DNA hybridization to mismatched templates: A chip study

Felix Naef,[1] Daniel A. Lim,[2] Nila Patil,[3] and Marcelo Magnasco[1]
[1]Center for Studies in Physics and Biology, Rockefeller University, 1230 York Avenue, New York, New York 10021
[2]Laboratory of Neurogenesis, Rockefeller University, New York, New York 10021
[3]Perlegen Sciences, 2021 Stierlin Court, Mountain View, California 94043

(Received 27 November 2001; published 9 April 2002)

High-density oligonucleotide arrays are among the most rapidly expanding technologies in biology today. In the GeneChip system, the reconstruction of the sample mRNA concentrations depends upon the differential signal generated by hybridizing the RNA to two nearly identical templates: a perfect match probe (PM) containing the exact biological sequence; and a single mismatch (MM) differing from the PM by a single base substitution. It has been observed that a large fraction of MMs repeatably bind targets better than the PMs, against the obvious expectation of sequence specificity. We examine this problem via statistical analysis of a large set of microarray experiments. We classify the probes according to their signal to noise (S/N) ratio, defined as the eccentricity of a (PM,MM) pair's ''trajectory'' across many experiments. Of those probes having large S/N (>3) only a fraction behave consistently with the commonly assumed hybridization model. Our results imply that the physics of DNA hybridization in microarrays is more complex than expected, and suggest estimators for the target RNA concentration.

DOI: 10.1103/PhysRevE.65.040902 PACS number(s): 87.15.−v, 82.39.Pj

Interest in the detailed physics of DNA hybridization is rooted in both purely theoretical and practical reasons. Studies of the denaturing transition started with models of perfectly homogeneous DNA [1], soon followed by studies of sequence-specific disorder [2–4]. The specificity with which DNA binds to its exact complement as opposed to a mismatched copy (a ''defect'') has been studied experimentally [5,6] and theoretically [7–9]. In this context it has been found that a fair fraction of the energetics of DNA hybridization is related to *stacking* interactions between first-neighbor bases, in addition to the obvious strand-strand contact [10–12]. We present a study of mismatch hybridization stemming from a very practical problem, hybridization in DNA microarrays. We shall show experimental evidence that the system behaves inconsistently with current models of hybridization specificities.

DNA microarrays provide an experimental technique for measuring thousands of individual mRNA concentrations present in a given target mixture. They are made by depositing DNA oligonucleotide sequences (probes) at specific locations on solid substrates. The probes can be either premade sequences as in cDNA spotted arrays, or they can be grown *in situ*, letter by letter, as in high-density oligonucleotide arrays [13]. The target mRNA is amplified (into either cDNA or cRNA depending on the protocol) and the product labeled fluorecently before being hybridized onto the array. The spatial distribution of fluorescence is then measured, providing estimates for the RNA concentrations. In GeneChip arrays, the synthesis of probe sequences by photolithography requires a number of different masks per added base, so it is impractical to grow more than a few dozen nucleotides. For such lengths, hybridization specificity is not expected to be high enough. To solve this conundrum, GeneChip technology is based on a twofold approach, involving *redundancy* and *differential signal* [13–15]. First, several different sequence snippets (each 25 bases long) are used to probe a single transcript; and second, each of these probes comes in two flavors. The perfect match (PM) is perfectly complementary to a portion of the target sequence whereas the single mismatch (MM) carries a substitution to the complementary base at its middle (13th) position. MM sequences are expected to probe for nonspecific hybridization as detailed below.

In current incarnations of the chips, each gene is probed by 14–20 (PM,MM) pairs (a probeset), and the task is, therefore, to reconstruct a single number (the RNA concentration) from these 28–40 measurements. There are many ways in which this can be done, with various degrees of noise rejection. The standard algorithm provided in the software suite [16] offers one method. However, as independent measurements of mRNA concentrations showed that the analysis process should be improved upon, many researchers attempted to do so [17,18]; it was then discovered that a fair number of MM probes consistently report higher fluorescence signal than their PM counterpart [18]. This observation is most intriguing because it violates the standard hybridization model outlined below. Thus, the notion that the specific binding signal alone can be obtained as a differential of the PM and MM signals appears to fail in a significant subset of the probes.

We shall show below, by carefully examining the statistics of PM-MM pairs, that it is not a matter of a few stray probes. Our statistics show that *most* of the probes misbehave to various degrees. Given the number of laboratories currently carrying out such experiments, squeezing out even one extra bit of signal to noise ratio from the data would be very valuable. It is clear that this shall not happen in the absence of a better understanding of DNA hybridization to slightly mismatched templates. We shall now attempt the first step toward this goal, which is to characterize the problem.

The rationale behind the use of MM probes follows from the standard hybridization model [17],

$$I_{\mathrm{PM}}=I_S+I_{NS}+B, \tag{1}$$

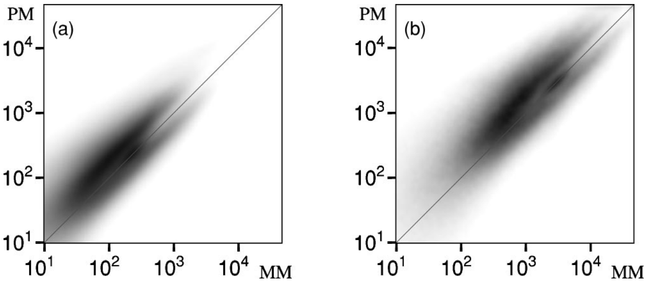

FIG. 1. Joint probability distribution $P(\ln(I_{PM},\ln I_{MM}))$ after background subtraction for (a) 86 HG-U95A human chips, human blood extracts; (b) 24 Mu11KsubA chips, mouse brain extracts. Three features are present in both: the probability cloud forks into two lobes at high intensity, and an intense "button" lies between the two forks right in the middle of the range. The lower lobe *fully* lies below the diagonal.

$$I_{\mathrm{MM}}=(1-\alpha)I_S+I_{NS}+B, \tag{2}$$

$$I_{\mathrm{PM}}-I_{\mathrm{MM}}=\alpha I_S\,. \tag{3}$$

Here I_{PM} (I_{MM}) are the measured brightness of the PM (MM) probe, I_S the contribution from specific complementary binding, I_{NS} the amount from nonspecific binding assumed to be insensitive to the substitution, and B a background of physical origin, i.e., the photodetector dark current or light reflections from the scanning process. Then α is the reduction of specific binding due to the single mismatch. These brightnesses are related to the quantity of interest (the RNA concentration in the sample) through

$$I_S=k\,[\mathrm{RNA}]_{\mathrm{spec}},$$

$$I_{NS}=h\,[\mathrm{RNA}]_{\mathrm{nonspec}},$$

where $[\mathrm{RNA}]_{\mathrm{spec}}$ denotes the concentration of target RNA, $[\mathrm{RNA}]_{\mathrm{nonspec}}$ the concentration of RNAs contributing to nonspecific hybridization. k and h are probe dependent specific and nonspecific susceptibilities and include effects such as the areal density of probe, various affinities, transcript length dependent effects (longer transcripts are likely to carry more fluorophors depending on the labeling technique).

While obviously the physics of hybridization is much more complex than this simplistic model, one could still hope that it would provide an essentially correct picture of GeneChip hybridizations. Let us summarize the basic assumptions made so far: (i) nonspecific binding is identical in PM and MM, so I_{NS} does not see the letter change; (ii) $\alpha>0$; (iii) k and h identical for PM and MM; (iv) k, h, and α are reasonably uniform numbers across a probeset.

So from these assumptions it follows that PM>MM for all probe pairs. But experimentally one observes a vast number of probe pairs violating this assumption consistently for a broad range of conditions. In our experience, most people in the know think of this problem as an imperfect adherence to the standard model. In other words, this problem is usually characterized as "there is a few probe pairs that do not work and we do not understand why." We shall show that this is not so: the MM>PM pairs are so abundant that we want to propose an alternate view: the model is simply inadequate for describing the hybridization process, and we do not understand the basic physics of MM hybridization.

TABLE I. Statistics of probe pairs with MM>PM taken across a large GeneChip data collection. "%PS with >1" means "percent of probesets with more than one MM>PM pair." The yeast chip (last column) is *noticeably* different and better behaved than the other cases.

Chip	Dros	HG-U95A	Mu11K	U74A	YG_S98
No. of pairs per PS	14	16	20	16	16
Chips analyzed	36	86	24	12	4
% MM>PM	35	31	34	34	17
% PS with>1	95	91	95	92	73
% PS with>5	58	56	71	64	21
% PS with>10	4	7	26	10	2

The human HG-U95A chip series, for instance, has 400 K probes for 12 K different probesets. Across a wide variety of conditions, we have observed approximately 30% of all probe pairs have MM>PM. This huge figure could be dismissed if most of them were in the low-intensity range, where the noise is relatively higher, or if they were clustered in a small set of problematic probesets. Neither is true: 91% of all probesets have at least 1 MM>PM probe pair, and 60% of probesets have five such probe pairs out of 16. In addition, the MM>PM pairs are fairly distributed with respect to brightness (cf. Fig. 1). Table I summarizes the statistics for various chip series.

What could give rise to those MM>PM? A perplexing extra bit of information lies in a simple statistic, the joint probability distribution $P(\ln I_{PM},\ln I_{MM})$. According to the standard model,

$$\frac{I_{PM}}{I_{MM}}=\frac{I_S+I_{NS}+B}{(1-\alpha)I_S+I_{NS}+B}.$$

So if $(1-\alpha)I_s \gg I_{NS}+B$ then $I_{PM}/I_{MM}\rightarrow 1/(1-\alpha)$, while if I_S vanishes (e.g., the transcript is not there) then $I_{PM}/I_{MM}\rightarrow 1$. Thus we expect

$$1\leqslant\frac{I_{PM}}{I_{MM}}\leqslant\frac{1}{1-\alpha}. \tag{4}$$

So the standard model predicts that $P(\ln I_{PM},\ln I_{MM})$ should lie in a band, with an upper limit given by $I_{MM}=(1-\alpha)I_{PM}$ for fully specific binding, and with lower limit in the diagonal PM=MM when cross hybridization dominates. Naively one would further assume that for low brightness most of the signal comes from nonspecific binding, while most would come from specific binding for high brightness. Figure 1 shows something quite otherwise: as brightness increases, the joint probability distribution forks into two branches. The crest of the lower one lies fully below the MM=PM diagonal.

The characteristic shapes of $P(\ln I_{PM},\ln I_{MM})$ are likely signatures of sequence-dependent effects. However, any hypothesis is impossible to verify as the probe sequences are not released to the public. Nevertheless, there are some obvious suspects. First, the nontrivial susceptibilities k and h mentioned above depend on the areal density of probe, which is sequence dependent by virtue of the varying efficiencies of the lithography process. Second, nucleic acids need to unstack the single-stranded probes in order to form each new duplex as they hybridize. Further, stacking energies are extremely sensitive to sequence details, which might result in large energy barriers. This would translate into kinetics constants varying exponentially (following Arrhenius' law in these energies, and lead to important consequences as the hybridization reactions are not carried to full thermodynamic equilibrium in the standard Affymetrix hybridization protocol, since the signal still increases if the hybridization is extended.

Given a single probe pair measured in N experiments with possibly different mRNA concentrations, further insight can be gained by following the trajectory of that pair $\vec{P}^i = (\ln I^i_{PM}, \ln I^i_{MM})$, with $i = 1, \ldots, N$ (after subtracting B). Ideally, these points would fall on a curve parametrizable by the mRNA concentration. In reality, however, the observed patterns range from nearly one dimensional to almost circular clouds. To classify probe pairs, we computed the center of mass (c.m.) and inertia tensor $\mathcal{I}$ of the set of points $\{\vec{P}_i\}$. The positive eigenvalues of $\mathcal{I}$, $I_1 \geqslant I_2$ define the eccentricity $e = \sqrt{I_1/I_2}$ and largest excursion $\lambda_1 = \sqrt{I_1}$. Highest eccentricities characterize probe pairs with largest S/N, whereas $e \sim 1$ would be typical for very noisy pairs, or pairs that did not move in the considered dataset because the mRNA concentration for that particular transcript was roughly constant across the N experiments.

The resulting distribution of centers of mass is shown in Fig. 2. Even though each point in Fig. 2(a) is an average over 86 points of Fig. 1(a), still Fig. 2(a) looks very similar to Fig. 1(a), proving that most probes behave *reproducibly*. For instance, a probe pair lying below the PM$=$MM diagonal in one experiment stays so in most of the 86 experiments; thus its c.m. stays below the diagonal too. Selecting for $e > 3$ eliminates most of the low-intensity probes pairs [Fig. 2(b)], and the remaining set contains two components: one consisting of the large λ_1 pairs [Fig. 2(c)] lying mostly in the PM $>$MM region; while the small λ_2 component forming an almost perfectly symmetric ''tulip'' structure [Fig. 2(d)], containing two forked branches plus the button mentioned in Fig. 1. Notice that only the probe pairs of Fig. 2(c) behave as we discussed following Eq. (4).

We have so far discussed the characteristics of single probe pairs, without grouping them into their respective probesets. Turning to properties of entire probesets, a feature that deeply affects attempts at analysis is the very broad brightness distributions within probes belonging to the same probeset (Fig. 3). Possible reasons for such behavior are sequence specific effects similar to those discussed in the context of the MM behavior.

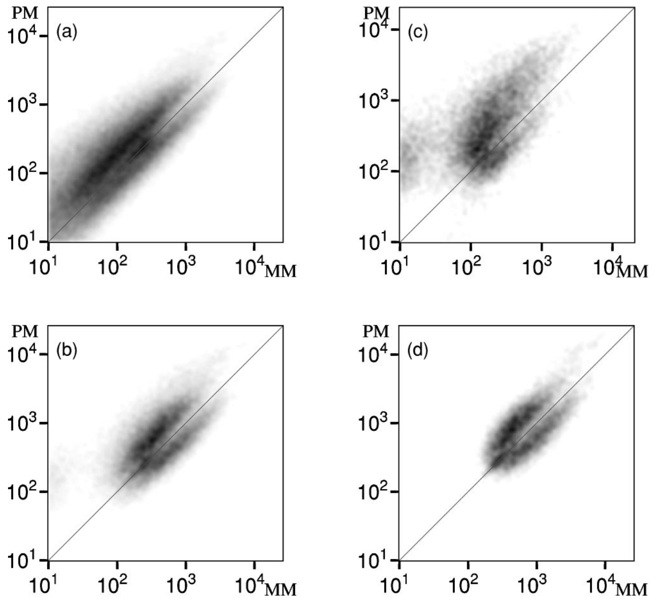

FIG. 2. Histogram of probe pair center of mass [same data as in Fig. 1(a)]. (a) All probe pairs. (b) Only those probe pairs with eccentricities $e > 3$. (c) The probe pairs of (b), further restricted to large excursions [$\lambda_1 > 0.133$, the top third of (b)]. (d) Same as (c) for small excursions ($\lambda_1 < 0.108$, the bottom third). Notice that (c) consists of all probe pairs with large S/N and large signal, while (d) consists of pairs having large S/N but small signal.

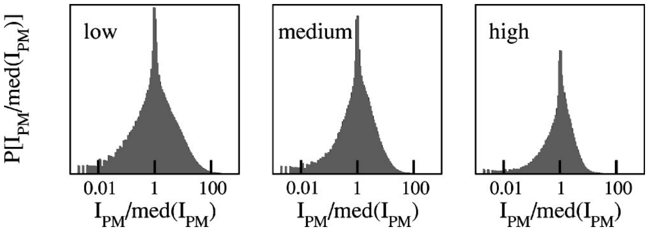

FIG. 3. Relative PM intensity distributions within probesets (after subtracting B). The data are identical to those in Fig. 1(a). Probesets are split into three groups of equal size according to their median PM intensity. In all cases, the distributions of $I_{\mathrm{PM}}/\mathrm{median}(I_{\mathrm{PM}})$ span up to four decades. Notice the signs of saturation in the right tail of the high-intensity panel.

Although we have so far pointed out caveats of the current hybridization model, our primary interest in the topic is to suggest *improvements* for the reconstruction of the sample mRNA concentration from the probeset data. Because the variability in the hybridization properties of the probes is larger than naively anticipated, it is unlikely that a single definitive procedure will be appropriate in all cases. For instance, the differentials PM-MM will not consistently be a good estimator of the true signal [18]. Given the unclear information contained in the MM, one alternative we studied is not using them at all as nonspecific controls. The mRNA expression level is then obtained from an ''outlier robust'' geometric average of the background subtracted PM values $(I_{\mathrm{PM}} - B)$, after a careful estimation of B [18]. The use of geometric averages is suggested by the shape of the distributions in Fig. 3, which are nearly symmetric in logarithmic coordinates. Arithmetically averaging numbers so distributed would result in the estimator being dominated by the largest measurements, and there would be no reduction of the noise level with the number of data points being used. In other words, the redundancy provided by having 14–20 probe pairs for each gene would not be exploited to improve the quality of the measurement. Of course, using only PM probes neglects cross-hybridization effects that would be detectable by a working MM probe, and hence tends to be less sensitive at the low-intensity end. One the other hand, it allows to rescue probesets with a high number of misbehaving MMs.

The introduced trajectories suggest a different approach for deriving expression levels from GeneChips, by extending the ellipsoid of inertia idea to the full probeset. The resulting method is close in spirit to model-based methods [17], which attempts to determine the susceptibilities α in Eq. (3) by a least-square fitting procedure of the differentials PM-MM in linear coordinates. Here, all probes (PM and MM) are used on an equal footing, and the intensities are log transformed. Concretely, one would consider the principal components of the matrix $A^{ij}=(\ln I^{ij}_{PM}, \ln I^{ij}_{MM})$ $(j=1,\ldots,N_p$ is the probe and i the experiment index) to identify the modes carrying the most signal. After singular value decomposition $\hat{A}=U\Lambda V^{T}$, where $\hat{A}^{ij}=A^{ij}-m^{j}$ and $m^{j}=(1/N)\Sigma_{i}A^{ij}$ is the center of mass, the signal $s_i=\Sigma_j(m_j+\hat{A}^{ij})V^{1}_{j}$ is given by projecting onto the largest direction of variation. A signal-to-noise measure for the entire probeset can be obtained from $S/N=\lambda_1/\sqrt{\Sigma_{j=2}^{N_p}\lambda_j^2}$. Preliminary testing of the method has lead to very promising results, especially in the high-intensity regime [19].

We showed that the hybridization of short length DNA sequences to single mismatched templates exhibits a far more complicated picture than what is usually assumed. Our observations do not only point at interesting physics in the DNA hybridization process to short sequences with defects, attached to a glass surface; they also have strong consequences for designers of GeneChip analysis tools. We hope this will bolster interest in the physics of hybridization and mismatch characterization, and eventually help improve current microarray designs.

Note added. Affymetrix Corporation has now released the probe sequences. Quick inspection shows the branches of Fig. 1 to correspond to whether the middle nucleotide is a pyrimidine (top) or purine (bottom). This suggests a role for the biotinilated bases uracil and cytosine, which are used in the fluorescent labeling of the RNA.

We would like to thank S. Bekiranov, G. Bonnet, E. van Nimwegen, and E. Siggia for sharing interesting ideas. We acknowledge support of the Meyer Foundation (MM) and the Swiss National Science Foundation.

[1] M. Peyrard and A.R. Bishop, Phys. Rev. Lett. **62**, 2755 (1989).
[2] D. Cule and T. Hwa, Phys. Rev. Lett. **79**, 2375 (1997).
[3] Y. Zhang, W.-M. Zheng, J.-X. Liu, and Y.Z. Chen, Phys. Rev. E **56**, 7100 (1997).
[4] D.K. Lubensky and D.R. Nelson, Phys. Rev. Lett. **85**, 1572 (2000).
[5] A. Bonincontro, M. Matzeu, F. Mazzei, A. Minoprio, and F. Pedone, Biochim. Biophys. Acta **1171**, 288 (1993).
[6] G. Bonnet, S. Tyagi, A. Libchaber, and F.R. Kramer, Proc. Natl. Acad. Sci. U.S.A. **96**, 6171 (1999).
[7] M. Salerno, Phys. Rev. A **44**, 5292 (1991).
[8] N. Singh and Y. Singh, Phys. Rev. E **64**, 042901 (2001).
[9] J.A.D. Wattis, S.A. Harris, C.R. Grindon, and C.A. Laughton, Phys. Rev. E **63**, 061903 (2001).
[10] G. Vesnaver and K.J. Breslauer, Proc. Natl. Acad. Sci. U.S.A. **88**, 3569 (1991).
[11] N.L. Goddard, G. Bonnet, O. Krichevsky, and A. Libchaber, Phys. Rev. Lett. **85**, 2400 (2000).
[12] Z. Haijun, Z. Yang, and O.-Y. Zhong-can, Phys. Rev. Lett. **82**, 4560 (1999).
[13] J. Lockhart and E.A. Winzeler, Nature (London) **405**, 827 (2000).
[14] M. Chee *et al.*, Science **274**, 610 (1996).
[15] R.J. Lipshutz, S.P. Fodor, T.R. Gingeras, and D.J. Lockhart, Nat. Genet. **21**, 20 (1999).
[16] Affymetrix Microarray Suite 4.0 User Guide (2000), Affymetrix, Inc., Santa Clara, CA.
[17] C. Li and W.H. Wong, Proc. Natl. Acad. Sci. U.S.A. **98**, 31 (2001).
[18] F. Naef, D. A. Lim, N. Patil, and M. Magnasco, Proceedings of the DIMACS Workshop on Analysis of Gene Expression Data 2001 (unpublished); also e-print physics/0102010.
[19] F. Naef, N. Socci, and M. Magnasco (unpublished).