Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays

Felix Naef and Marcelo O. Magnasco*
Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

(Received 6 November 2002; revised manuscript received 24 March 2003; published 16 July 2003)

RNA binding to high-density oligonucleotide arrays has shown tantalizing differences with solution experiments. We analyze here its sequence specificity, fitting binding affinities to sequence composition in large datasets. Our results suggest that the fluorescent labels interfere with binding, causing a catch-22. To be detected, the RNA must both glow and bind: without labels it cannot be seen even if bound, while with too many it will not bind. A simple model for the binding of labeled oligonucleotides sheds light on the interplay between binding energies and labeling probability.

DOI: 10.1103/PhysRevE.68.011906 PACS number(s): 87.15.−v, 82.39.Pj

Hybridization-based DNA microarrays have recently been developed for large-scale measurements of messenger RNA (mRNA) transcript abundance in biological systems [1–3]. Such DNA arrays permit the measurement of thousands of mRNA species simultaneously, providing a *global* snapshot of transcriptional activity in a given cellular state. Although they are mainly used as genetic screening devices, they hold the promise to unravel some aspects of the tangled web of transcriptional controls [4,5]. However, it has been argued [6–8] that progress in this expanding technology needs a better understanding of the system's basic hybridization physics. Previous studies in oligonucleotide hybridization have relevance for microarrays; in particular, experimental and theoretical work has investigated the binding specificities of exactly complimentary strands versus strands with a number of mismatches or defects [9–13]. However, the geometrical constraints of surface hybridization and the use of labeled nucleotide add array-specific particularities that require deeper study.

Such studies are primarily motivated by practical considerations. Array-hybridization signal, being the result of a trade off of quality for quantity, is intrinsically imperfect, and analysis algorithms need to achieve high levels of noise rejection. The term ''noise'' refers to a complex superposition of effects ranging from fluorescence background, nonspecific and concentration dependent hybridization from competing RNA species in the mixture, to systematic contributions related to the probe sequences and labeled nucleotides. Our purpose is to focus on the latter aspects of hybridization in high-density oligonucleotide arrays [HDONAs, a.k.a. GeneChip(r)].

HDONA probes consist of 25-bases oligonucleotides (25-mers) grown photolithographically onto a glass surface, and at current densities, about a million different such probes can be synthesized on each array. Because 25-mers can exhibit considerable cross hybridization to a complex background, the system was designed on two layers. First, a ''differential signal'' approach performs the first level of rejection of spurious signal by computing the difference between the brightness of a perfect-match (PM) probe complimentary to a 25-mer in the mRNA sequence, and a single-mismatch (MM) probe in which the middle nucleotide has been changed to its complement. Second, redundancy is introduced by using between 10 and 20 probe pairs per transcript, corresponding to distinct 25-mers along the length of the transcript, as shown in Fig. 1. The full set of probes for one transcript is called a probeset. At the lowest level, analysis must translate the patterns of light and dark recorded by a laser beam into the best possible estimate of the specific mRNA concentration [14,15]. Any inaccuracies introduced at that level (i.e., loss of signal or false positive assignments) cannot be recovered from thereafter.

From the thermodynamics of DNA-RNA hybrids in solution [16], it was expected that the PM probe should have a higher affinity for the specific target than the MM probe, while cross hybridization should be roughly equal for both. But these ideas do not translate that easily from hybridization in solution to HDONAs. An issue long noticed was the large number of probe pairs for which the single mismatch brightness was higher than the perfect match, up to a third of all probe pairs in some chip models [7]. A two-dimensional histogram of PMs versus corresponding MMs shows a joint probability distribution with two branches, and so it was suggested that sequence specific effects are playing a crucial role [7]. However, this could not be verified in the absence of sequence information. Now that this information is available [17], we can address the problem explicitly.

We show in Fig. 2 joint probability distributions of PMs and MMs, obtained from all probe pairs in a large set of experiments. Actually, two separate probability distributions are superimposed: in red, the distribution for all probe pairs whose 13th letter is a purine, and in cyan those whose 13th letter is a pyrimidine. The plot clearly shows two distinct branches in two colors, corresponding to the basic distinction between the shapes of the bases: purines are large, double ringed nucleotides while pyrimidines have smaller single rings. This underscores that by replacing the middle letter of the PM with its complementary base, the situation on the

*Present address: The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, Trieste I-34100, Italy.

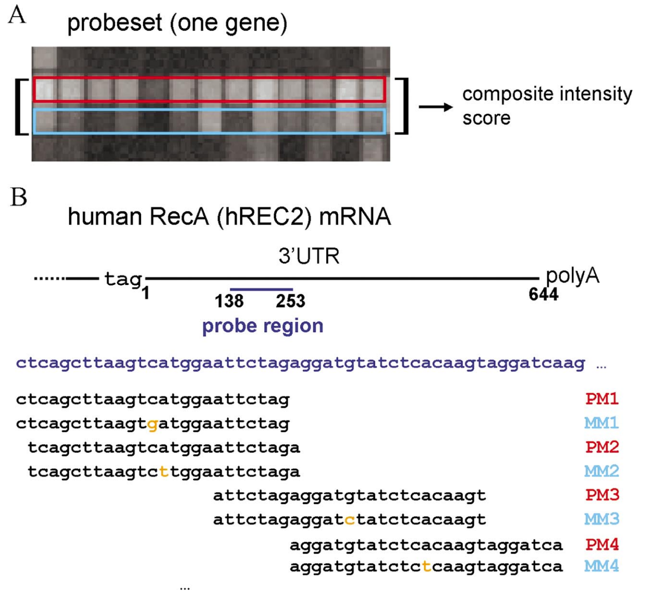
A
probeset (one gene)
composite intensity score
B
human RecA (hREC2) mRNA
3'UTR
tag
1
138
253
644
polyA
probe region
ctcagcttaagtcatggaattctagaggatgtatctcacaagtaggatcaag ...
ctcagcttaagtcatggaattctag PM1
ctcagcttaagtgatggaattctag MM1
tcagcttaagtcatggaattctaga PM2
tcagcttaagtcttggaattctaga MM2
attctagaggatgtatctcacaagt PM3
attctagaggatctatctcacaagt MM3
aggatgtatctcacaagtaggatca PM4
aggatgtatctctcaagtaggatca MM4
...

FIG. 1. (Color) Probeset design. (A) The raw scanned image of a typical probeset, with the PM (MM) on the top (bottom) row; higher brightness (white) corresponds to higher abundance of bound RNA molecules. The large variability in probe brightness is clearly visible. (B) Arrangement of probe sequences along the target transcript for the human recA gene in the HG-U95A array. Here the probe region (blue) is 116 bases long; it is typical that probes lie in the 3' UnTRanslated region, namely, between the stop triplet (codon) "tag" and the polyadenylation signal. The first four probes are shown explicitly; notice the overlap in their sequences.

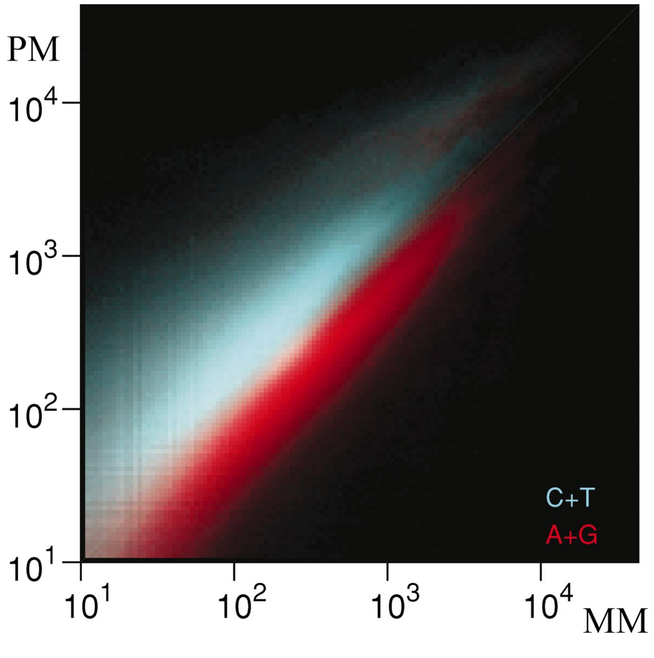

FIG. 2. (Color) PM vs MM histogram from 86 human HG-U95A arrays. The joint probability distribution for PM and MM shows strong sequence specificity. In this diagram, all 17×10^6 (PM,MM) pairs in a dataset were used to construct a two-dimensional histogram. Pairs whose PM middle letter is a pyrimidine (C or T) are shown in cyan, and purines (A or G) in red. 33% of all probe pairs are below the PM=MM diagonal; 95% of these have a purine as their middle letter.

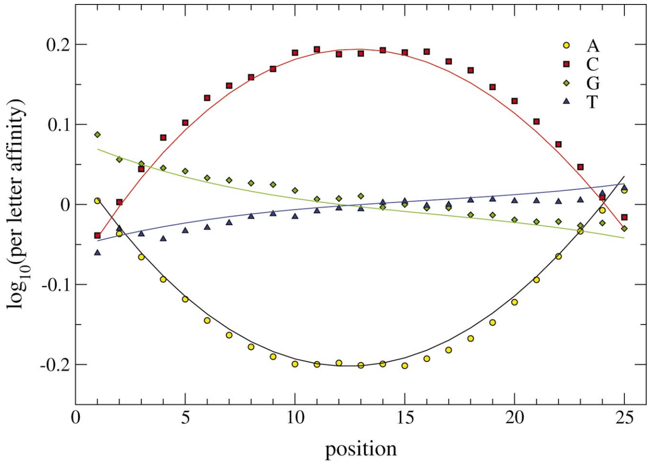

FIG. 3. (Color) Sequence specificity of brightness in the PM probes. PM probes from the same data as in Fig. 2 were fit (multiple linear regression) to the probe sequence composition. The resulting site-specific affinities A_{li} are shown as dots; position 1 corresponds to the first base on the glass side. The spatial smoothness of the A_{li} permits the use of $A_{l\alpha}$ as fitting variables, thereby reducing the number of parameters. The solid lines show the position dependence obtained from a cubic expansion ($\alpha=0,\ldots,3$). 13 (four parameters×three independent letters + offset) variables were fit to 17×10^6 data points, with the following statistics: $r^2=0.44$, $F=1071045$, and $p<10^{-16}$. In our data, the variance in brightness in 96% of all probesets is reduced after the predicted sequence-specific part is subtracted, and the reduction is larger than a factor of 2 for 65% of the probesets.

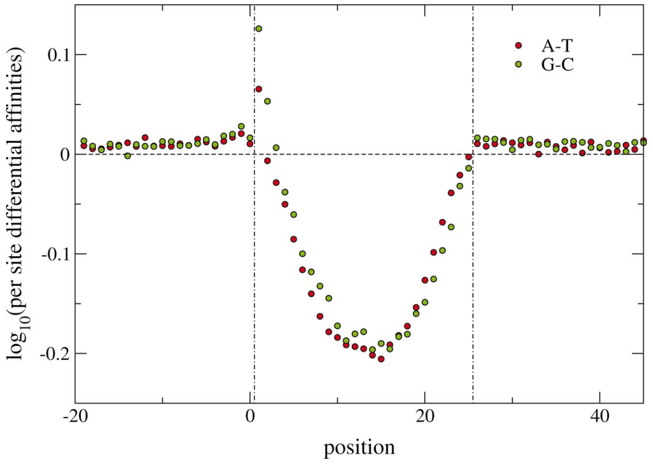

FIG. 4. (Color) Reduction in brightness due to labeled U's and C's. Here fits have been extended to also include sequence information from 20 flanking bases on each end of the probe. The asymmetry of (A,T) and (G,C) affinities in Fig. 3 can be explained because only A-U and G-C bonds carry labels (purines U and C on the mRNA are labeled). Notice the nearly equal magnitudes of the reduction in both type of bonds. Additionally, one can observe the change in sign at the boundaries of the probes, reflecting the fact that carrying labels outside the probe region tends to contribute positively to the brightness, while carrying labels inside the probe region is unfavorable because labels interfere with binding.

MM probe is that the middle letter always faces itself, leading to two quite distinct outcomes according to the size of the nucleotide. If the letter is a purine, there is no room within an undistorted backbone for two large bases, so this mismatch distorts the geometry of the double helix, incurring a large steric and stacking cost. But if the letter is a pyrimidine, there is room to spare, and the bases just dangle. The only energy lost is that of the hydrogen bonds. So the existence of two branches agrees with basic hybridization physics, but it still does not explain why the MMs are actually brighter than the PMs in many sequences with a purine middle letter.

To understand this we concentrate momentarily only on the PM sequences. It has been pointed out that the PMs within a probeset are very broadly distributed, typically spanning two decades or more. We can try to determine whether this breadth is similarly sequence dependent by fitting brightness B of PM probes (divided by a surrogate for the RNA concentration: the median of the PM brightnesses) against their own sequence composition:

$$\ln\left(\frac{B}{[RNA]}\right)=\sum_{li} S_{li}A_{li}=\sum_{l\alpha} S_{l\alpha}A_{l\alpha}, \tag{1}$$

where $l=A,C,G,T$ is the letter index and $i=1,\ldots,25$ the position along the 25-mer; S is a Boolean variable equal to 1 if the probe sequence has letter l at site i and 0 otherwise, and thus A_{li}'s are per-site, per-letter affinities. Note that $\Sigma_l S_l=1$ for all i, so that the addition of an intercept in Eq. (1) can be absorbed in a redefinition of the A's. The last equality uses an expansion of the spatial dependence in orthonormal polynomials $P_{i\alpha}$ on interval [1,25], so that $A_{l\alpha}=\Sigma_i A_{li}P_{i\alpha}$.

Finer models would include stacking energies involving adjacent letters (nearest-neighbor interactions along the transcript length); while this contribution is important for hybridization experiments in solution [18,19], we found that it does not improve our fit enough to justify the increase in number of parameters. On the other hand, we were surprised to discover that the major improvement comes from introducing position-dependent affinities, as opposed to affinities depending only on the total number of occurrences of each letter. The fitted per-site affinities are shown in Fig. 3. Note the strength of letter-specific contributions: changing an A to a C in the middle of the sequence changes the brightness of the probe by 250%. Notice the prominent edge effects, indicating breathing of the duplex. The left-right asymmetry could be due to both attachment to the glass and fabrication efficiency effects, e.g., premature termination. Performing identical fits on mouse, drosophila, and yeast arrays lead to affinities virtually identical to those shown in Fig. 3. An unexpected aspect of the above fits is the asymmetry of A versus T (and G versus C) affinities, which goes against the zeroth order energetic consideration that A-T and T-A bonds (or G-C and C-G) would contribute equally to the binding. The asymmetry is shown clearly in Fig. 4.

The obvious culprits for this effect are *the fluorescent labels*. The standard recommended protocol entails labeling the amplified mRNA with biotinilated nucleotides, more specifically, U and C, the pyrimidines. This suggests a rather simple explanation, namely, that the biotinilated bases somehow impede the binding; the effect diminishing to zero toward the probe edges, where the double strand breathes enough to be able to accommodate the linkers, and being maximal near the center, where the largest disruption would be effected, where the largest disruption would occur. This would cause a catch-22 in terms of obtaining the maximal fluorescence: if a sequence has too few bases that can be labeled, it will not shine even if it binds strongly, while if it has too many labels it will not shine because it does not bind.

To understand how these labels interfere with brightness, and to shed light on the dependence on the physical parameters, we introduce a simplified model. Consider a given RNA transcript and let $N=\Sigma_i S_{Ai}+S_{Gi}$ be the number of potentially labeled sites. If p is the probability that such a site be labeled (in the standard protocol, $p\sim 1/4$), then the fraction of the RNA molecules carrying n labels is given by binomial distribution $B_p(n,N)=\binom{N}{n}p^n(1-p)^{N-n}$. Typically, $N\sim 12$ for 25-bases probes. The binding energy of a hybridization duplex with $n=0,\ldots,N$ labels can be approximated as $E_B(n)=E_B^0+nE_L$, where E_B^0 is the bare binding energy and E_L is a penalty for each label. Assuming an ideal solution, the chemical potentials are given as $\mu(c_n)=\beta\ln(c_n/c_0)$, with $c_n=[\mathrm{RNA}]B_p(n,N)$ and $\beta=1/k_BT$. Then, in the limit of low coverage, $\Sigma_{n=0}^{N}e^{-\beta(E_B(n)-\mu(c_n))}\ll 1$, the average number of labeled nucleotides per probe $\langle n\rangle\in[0,N]$ reads

$$\langle n\rangle=\frac{[\mathrm{RNA}]}{c_0}\sum_{n=0}^{N} n e^{-\beta[E_B(n)-\mu(c_n)]} \tag{2}$$

$$=\frac{[\mathrm{RNA}]}{c_0}e^{-\beta E_B^0}\frac{\partial}{\partial(-\beta E_L)}[pe^{-\beta E_L}+(1-p)]^N. \tag{3}$$

The connection between the affinities in Eq. (1) and the physical parameters introduced above is obtained via

$$\ln\left(\frac{B}{[\mathrm{RNA}]}\right)=\ln(\langle n\rangle)+\ln(\Lambda) \tag{4}$$

$$=\ln(\Lambda)-\beta E_B^0+(N-1)\ln[pe^{-\beta E_L}+(1-p)] + \ln(Npe^{-\beta E_L}), \tag{5}$$

where Λ is the constant relating the number of fluorophores to the observed reduced brightness [Eq. (1)]. Two limits help shed some light on the intricate interplay between energetic costs and labeling probability. The easiest case is $p\rightarrow 1$,

$$\ln(\langle n\rangle)=-\beta(E_B^0+NE_L)+\ln(N)+\ln\left(\frac{[\mathrm{RNA}]}{c_0}\right), \tag{6}$$

in which all sites are labeled and the maximum labeling penalty has to be paid. We can also investigate limit $p\rightarrow 0$, keeping pN finite. Then,

$$\ln(\langle n\rangle)=-\beta(E_B^0+E_L)+\ln(pN)+pN(e^{-\beta E_L}-1)+\ln\left(\frac{[\mathrm{RNA}]}{c_0}\right). \tag{7}$$

Here, the first term indicates that only RNA molecules with a single label contribute to brightness. In the real situation, $p\sim 1/4$ and $pN\sim 3$, indicating that two-label corrections may be needed for more accuracy. Still, the large value of E_L restricts the largest contribution to single-label contributions, thus justifying a linear form of the fit in Eq. (1).

But this catch-22 has a curious loophole. The optimal region to have the fluorophores should then be outside the 25-mer, since the RNA fragment being hybridized is usually longer than the 25-mer it is binding to. Figure 4 confirms this: when including the contribution to brightness from sequence composition outside the 25-mer, we find the purine contribution to be strictly positive, while negative inside the binding region.

Interference with binding by the biotinilated bases also suggests a solution to the MM>PM riddle. As we mentioned, a purine in the middle of the PM probe implies a gap between the two nucleotides on the MM probe; thus one could conjecture that this gap permits the linker between nucleotide and biotin not to interfere with the binding. If this were so, when considering the effective contribution of a middle bond to brightness, a *G*-*C** bond on the PM probe should be dimmer than a *C*-*C** bond on the MM, which, in turn, should be dimmer than a *C*-*G* bond on the PM (where * denotes a labeled nucleotide on the RNA strand). This conjecture is quantitatively compatible with the data: according to Fig. 4, the energetic penalty for a label in the middle of the sequence is 0.2 in $\log_{10}$ units (an estimate for the *G*-*C** to *C*-*G* loss), which should be comparable (but not smaller than) the median excess brightness of the MMs in the purine (red) lobe of Fig. 2, which we measure to be about 0.1.

We have shown how the vast amount of data from hybridization experiments can be used to further our understanding of the physics of the measurement device itself. In addition to providing insight into position- and label-dependence of the binding, the predicted affinities also bear practical value as they permit to effectively reduce the variability in the probe intensities within a probeset (cf. Fig. 4). Consequently, averaging the redundant probes will lead to lower noise levels in absolute concentration estimates. While it is usually emphasized that high-throughput techniques, such as microarrays, pose analytical challenges in terms of global biological interpretation, our work exemplifies that to reach a level where analysis can be abstracted to such heights, one should first understand in some detail the physics of the instrument and how it affects the raw data.

We thank Herman Wijnen, Edward Yang, Nila Patil, Coleen Hacker, Adam Claridge-Chang, and Doeke Hekstra for helpful discussions.

[1] D.J. Lockhart and E.A. Winzeler, Nature (London) **405**, 827 (2000).

[2] M. Chee *et al.*, Science **274**, 610 (1996).

[3] R.J. Lipshutz *et al.*, Nat. Genet. **21**, 20 (1999).

[4] D.K. Gifford, Science **293**, 2049 (2001).

[5] N. Banerjee and M.Q. Zhang, Curr. Microbiol. **5**, 313 (2002).

[6] H. Dai, M. Meyer, S. Stepaniants, M. Ziman, and R. Stoughton, Nucleic Acids Res. **30**, E86 (2002).

[7] F. Naef *et al.*, Phys. Rev. E **65**, 040902 (2002).

[8] D. Hekstra, A.R. Taussig, M. Magnasco, and F. Naef, Nucleic Acids Res. **31**(7) 1962 (2003).

[9] A. Bonincontro, M. Matzeu, F. Mazzei, A. Minoprio, and F. Pedone, Biochim. Biophys. Acta **1171**, 288 (1993).

[10] G. Bonnet, S. Tyagi, A. Libchaber, and F.R. Kramer, Proc. Natl. Acad. Sci. U.S.A. **96**, 6171 (1999).

[11] M. Salerno, Phys. Rev. A **44**, 5292 (1991).

[12] N. Singh and Y. Singh, Phys. Rev. E **64**, 042901 (2001).

[13] J.A.D. Wattis, S.A. Harris, C.R. Grindon and C.A. Laughton, Phys. Rev. E **63**, 061903 (2001).

[14] F. Naef, C. Hacker, N. Patil, and M. Magnasco, Genome Biol **3**, 18 (2002).

[15] C. Li and W.H. Wong, Proc. Natl. Acad. Sci. U.S.A. **98**, 31 (2001).

[16] N. Sugimoto *et al.*, Biochemistry **34**, 11 211 (1995).

[17] Available from http://www.affymetrix.com/analysis/download_center.affx

[18] G. Vesnaver and K.J. Breslauer, Proc. Natl. Acad. Sci. U.S.A. **88**, 3569 (1991).

[19] N.L. Goddard *et al.*, Phys. Rev. Lett. **85**, 2400 (2000).