differ by approximately 20% in GC content (since T_m rises by ~4.1 K with a 10% GC increase[19]). This composition range is insufficient and could be improved to some extent by using buffers containing tetramethylammonium chloride or chemically modified oligonucleotides (reviewed in Refs. 4 and 10). Another interesting possibility for equalizing T_M values is using variable length probes.

In a single microarray experiment, the interrogated set of SNPs is restricted by the melting temperature range

$$\delta T_M = \delta T + (T_M^{PM} - T_M^{MM}) \qquad (17)$$

centered at average T_M. For the case considered where $n_D \ll n_P$, the melting width δT in Eq. (10) only slightly depends on n_P as Figure 1 shows. This explains the practically unchanged width of $\delta(T)$ curves for different n_P in Figure 5. However, when $n_D \sim n_P$, ΔT increases considerably at high n_P, as shown previously.[37] Therefore, the range of T_M values can be expanded to allow reading of SNPs in both GC- and AT-rich DNA regions. This regime is of interest for reliable high-throughput SNPs screening and will be more completely described in a forthcoming paper.

CONCLUSION

We derived a simple phenomenological isotherm for an on-array hybridization of a small amount of nucleic acid. This approach describes the hybridization yield as a function of solution (ionic strength and hybridization temperature), target (sequence, length and concentration) and array (probe sequence, length and surface density) parameters and captures much of the details for DNA array hybridization experiments.

We considered, in detail, a typical assay when the quantity of assayed genomic DNA is small compared with the hybridization capacity of the probe spots. Under this condition, we derived the simplified isotherm Eq. (5) and applied it to find optimal conditions for SNP detection. Highly parallel SNPs screening is restricted by variation in the melting temperature preventing SNPs detection in both GC- and AT-rich DNA regions. We estimated that SNPs can be scored in a single temperature experiment if their GC content differ by less than 20%. However, we found that the variety of assayed SNPs can be drastically extended if majority of probes are hybridized.

A. V. thanks Michael Shortreed for valuable comments on on-array hybridization of complex genomic nucleic acids mixtures. The authors thank Lloyd Smith for stimulating discussions.

REFERENCES

1. Guo, Z.; et al. Nucleic Acids Res 1994, 22, 5456–5465.
2. Shchepinov, M. S.; CaseGreen, S. C.; Southern, E. M. Nucleic Acids Res 1997, 25, 1155–1161.
3. Wang, D. G.; et al. Science 1998, 280, 1077–1082.
4. Hacia, J. G. Nature Genet 1999, 21, 42–47.
5. Roses, A. D. Nature 2000, 405, 857–865.
6. Cutler, D. J.; et al. Genome Res 2001, 11, 1913–1925.
7. Yu, C. J.; et al. J Am Chem Soc 2001, 123, 11155–11161.
8. Guo, Z.; et al. Genome Res 2002, 12, 447–457.
9. Kivi, M.; et al. J Clin Microbiol 2002, 40, 2192–2198.
10. Kolchinsky, A.; Mirzabekov, A. Hum Mutat 2002, 19, 343–360.
11. Lindroos, K.; et al. Nucleic Acids Res 2002, 30.
12. Lu, M. C.; et al. J Am Chem Soc 2002, 124, 7924–7931.
13. Stickney, H. L.; et al. Genome Res 2002, 12, 1929–1934.
14. Straub, T. M.; et al. Appl Environ Microbiol 2002, 68, 1817–1826.
15. Warrington, J. A.; et al. Hum Mutat 2002, 19, 402–409.
16. Vainrub, A.; Pettitt, B. M. Phys Rev E 2002, 66.
17. Vainrub, A.; Pettitt, B. M. J Am Chem Soc 2003, 125, 7798–7799.
18. Vainrub, A.; Pettitt, B. M. Biopolymers 2003, 68, 265–270.

18a. Vainrub, A.; Pettitt, B. M. Chem Phys Lett 2000, 323, 160–166.

19. Wetmur, J. G. DNA Crit Rev Biochem Mol Biol 1991, 26, 227–259.
20. Manning, G. S.; Ray, J. J Biomol Struct Dyn 1998, 16, 461–476.
21. Jin, L.; Horgan, A.; Levicky, R. Langmuir 2003, 19, 6968–6975.
22. Podyminogin, M. A.; Lukhtanov, E. A.; Reed, M. W. Nucleic Acids Res 2001, 29, 5090–5098.
23. Charles, P. T.; et al. Langmuir 2003, 19, 1586–1591.
24. Okahata, Y.; et al. Anal Chem 1998, 70, 1288–1296.
25. Nelson, B. P.; et al. Anal Chem 2001, 73, 1–7.
26. Peterson, A. W.; Wolf, L. K.; Georgiadis, R. M. J Am Chem Soc 2002, 124, 14601–14607.
27. Held, G. A.; Grinstein, G.; Tu, Y. Proc Natl Acad Sci USA 2003, 100, 7575–7580.
28. Hekstra, D.; et al. Nucleic Acids Res 2003, 31, 1962–1968.
29. SantaLucia, J.; Allawi, H. T.; Seneviratne, A. Biochemistry 1996, 35, 3555–3562.
30. Forman, J. E.; et al. ACS Symposium Series 1998, 682, 206–228.
31. Watterson, J. H.; et al. Sens Actuators B 2001, 74, 27–36.
32. Watterson, J. H.; et al. Langmuir 2000, 16, 4984–4992.
33. Walsh, M. K.; Wang, X. W.; Weimer, B. C. J Biochem Biophys Methods 2001, 47, 221–231.
34. Herne, T. M.; Tarlov, M. J. J Am Chem Soc 1997, 119, 8916–8920.
35. Peterson, A. W.; Heaton, R. J.; Georgiadis, R. M. Nucleic Acids Res 2001, 29, 5163–5168.
36. Steel, A. B.; Herne, T. M.; Tarlov, M. J. Anal Chem 1998, 70, 4670–4677.
37. Vainrub, A.; et al. In Biomedical Technology and Devices Handbook; Moore, J.; Zouridakis, G., eds. CRC Press: Boca Raton, FL, 2004; p 14-1–14-14.

Reviewing Editor: Dr. David A. Case