# Thermodynamics of RNA/DNA hybridization in high density oligonucleotide microarrays

Enrico Carlon, Thomas Heim

# Thermodynamics of RNA/DNA hybridization in high density oligonucleotide microarrays

Enrico Carlon and Thomas Heim
*Interdisciplinary Research Institute c/o IEMN, Cité Scientifique BP 60069, F-59652 Villeneuve d'Ascq, France*
(Dated: October 23, 2018)

We analyze a series of publicly available controlled experiments (Latin square) on Affymetrix high density oligonucleotide microarrays using a simple physical model of the hybridization process. We plot for each gene the signal intensity versus the hybridization free energy of RNA/DNA duplexes in solution, for perfect matching and mismatching probes. Both values tend to align on a single master curve in good agreement with Langmuir adsorption theory, provided one takes into account the decrease of the effective target concentration due to target-target hybridization in solution. We give an example of a deviation from the expected thermodynamical behavior for the probe set 1091_at due to annotation problems, i.e. the surface-bound probe is not the exact complement of the target RNA sequence, because of errors present in public databases at the time when the array was designed. We show that the parametrization of the experimental data with RNA/DNA free energy improves the quality of the fits and enhances the stability of the fitting parameters compared to previous studies.

PACS numbers: 87.15.-v,82.39.Pj

## I. INTRODUCTION

DNA microarrays [1, 2] are devices capable of measuring the gene expression levels on a genome-wide scale and are based on the *hybridization* between surface-bound DNA sequences (the probes) and DNA, or RNA, sequences in solution (the targets). While the specificity of the interaction between complementary base pairs A–T and C–G suggests that the hybridization of a single stranded DNA target with its perfect matching probe would be dominant, often strong non-complementary hybridization effects are observed (see Figure 1). As the targets are fluorescently labeled, the amount of hybridized DNA from each probe can be determined from optical measurements. The presence of strong cross-hybridizations is one of the reasons why one cannot interpret the fluorescent light intensities as direct measures of the gene expression levels.

One of the most popular platform for DNA microarrays is provided by Affymetrix [2], which produces high-density oligonucleotide arrays. In Affymetrix chips short single stranded DNA sequences (25 nucleotides) are grown *in situ* using photolitographic techniques. As a single probe of just 25 nucleotides may not provide enough specificity for a reliable measurement of the gene expression level, a set of 10-16 probes (the probe set) complementary to different regions of the same target sequence are present in the chip. For each perfect matching (PM) probe there is a sequence differing by a single nucleotide. These are referred to as mismatching (MM) probes and are used to quantify the effects of cross-hybridization [2].

Most of the available software packages for the calculation of gene expression levels from the fluorescence intensities rely on algorithms of purely statistical or empirical nature [3, 4]. In the past two years, however, several algorithms based on physical properties of the hybridization process were proposed [5, 6, 7, 8, 9]. The basic idea behind the latter approach is that the intensities are linked to the hybridization free energies for the formation of the probe-target duplexes. For instance, for equal target concentration in solution, binding to CG-rich probes will provide a stronger signal compared to CG-poor probes (CG nucleotides are more strongly bound than AT pairs [10]).

In this paper we investigate a set of controlled experiments known as Latin square experiments and performed by Affymetrix in the human HGU95a chipset. In these experiments some target sequences are added at controlled ("spike-in") concentrations on a background reference solution. The target concentrations range from 0 to 1024 pM increasing as a power of 2 and following the scheme depicted in Fig. 2, covering all concentrations of biological interest. Note that in the table experiment vs. targets of Fig. 2 equal concentrations are found along the lower left-upper right diagonals, following thus a pattern known as Latin square. The data, which are pub-

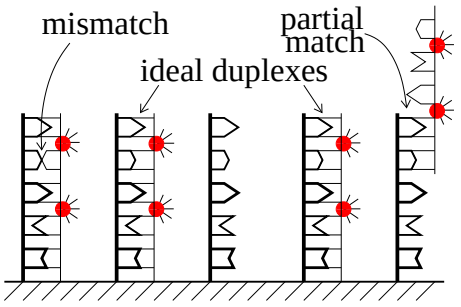


FIG. 1: DNA microarrays are based on the hybridization of surface-bound DNA probes (thick) with target sequences in solution carrying fluorescent labels (thin). Besides perfect matching probe-target pairs forming ideal duplexes, partial hybridizations, or mismatches are possible, although they are expected to be thermodynamically less stable.

| Experiments \ targets | 1 | 2 | 3 | 4 | 5 | ..... |
|---|---|---|---|---|---|---|
| 1 | 0 | 0.25 | 0.5 | 1 | 2 | ..... |
| 2 | 0.25 | 0.5 | 1 | 2 | 4 | ..... |
| 3 | 0.5 | 1 | 2 | 4 | 8 | ..... |
| ..... | ..... | ..... | ..... | ..... | ..... | ..... |
| N | 1024 | 0 | 0.25 | 0.5 | 1 | ..... |

FIG. 2: In the Latin square experiments some selected target sequences are added at known concentrations following the scheme indicated in the figure. Affymetrix considered 14 different concentrations ranging from 0 to 1024 pM (picomolars) and varying by a factor 2. In the Experiment 1, for instance, the RNA target 1 is absent from the solution (0 pM), the target 2 is present at a concentration of 0.25 pM ... In this scheme all possible 14 concentrations for each of the 14 target sequences are explored in 14 different experiments.

licly available from the Affymetrix web site [11], are important references for testing new algorithms that calculate gene expression levels from "raw" microarrays data. Not surprisingly, due to their central importance, the Latin square data have been analyzed by several groups through various approaches [3, 4, 5, 6, 7, 8, 9, 12, 13]. Our analysis is based on a simple physical model of target-probe hybridization. Although the modeling of microarray data with the physics of hybridization has been followed by other groups in the past couple of years [5, 7, 8], our approach differs from what has been done so far in the following ways: 1) We use the free energy parameters of formation of RNA/DNA duplexes in solution, and not the DNA/DNA parameters as in [8]. 2) We include the analysis of mismatches. 3) We include the effect of target-target hybridization in solution.

The latter effect turns out to be an essential feature of our approach: When target-target hybridization is neglected the fit of the experimental data is very poor for half of the 14 spike-in genes. On the contrary, when hybridization in solution is included we obtain good fits of the experimental data with a simple theory containing four fitting parameters. The ultimate test of the validity of our approach is through the analysis of scaling collapses: when plotted as a function of an appropriate rescaled thermodynamic variable, which depends on an effective temperature, on the hybridization free energies and on the target concentration, the Latin square data for different experiments tends to collapse into a single master curve. Although the noise level can still be large, significant deviations from this master curve are very rare. As we shall see, the deviations from the expected isotherm can be understood in several cases.

This paper is organized as follows: in Section II we discuss the thermodynamic parameters used in this paper. In Sec. III we present the analysis of the Latin square data and the model used. In Section IV we present few examples of probes deviating from the expected behavior and discuss the origin of these deviation. Section V concludes the paper with an overview of the results, open issues and a discussion of related works.

TABLE I: The stacking free energy parameters $\Delta G_{37}$ for RNA/DNA hybrids measured in solution at a salt concentration 1 M NaCl and $T = 37^\circ$ C [14]. The upper strand is RNA (with orientation 5′-3′) and lower strand DNA (orientation 3′-5′). Between parenthesis we give the DNA/DNA parameters.

| Seq. | $-\Delta G_{37}$ (kcal/mol) | Seq. | $-\Delta G_{37}$ (kcal/mol) |
|---|---|---|---|
| rAA/dTT | 1.0 (1.00) | rAC/dTG | 2.1 (1.44) |
| rAG/dTC | 1.8 (1.28) | rAU/dTA | 0.9 (0.88) |
| rCA/dGT | 0.9 (1.45) | rCC/dGG | 2.1 (1.84) |
| rCG/dGC | 1.7 (2.17) | rCU/dGA | 0.9 (1.28) |
| rGA/dCT | 1.3 (1.30) | rGC/dCG | 2.7 (2.24) |
| rGG/dCC | 2.9 (1.84) | rGU/dCA | 1.1 (1.44) |
| rUA/dAT | 0.6 (0.58) | rUC/dAG | 1.5 (1.30) |
| rUG/dAC | 1.6 (1.45) | rUU/dAA | 0.2 (1.00) |

## II. THERMODYNAMICS OF RNA/DNA HYBRIDS

The thermodynamics of duplex formation of nucleic acids in solution is well described by the nearest neighbor model according to which the free energy difference between a duplex and two separated strands is given by the sum of the local terms which keep into account hydrogen bonding and base stacking [10]. In melting experiments in solution one usually determines $\Delta H$ and $\Delta S$ the enthalpy and entropy differences between a duplex and two separate strands, from which the free energy difference $\Delta G = \Delta H - T\Delta S$ is obtained. The Table I gives the free energy parameters at 1 M of NaCl and at $T = 37^\circ$ C (data taken from Ref. [14]). The calculation of $\Delta G$ for a given sequence is obtained by summing up the data on Table I and adding to that a contribution of helix initiation $\Delta G_{37}^{\rm init.} = 3.1$ kcal/mol [14].

The thermodynamic parameters for RNA/DNA hybrids containing a single mismatch have recently been

determined [15]. The simple nearest neighbor model with stacking free energy parameters is no longer accurate for mismatches. For RNA/DNA single mismatches it has been found that the trinucleotide model, in which distinct free energies are associated to the triplet formed by the mismatch and the two neighboring nucleotides, fits the experimental data reasonably well [15]. As a MM probe in Affymetrix chips is realized by interchanging C with G and A with T in the middle nucleotide of a PM probe, there are four types of mismatches rGdG, rCdC, rUdT and rAdA. Taking into account the four possible combinations of neighboring nucleotides there are thus in total 64 different mismatches that should be considered. The free energy of only part of these 64 triplets can be found in the present literature [15]. The full list of mismatch free energies used in this paper is given in the Appendix A.

## III. LATIN SQUARE DATA

Usually, for the intensities $I$ measured in the Affymetrix experiment one distinguishes the two contributions from non-specific ($N$) and specific ($S$) hybridizations [3]. We follow the same idea here and write:

$$I(c, \Delta G) = N + S(c, \Delta G) + \varepsilon \tag{1}$$

where $\varepsilon$ denotes some experimental noise. Here $I$ is the intensity from the probe whose complementary RNA target is in solution at a concentration $c$ (known for the "spike-in" genes) and $\Delta G$ the hybridization free energy. Note that we did not make any distinction between PM and MM probes, as we assume that their specific binding will depend only on $c$ and $\Delta G$. The non-specific hybridization $N$ depends on the total RNA concentration in solution and possibly on other free energy parameters describing the partial matching with all RNA in solution. However, the precise form of $N$ is not relevant for the analysis performed in this paper, as we focus here on

$$\Delta I \equiv I(c) - I(0) \approx S(c, \Delta G) \tag{2}$$

The background subtraction is only possible in the Latin square set as there is always a reference measurement at zero spike-in concentration. The problem of calculating the "background" ($N$ in Eq. (1)) from first principles approaches will be addressed elsewhere.

As in Ref. [8], we model the specific hybridization as a two-state process where the target is either unbound in solution or fully hybridized to the probe forming a 25 nucleotides double helix at the surface, with one mismatch at position 13 for the MM probes. The Langmuir model predicts that:

$$S(c, \Delta G) = \frac{Ace^{-\beta \Delta G}}{1 + ce^{-\beta \Delta G}} \tag{3}$$

here $\beta = 1/RT$, where $T$ is the temperature and $R = 1.99$ cal/mol $\cdot$ K is the gas constant. The parameter $A$ sets the scale of the intensity and corresponds to the saturation value in the limit where $c \gg e^{\beta \Delta G}$, i.e. where the concentration is high or the binding is strong.

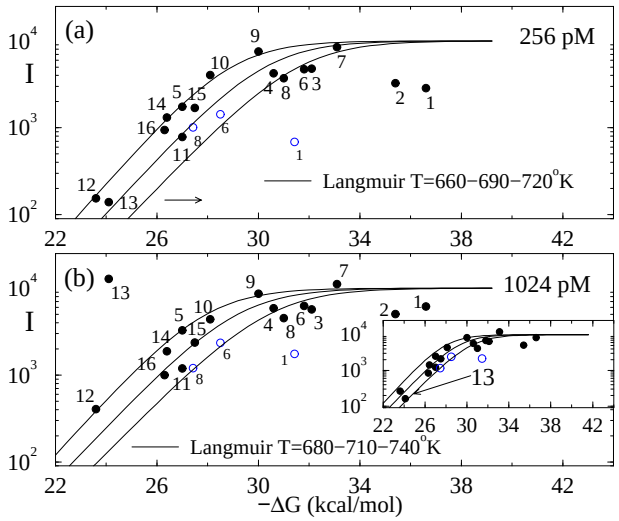


FIG. 3: Signal intensities versus the bulk hybridization free energy for the probe set 1708_at, "spiked-in" at concentration 256 pM (a) and 1024 pM (b). The data are obtained from the Affymetrix experiments 1521m99hpp_av06 and 1521a99hpp_av06, respectively. Filled and empty symbols refer to PM and MM probe sets. The solid lines are curves from Eq. (3) where $c$ is given in the experiment, $A = 10^4$ and for three values of the temperature (the arrow in (a) indicates the direction of increasing temperature). The inset shows a plot for a replicate of the 1024 pM experiment taken from the file 1532a99hpp_av04. Notice that in the latter the intensity of the PM probe 13 (indicated by the arrow) agrees very well with the Langmuir isotherm.

### A. High "spike-in" concentrations

It is convenient to analyze first the limit of high "spike-in" concentrations, which we find to correspond to $c \geq$ 256 pM. At such high concentrations typically $I(c) \gg I(0)$ thus the contribution of the background signal can be safely ignored. It is therefore equivalent to plot $I(c)$ or $\Delta I(c)$, as defined in Eq. (2).

Figure 3 shows a plot of $I$ vs. $\Delta G$ for the probe set 1708_at for the concentrations of 256 (a) and 1024 pM (b). Both PM and MM probes (filled and empty symbols) are shown. The numbers label the probes, following the notation chosen by Affymetrix. For the MM probes we could calculate 3 out of 16 free energies, using the data given in Table III. Although fluctuations are quite strong, the intensities shown in Figure 3 tend to align to a single master curve both for PM and MM probes.

The solid lines in Figure 3 are plots of the Langmuir isotherm given in Eq. (3). Assuming that the probe density is roughly constant for the whole array, we expect that the value of the saturation amplitude $A$ is the same

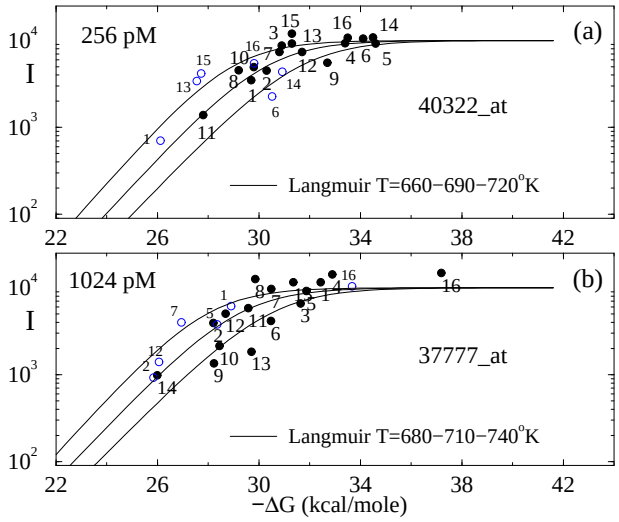


FIG. 4: As in Figure 3 for the probe set 40322_at. The data are obtained from Affymetrix experiments given in the files 1521b99hpp_av06 (a) and 1521d99hpp_av06 (b).

for all probes. We analyzed the histograms of intensities for the whole set of Latin square experiments and found that the probability of finding an intensity $I$ drops sharply beyond $I_{\max} \approx 10^4$, therefore we fix in the whole paper $A = 10^4$. As the concentration $c$ is known, the only free parameter is the temperature $T$. We find that best fits of the Langmuir isotherm with the experimental data are obtained for a temperature $T = 700 \pm 30°K$, roughly twice as large as the temperature in the Affymetrix experiment, which is of $T = 45°C \approx 320°K$.

The "discrepancy" between fitted and experimental temperatures deserves some discussion. We have estimated $\Delta G$ from a two state model summing up over the stacking parameters of Table I. In reality the binding of a target with a PM or MM probe can also involve fewer nucleotides. Moreover, the photolitographic process is not perfect and the surface bound probes have varying lengths (see Ref. [16]). These remarks indicate that the binding free energies are lower than those we have estimated on the basis of a simple two state process assuming that all probes have a fixed length of 25 nucleotides. However, we note that in plots of intensities versus $\Delta G$, the hybridization free energies calculated from Table I, the experimental data tend to align along a single master curve, as shown in Fig. 3 and in the rest of the paper. This suggests that $\Delta G$ is a good thermodynamic variable to parametrize the experimental data. The fact that the data follow a Langmuir isotherm suggests also that differences with the true hybridization free energy in the array can be reabsorbed in a rescaling of the temperature, as $\Delta G$ enters in the analysis through a Boltzmann weight $\exp(-\beta \Delta G)$. An "effective" temperature of about 700° K implies that on average $\Delta G_{\rm array} \approx \Delta G_{\rm sol}/2$. Being an "effective" parameter $T$ should not be compared directly to the experimental value. More important, for the purposes of this work, is the stability of $T$ as a fitting parameter: our analysis indicates that $T = 700°$ K fits rather well the experimental data for different probe sets and spike-in concentrations.

A rather high effective temperature ($T = 2100°$ K) was found in the fit of the Latin square data of Ref. [8]. The difference between our estimate and that of Ref. [8] is due to a different free energy parametrization (we use the more appropriate RNA/DNA values) and a different fitting procedure. Here we focus on fits of Langmuir isotherms as function of $\Delta G$, rather than as function of the concentration as done in [8]. These issues are discussed in Appendix B.

In Figure 3 one notices the presence of few "outliers", i.e. those probes whose intensities strongly deviate from the Langmuir isotherm, for instance as the probe 13 in Figure 3(b). The inset of Figure 3 shows a replicate of the experiment at a concentration of 1024 pM. In that case the intensity of probe 13 is in agreement with the Langmuir isotherm. The intensities from the probes 1 and 2 instead deviate systematically from the Langmuir isotherm in all replicates of the 256 pM and 512 pM experiments. The origin of these deviations is discussed below.

Figure 4 shows the intensities for the probe set 40322_at for "spike-in" concentrations of 256 pM (a) and for the probe set 37777_at at a concentration of 1024 pM (b). Again the trend of the PM and MM data is to align into a single master curve fitting quite well Eq. (3), when the same effective temperatures as in Figure 3 are used. Similar behavior is found for the other spike-in genes [17].

### B. General case

In order to test the global functional form of the Langmuir isotherm we turn now to the analysis of the full range of concentrations. From Eq. (3) we expect that the experimental data should "collapse" into a simple master curve when plotted as a function of the scaled variable $x = c \exp(-\beta \Delta G)$. A preliminary analysis at various temperatures at around $T = 700°$ K shows that the best fits are obtained for an effective temperature $T = 680°$ K, which we fix now once for all.

Figure 5 shows a plot of $\Delta I$, as defined in Eq. (2), vs. $x$ for the probe set 37777_at. Note that the large majority of probes follow indeed the Langmuir isotherm which takes the form $Ax/(1+x)$, and which is shown as a solid line in Figure 5. Only the intensities of the probe 16 deviate substantially from it. Quite interestingly, probe 16 still follows a Langmuir isotherm shifted along the horizontal axis. This shift is equivalent to a probe-dependent rescaling of the variable $x$. One can thus collapse all the data onto the curve $Ax'/(1+x')$ by plotting $\Delta I$ as a function of

$$x' = \alpha_k \; c \; e^{-\beta \Delta G}, \tag{4}$$

with $\alpha_k$ probe dependent. For instance, for the probe set

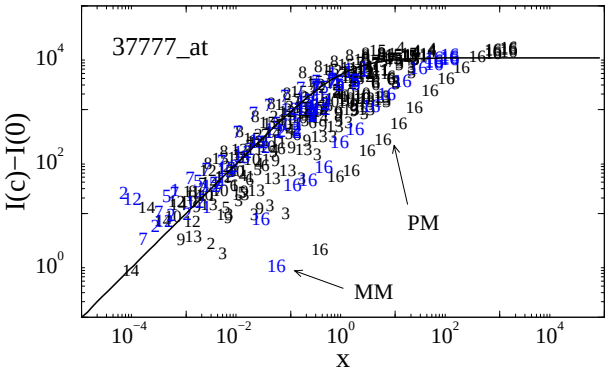


FIG. 5: Intensities for the Latin square experiment (set 1521 [11]) for the probe set 37777_at plotted as function of the rescaled variable $x = c\exp(\beta\Delta G)$. The probe numbers for both PM (smaller characters) and MM (bigger characters) are given.

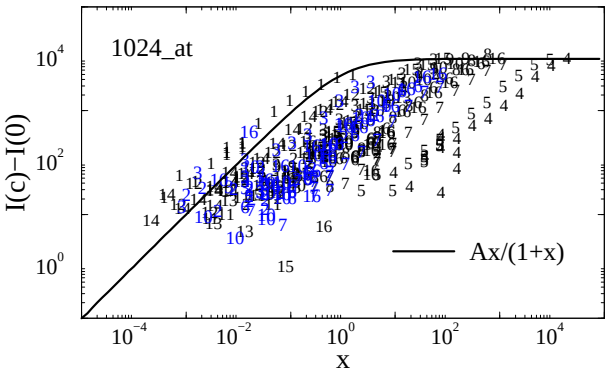


FIG. 6: Plot of $\Delta I$ vs. $x = c\exp(-\beta\Delta G)$ for the probe set 1024_at.

37777_at one could take $\alpha_k \approx 1$ for all probes except for probe 16 for which $\alpha_{16} \approx 10^{-3}$. While in Figure 5 only one probe deviates sensibly from the Langmuir isotherm, in other cases the disagreement involves the majority of the probes. An example is given in Figure 6, which shows a plot of $\Delta I$ vs. $x$ for the probe set 1024_at. We note that the shift along the $x$-axis is predominantly to the right side of the Langmuir isotherm, corresponding to a rescaling parameter $\alpha_k < 1$. An analysis of all the 14 "spike-in" genes shows that half of them are quite well-behaving in the sense that most of the data on a plot of $\Delta I$ vs. $x$ align along the Langmuir isotherm, as in Figure 5. The remaining half resembles more the example of Figure 6 (all figures are shown in [17]). A closer inspection to these defective probes shows that most of them have a rather high hybridization free energy, typically larger than $30 - 35$ kcal/mol.

A rescaling factor $\alpha < 1$ can also be interpreted as a lowering of the target concentration in solution $c' = \alpha c$. The most plausible explanation of this reduction of the concentration is the target-target hybridization in solution, or RNA secondary structure formation, as schematically illustrated in Figure 7. The figure shows an example four 25 nucleotides long regions of the target RNA, which are complementary to probes 1, 2, 3 and 4 of some probe set. The regions richer in CG, which have therefore higher hybridization free energy (in the example 2 and 4), tend to form stable duplexes with other RNA fragments or to form some secondary structure. Once hybridization in solution has occurred the amount of target RNA available for hybridization to the probe sequences is reduced.

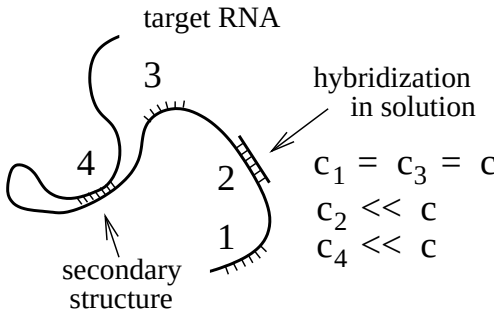


FIG. 7: Example of reduction of the effective concentration in solution of the target sequences due to hybridization with other RNA fragments in solution (2) or due to secondary structure formation (4). For probes 2 and 4 the target sequences available for hybridization is decrease. We model this effects with a simple function $\alpha$ decreasing exponentially in the RNA/RNA hybridization free energy [Eq. (5)].

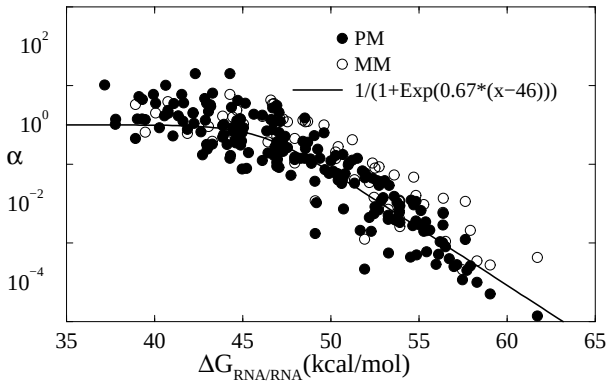


FIG. 8: Plot of the rescaling factor $\alpha$ needed to shift the data points to the Langmuir isotherm $Ax/(1+x)$ as function of the RNA/RNA hybridization free energies for each probe. The solid line is the fit to the Eq. (5) as expected from a simple model of bulk hybridization. The circles and diamonds emphasize the data from the probe sets 408_at and 36889_at which contain some defective probes. All these probes tend to deviate more strongly from the average behavior.

Figure 8 shows the reduction of the effective target concentration $\alpha_k$ for all the "spike-in" genes of the experiments 1521 as a function of the free energy of RNA/RNA duplex for each probe. The parameter $\alpha_k$ is determined from the distance of the experimental data to the Langmuir isotherm $Ax/(1+x)$ in $\Delta I$ vs $x$ plots. In Figure 8 the data follow two different behaviors below and above 45 kcal/mol. For low hybridization free energies $\alpha_k$ is

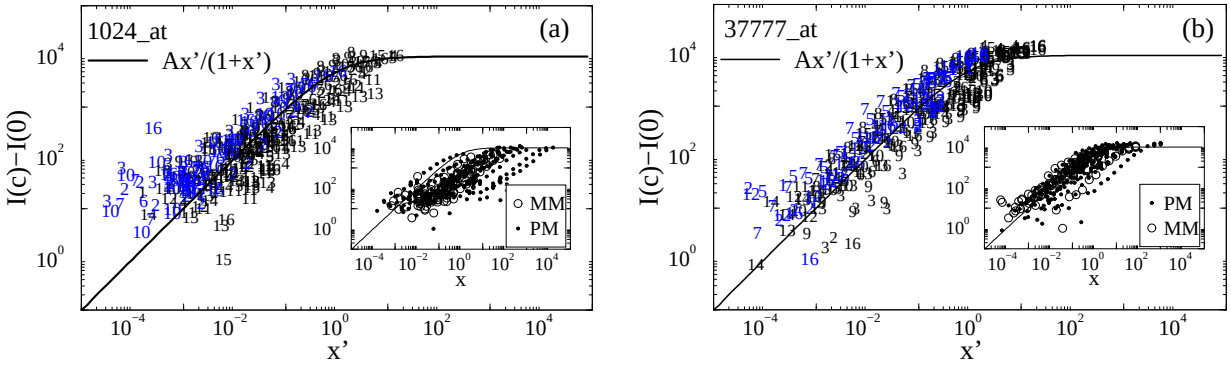


FIG. 9: Intensities for the probe sets 1024_at (a) and 37777_at (b) plotted as functions of the scaled variable $x' = \alpha c \exp(\beta \Delta G)$, which takes into account of $\alpha$, the fraction of target sequences hybridizing in solution. As a comparison we show in the insets the same quantity plotted as a function of $x = c \exp(\beta \Delta G)$.

constant and roughly equal to 1, indicating that those regions of the target RNA are single stranded and available for binding to the probes. For free energies larger than 45 kcal/mol $\alpha_k$ diminishes following approximately an exponential decay, due to the possible effect of enhanced hybridization in solution. We stress that the free energies shown in Figure 8 are for RNA/RNA duplexes and these are typically stronger than RNA/DNA or DNA/DNA counterparts. We fit the global behavior of $\alpha_k$ with the following equation

$$\alpha_{\rm k} = \frac{1}{1 + \tilde{c} \exp(-\beta' \Delta G_{\rm R})} \qquad (5)$$

where $\Delta G_{\rm R}$ is the RNA/RNA hybridization free energy in solution. The best fit of the data is shown as a solid line in Figure 8, which leads to $\tilde{c} \approx 10^{-2}$ pM and $\beta' = 0.67$ mol/kcal, i.e. $T' = 725°$ K.

The Eq. (5) resembles that for a two state process in which the target RNA reacts with a fragment with concentration $\tilde{c}$. In reality there are many different matching fragments hybridizing with the same target region. One should not view $\tilde{c}$ as a real concentration, rather the whole $\tilde{c} \exp(-\beta \Delta G_{\rm R})$ as a global relative probability for hybridization in solution, which is obtained by averaging over all these processes.

Having now fixed the four fitting parameters $A$, $T$, $\tilde{c}$ and $T'$, we can reanalyze the data collapse by using as a scaling variable $x' = \alpha_{\rm k} c \exp(\beta \Delta G)$, with $\alpha_k$ given in Eq. (5). Figure 9 shows the plot of $\Delta I$ with the new scaling variable for the probe set 1024_at (left) and 37777_at (right). Notice the nice collapse of all PM and MM intensities into a single master curve, now in much better agreement with the Langmuir isotherm. Similar plots for all "spike-in" genes show equally good collapses (plots for all 14 genes of the Latin square set are given in [17]).

We can use the proposed model to fit the experimental data keeping the absolute target concentration as the only fitting parameter. A plot of the fitted concentration as a function of the spike-in concentration is given in the inset of Fig. 10. The fitted concentration is, apart from the region $c_s < 1$ pM, within a factor two from the spike-in value $c_s$. A result which compares favorably with other algorithms [6, 8, 9] (for more details see [17]).

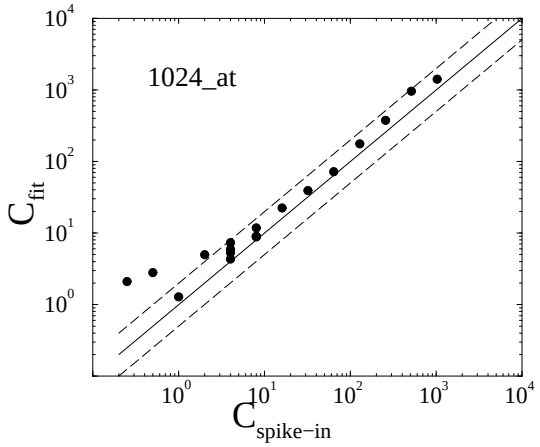


FIG. 10: Fitted concentration vs. spike-in concentration for the probe set 1024_at. The solid line is $y = x$, while the dashed lines correspond to $y = x/2$ and $y = 2x$.

## IV. DEVIATIONS FROM THE MODIFIED LANGMUIR ISOTHERM

An analysis of the 14 spike-in genes reveals that there are still a few probes deviating from the expected behavior of the modified Langmuir isotherm $Ax'/(1 + x')$, which takes into account the target-target hybridization in solution. Figure 11 shows two examples of such deviations: (a) the probe 9, both PM and MM, of the probe set 1091_at and (b) the probe 10 of the probe set 36202_at. These deviations are systematic as they are observed in other replicates of the Latin square experiments and at all concentrations. Note that the large majority of the probes are in quite good agreement with the Langmuir

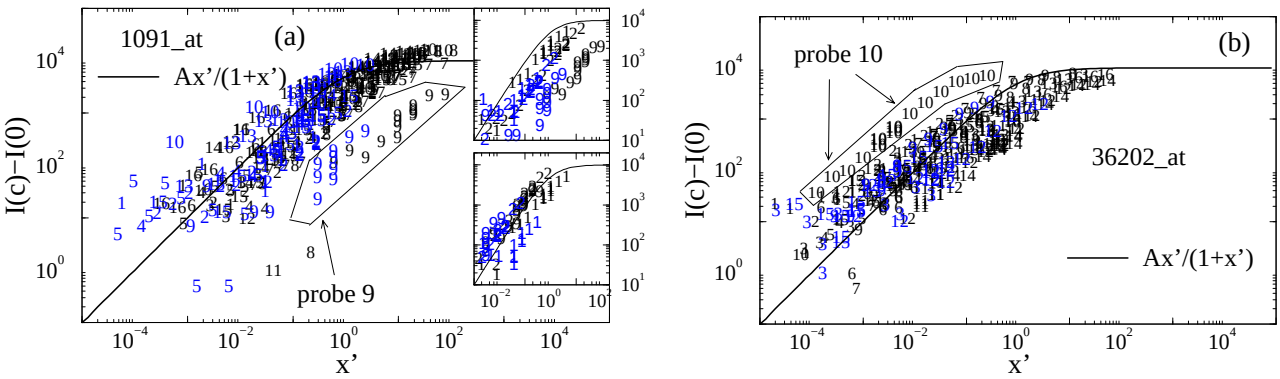


FIG. 11: Examples of deviations from the rescaled Langmuir isotherm $Ax'/(1+x')$. (a) The probe 9 of the probe set 1091_at has substantially lower signal than that expected. (b) The probe 10 of the probe set 36202_at has a significantly higher signal than that expected from the Langmuir isotherm. The insets of (a) show the intensities for the three "defective" probes, which do not align against the GenBank entries BC013368.2 and AL833563.1. All these probes have lower intensities than expected from the Langmuir curve $Ax'/(1+x')$ (upper inset). A recalculation of the hybridization free energies for these probes leads to a horizontal shift of the data, which are much closer to the Langmuir isotherm (lower inset).

isotherm as in the examples shown in Fig. 9. The deviations typically involve just one probe per probe set and they are observed in very few of the 14 spike-in genes of the Latin square set.

It is very instructive to look at these deviations more in detail. We performed a systematic sequence alignment against the whole human genome stored in public data banks (as GenBank) for all 14 probe sets of the Latin square experiments. Affymetrix arrays are produced by photolitographic techniques and each probe is synthesized in situ using the sequences taken from GenBank. However, the GenBank entries are continuously updated, and in some sequences errors may be present. The Affymetrix *NetAffx*™ Analysis center [18], provides information on the GenBank entries used to design the probe sequences.

Let us discuss first the probe set 1091_at. *NetAffx*™ indicates that this probe set was obtained from the GenBank entry M65066.1. A sequence alignment indeed shows that all the probe set sequences for the 1091_at match fully with the GenBank entry M65066.1. The alignment also shows that the two other sequences with GenBank entries BC013368.2 and AL833563.1 match perfectly with 13 of the 16 probes of probe set 1091_at, while the match is only partial for the probes 1,2 and 9. The Table II summarizes the results of the alignment for the probe set 1091_at. The difference is a single nucleotide close to the 5' and 3' ends for the probes 1 and 2, while there are 5 mismatching nucleotides for the probe 9. Note also that the GenBank sequence M65066.1 dates from 1994 (see Table II), while the two other entries are much more recent. We therefore suspect that the entry M65066.1 contains some annotation errors. As Affymetrix probe sequences are obtained by public databases, which are constantly updated, inconsistencies between probes and actual mRNA sequences may be present in some GenBank entries. If we assume that BC013368.2 and AL833563.1 contain the correct mRNA sequence, then the hybridization free energies that were used in Fig. 11(a) for the probes 1, 2 and 9 are overestimated and need to be corrected. Note that the three probes 1,2 and 9 have all intensities lower than expected from the Langmuir model as shown in the upper inset of Fig. 9(a).

Before discussing the free energies corrections, we recall that the Affymetrix RNA target in solution is actually anti-sense RNA, complementary to the usual mRNA sequences. Therefore the surface-bound probes have the same sequences as mRNA's, apart from the substitution of U with T.

*Probe 1*: The new mRNA annotation from the sequences BC013368.2 and AL833563.1 of Table II implies that a CA mismatch with a triplet rA<u>C</u>A/dT<u>A</u>T is formed when the target RNA hybridizes with the probe 1. There is no information in the present literature [15] about this mismatch triplet, therefore we cannot assign a free energy to it. We notice anyhow that the mismatch occurs very close to the 5′ end of the probe sequence (see Table II), which is the "free" end as the probes are linked to the substrate at their 3′ end. It is plausible that close to its free end the double helix RNA/DNA can be substantially distorted without a large penalty in free energy. The bases in the mismatch rC/dA should be still able to form two hydrogen bonds, therefore we consider likely that this mismatch does not affect substantially the hybridization free energy. This is a plausible explanation of the fact that the probe 1 (both PM and MM) deviates only slightly from the Langmuir isotherm, as shown in the inset of Fig. 11(a).

*Probe 2*: The new annotation implies an extra mismatch of the type rG<u>C</u>C/dC<u>C</u>G, for which no free energy has been given in the literature. The free energy

TABLE II: Best alignments for the probes 1, 2 and 9 of the probe set 1091_at. For each probe the first line is the sequence found in the Affymetrix chip, which aligns perfectly with the sequence with GenBank entry M65066.1 (second line). The sequences with GenBank entries BC013368.2 and AL833563.1 align perfectly with 13 of the probes in the probe set 1091_at, but they have some differences with the probes 1, 2 and 9. The differing nucleotides are underlined. The last column of the Table shows the sequence submission date to the GenBank.

| Probe | Origin | Sequence | GenBank | Date |
|---|---|---|---|---|
| | Affymetrix | 5′-TATGAGATTGATCTTGCCCCTAATT-3′ | | |
| 1 | Blast 1 | 5′-TATGAGATTGATCTTGCCCCTAATT-3′ | M65066.1 | 10-NOV-1994 |
| | Blast 2 | 5′-T<u>G</u>TGAGATTGATCTTGCCCCTAATT-3′ | BC013368.2 | 19-NOV-2003 |
| | Blast 3 | 5′-T<u>G</u>TGAGATTGATCTTGCCCCTAATT-3′ | AL833563.1 | 13-MAY-2003 |
| | Affymetrix | 5′-GCAGAAGTCAAGCCAGCCGCGGCCC-3′ | | |
| 2 | Blast 1 | 5′-GCAGAAGTCAAGCCAGCCGCGGCCC-3′ | M65066.1 | 10-NOV-1994 |
| | Blast 2 | 5′-GCAGAAGTCAAGCCAGCCGCGG<u>G</u>CC-3′ | BC013368.2 | 19-NOV-2003 |
| | Blast 3 | 5′-GCAGAAGTCAAGCCAGCCGCGG<u>G</u>CC-3′ | AL833563.1 | 13-MAY-2003 |
| | Affymetrix | 5′-CTGTCCTTGGTCCG_CATGGCTCGTT-3′ | | |
| 9 | Blast 1 | 5′-CTGTCCTTGGTCCG_CATGGCTCGTT-3′ | M65066.1 | 10-NOV-1994 |
| | Blast 2 | 5′-CTGTCCTTGGTCCG<u>AGGCT</u>GCTCGTT-3′ | BC013368.2 | 19-NOV-2003 |
| | Blast 3 | 5′-CTGTCCTTGGTCCG<u>AGGCT</u>GCTCGTT-3′ | AL833563.1 | 13-MAY-2003 |

differences between perfect triplets and triplets with a CC mismatch, as given in Figure 13, suggest as a rough estimate for the CC mismatch of about 4.5 kcal/mol. This causes a shift of the data toward a lower value of the variable $x'$ and shifts them much closer to the curve $Ax'/(1+x')$.

*Probe 9*: For the probe 9 the alignment of the sequence M65066.1 with those in BC013368.2 and AL833563.1 differs the most. The duplexes formed in this case are as shown in Figure 12 and contain an inner asymmetric loop with two arms with 4 and 5 nucleotides. It is difficult to evaluate the hybridization free energies for these configurations. A rough estimate, taking $\Delta G^{\rm loop}=0$, yields a free energy shift of $8-10$ kcal/mol.

We have used the estimated free energies to correct for the $x'$ variables for the probes 2 and 9. For instance a shift of 8 kcal/mol for the probe 9 implies a correction factor of $\exp(-8/RT)\approx 3\cdot 10^{-3}$, where we have used $T=700^\circ$ K. Analogously for the probe 2 we find $\exp(-4/RT)\approx 0.05$. The insets in Fig. 12(a) show the intensities for the three probes with conflicting alignments in the case of un-normalized data (top) and data with rescaled factors for the probes 2 and 9 (bottom). In the latter case the agreement with the Langmuir isotherm is substantially improved.

In order to assess on the possible frequency of annotation errors we have performed an alignment analysis of all probe sets of the Latin square set (for more details see [17]). The only potential annotation problems were those detected for the probe set 1091_at, which suggests that these errors should not be too frequent, at least in the human genome.

We turn now to the probe 10 of Fig. 11(b), which has a signal significantly higher compared to the Langmuir prediction. In the whole set of Latin square data we found only another example of a similar high signal, namely

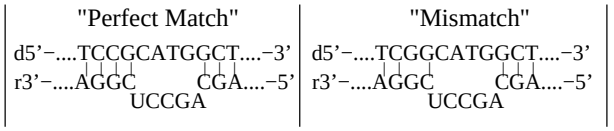


FIG. 12: Expected duplexes formed with the "defective" probe 9 of the set 1091_at for the PM and MM signal, if the mRNA sequence is taken from the entries BC013368.2 and AL833563.1. The upper strand is the surface-bound DNA, while the lower one is the RNA target.

the probe 16 of the probe set 36085_at. Analyzing these two sequences we find that they share a common feature: both are A-rich close to the 3′ end. The sequences are ...CACAAAAG-3′ (36202_at10) and ...CAATAAA-5′ (36085_at16). Note that the Table I gives for the combination rUU/dAA the lowest free energy. A possible improvement of the data collapse could be obtained by introducing position-dependent weights $w_i$ with $i=1,2\ldots$ so that

$$\Delta G=\Delta G^{\rm init.}+\sum_{i=1}^{24} w_i\Delta G_i \qquad (6)$$

where $\Delta G_i$ is taken from Table I and the sum is extended over the nucleotides of the target sequence. One could assume that far from the substrate $w_i\approx 1$, so that the hybridization free energy reaches the solution limit and that $w_i<1$ close to the substrate. A reweighting of this type lowers the global value of $\Delta G$ (and consequently decreases the effective fitting temperature) and could lower the anomalously high signal of the two probes A-rich at their 3′ end. It is not clear a priori which function to choose for the weight $w_i$, apart from the fact that it should increase far from the surface, and different possibilities will be explored elsewhere. We stress however

that the choice $w_i = 1$, used in this paper, provides already a quite good fit of the experimental data except in very few cases.

## V. DISCUSSION

In this paper we have analyzed a series of controlled experiments on Affymetrix microarrays using a simple model of RNA/DNA hybridization. We have shown that for each probe (PM and MM) of a probe set the intensities plotted as a function of the free energy of RNA/DNA hybridization tend to align along a single master curve in quite good agreement with a Langmuir isotherm. In fact the intensities of half of the "spike-in" genes are already well fitted with the simple form given in Eq. (3). For the other half, those containing probes with higher CG content, we found that one has to include the effect of target hybridization in solution, which diminishes the effective concentration of single stranded RNA sequences in solution. This effect is well described by a simple analytical form given by Eq. (5). Despite this very simplified model the fit with the experimental data is very satisfactory, as shown by the scaling collapses (i.e. plots of intensities as function of the rescaled variable $x'$ of Eq. (4)). Although the data are somewhat noisy the calculations of the target concentrations are in good agreement with the input spike-in values (see Fig. 10 and [17]). This is due to the fact that concentrations are obtained from averaging over the signal of each individual 16 PM probes and of the MM probes of which we were able to include in the analysis. The averaging over these data points yields quite robust and reliable estimated of target concentration values.

The feature that is still missing in our analysis is the calculation of the background level ($N$ in Eq. (1)). We circumvented this problem by subtracting from the intensities those measured at zero spike-in concentration. This is only possible for the Latin square data. A good estimate of the background level could help in improving the quality of the fits in the low concentrations limit. These issues will be considered elsewhere.

The physics of the hybridization in high density microarrays has been investigated recently by other groups. We comment now on the differences between the present approach and what has been done in the literature so far. In a recent paper Hekstra *et al.* [6] found nice agreement of the Latin square data with a Langmuir model. Their plots of rescaled intensities versus rescaled concentrations follow very well the curve $x/(1+x)$, with small fluctuations. For their rescaling they use probe-dependent values, a procedure which requires the use of 24 fitting parameters. The advantage of our approach is that we find good collapses of the experimental data using a very simple model with only few fitting parameters.

It has been recently claimed that the MM probes do not follow the behavior predicted by the standard hybridization theory [19]. Our analysis, instead, shows that MM probes intensities follow the same Langmuir isotherm as the PM probes. For the mismatches we used the trinucleotide free energies for RNA/DNA duplexes in solution [15]. A very important aspect of the MM hybridization, as highlighted in studies of RNA/DNA duplexes melting in solution [15], is that their free energy strongly depends on the type and order of the two nucleotides close to the mismatch. This is probably the reason why an analysis of the mismatches based on single base pairs energies, as in Ref. [7], shows deviations from the Langmuir isotherm of the PM probes. We believe that the strong dependence of the mismatch free energy on the two neighboring nucleotide is a very important aspect for the correct modeling of the hybridization of MM probes.

In Ref. [8], the Langmuir model for target-probe hybridization was used to fit Affymetrix Latin square data. The hybridization free energies were obtained from values of DNA/DNA duplexes in solution [20], and not for RNA/DNA duplexes, as we have done here. As discussed in Section II there are some differences between the two sets of parameters. An explicit example emphasizing the influence of these differences for the fitting procedure is shown in the Appendix B. Our results show that a correct free energy parametrization improves substantially the quality and stability of the fits.

In other studies [5, 7], the free energies were obtained directly from a fit of Affymetrix data assuming a input relationship $I(\Delta G)$. The binding free energies were taken dependent on the distance from the substrate [5, 7], while we have so far calculated free energies by summing up uniformly over all the stacking energies of the probe sequences. As pointed out before, a position dependent weight in the free energy calculation may improve the quality of our data collapses for those few A-rich probes close to the substrate which we found to deviate more strongly from the Langmuir model. The overall quality of the fits remains however quite good also in absence of position-dependent binding (see Ref. [17]).

Another effect which has been claimed to be relevant for hybridization in high density DNA microarrays is the Coulomb interaction between a highly negatively charged surface DNA layer and negatively charged target molecules [21, 22]. These effects may play a role for a system with monodisperse probe length distribution. However, in Affymetrix chips the probe lengths are widely distributed [16], An analysis of the electrostatic interaction [17], shows that its strength is much weaker compared to that of the systems studied in Refs. [21, 22]. It is thus possible to neglect electrostatic effects, as we did here and as done in other studies involved with the physical modeling of hybridization in Affymetrix arrays [5, 6, 7, 8].

Finally, one may wonder how representative the spike-in targets chosen by Affymetrix are for the overall behavior of the microarray. A recent investigation [23] of several human housekeeping genes (i.e. those which are expressed in virtually all tissues) and of the Affymetrix spike-in data for the chipset HGU133 shows that the in-

TABLE III: $\Delta G_{37}$ for triplet mismatches in RNA/DNA duplexes, where the upper strand is RNA (the orientation is 3′ to 5′ from left to right) and lower strand DNA. Only the mismatches which are realized in the Affymetrix chip are shown. *a*: Deduced from the rUdG mismatches, *b*: Deduced from the rAdA mismatch. *c*: Deduced from the rCdT mismatches. *d*: Deduced from the rCdA mismatches. Between parenthesis is the free energy of a perfectly matching triplet obtained by interchanging C with G and A with T in the central nucleotide of the DNA strand.

| Sequence | $\Delta G_{37}$(kcal/mol) | Sequence | $\Delta G_{37}$(kcal/mol) |
|---|---|---|---|
| GG-mismatches | | | |
| rC**G**G<br>dG**G**C | 0.11 (-4.6) | rC**G**C<br>dG**G**G | -0.97 (-4.4) |
| rG**G**G<br>dC**G**C | -1.26 (-5.8) | rG**G**C<br>dC**G**G | -2.25 (-5.6) |
| rA**G**A<br>dT**G**T | 0.48[a] (-3.1) | rA**G**C<br>dT**G**G | -0.62[a] (-4.5) |
| rC**G**U<br>dG**G**A | 1.24[a] (-2.8) | rU**G**G<br>dA**G**C | 0.67[a] (-4.5) |
| AA-mismatches | | | |
| rC**A**G<br>dG**A**C | 1.05 (-2.7) | rC**A**C<br>dG**A**G | 0.32 (-3.0) |
| rG**A**G<br>dC**A**C | 0.20 (-3.1) | rG**A**C<br>dC**A**G | 0.29 (-3.4) |
| UT-mismatches | | | |
| rC**U**G<br>dG**T**C | 1.05[b] (-2.5) | rC**U**C<br>dG**T**G | 0.25 (-2.4) |
| rG**U**G<br>dC**T**C | 0.44 (-2.7) | rG**U**C<br>dC**T**G | -0.22 (-2.6) |
| CC-mismatches | | | |
| rC**C**G<br>dG**C**C | 0.73[c] (-3.8) | rC**C**C<br>dG**C**G | - (-4.2) |
| rG**C**G<br>dC**C**C | 0.28[d] (-4.4) | rG**C**C<br>dC**C**G | - (-4.8) |

tensities within a given probe set follow a distribution which is very similar to that observed in this work.


### Acknowledgments

E.C. would like to thank the Isaac Newton Institute for Mathematical Sciences in Cambridge, where this work was started, for kind hospitality. We are grateful to G. Barkema and J. Klein-Wolterink for interesting discussions. We acknowledge financial support from the Van Gogh Programme d'Actions Intǵrées (PAI) 08505PB of the French Ministry of Foreign Affairs.


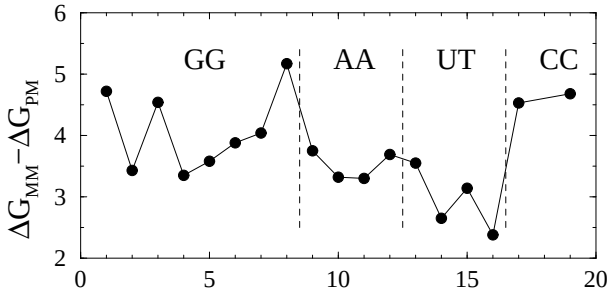


FIG. 13: Free energy differences between MM and PM triplets for the 18 mismatches given in Table III, where the labeling follows the same ordering as the Table. As an example the first point is the free energy difference between rCGG/dGGC and rCGG/dGCC, the second point the free energy difference between rCGC/dGGG and rCGC/dGCG ...

## APPENDIX A: FREE ENERGIES FOR MISMATCHES

The Table III shows all the free energies for triplets with a single mismatch used in this paper. Part of these free energies are obtained from experimental results of Ref. [15]. In some cases the free energies were deduced from the analogy with other mismatches. For instance, as pointed out in Ref. [15], the free energies for triplets with rAdA mismatches are close to those of triplets with rUdT mismatches. In absence of experimental determinations of mismatch free energies for the triplet rCUG/dGTC, we assign to the latter the same free energy as the mismatch rCAG/dGAC, which is of 1.05 kcal/mol. Of the 18 triplet free energies in Table III, 11 were obtained by direct experimental data inputs, while 7 from similarities with other mismatches.

Another interesting quantity is the free energy difference between a perfect matching triplet and one with a central mismatch, i.e. the free energy shift due to a single mismatch. Figure 13 reports the free energy differences for the 18 mismatches given in Table III, following the same order. These differences are obtained by subtracting from the data in Table III the values between parenthesis, corresponding to a perfect matching triplet. The Figure 13 shows that the free energy difference is quite sensitive to the type of mismatch and of its two neighboring nucleotides.

## APPENDIX B: COMPARING RNA/DNA WITH DNA/DNA HYBRIDIZATION FREE ENERGIES

The approach followed in Ref. [8] uses hybridization free energy for DNA/DNA duplexes in solution, while throughout this paper we used RNA/DNA parameters, as in the Affymetrix Latin square experiments the target is composed by RNA, while the probes are surface-bound DNA sequences. In order to illustrate the difference in the intensity vs. free energy plots we show in Figure 14

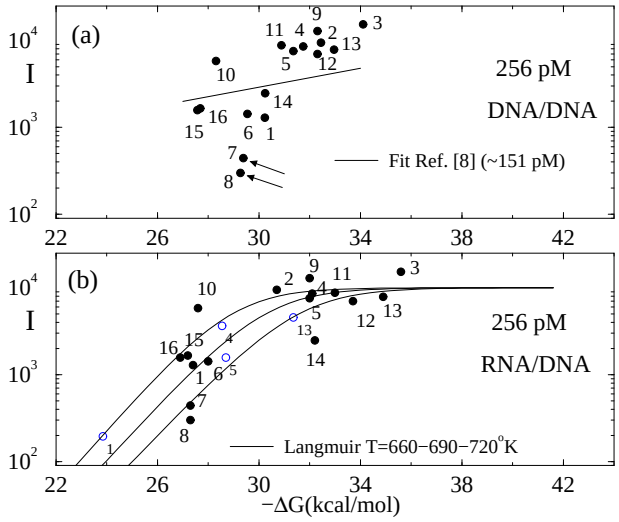


FIG. 14: Intensity for the probe set 36311_at for a target concentration of 256 pM plotted as function of (a) DNA/DNA and (b) RNA/DNA hybridization free energies. (a) reproduces Figure 5 of Ref. [8], and the solid line is the best fit with the parameters used in that reference; the probes 7 and 8 (indicated as arrows) are considered as outliers. In the case (b) the solid lines are the Langmuir isotherms from Eq. (3), $A = 10^4$, $c = 256$ pM and three values of the temperature.

the fit to the Langmuir model the intensities of the probe set 36311_at (from the Affymetrix file 1521g99hpp_av06) when (a) DNA/DNA and (b) RNA/DNA hybridization free energies are used. The figure shows the same data of Figure 5 of Ref. [8]. The best fit obtained from the parameters of Ref. [8] is the thick solid line shown in Figure 14(a), where the probes 7 and 8, indicated by arrows, were considered as outliers. In Figure 14(b) the solid lines are the Langmuir isotherms with the same parameters as in Figures 3(a) and 4(a). By comparing the two plots one can conclude that the fit in Figure 14(b), which is also consistent with plots for other probe sets (see Figs. 3 and 4), is more convincing than that shown in the case (a). Note that in Fig. 14(b) the four MM intensities follow the same behavior as the PM probes. The analysis of Ref. [8] is restricted to PM probes only.

The difference between the fits performed here and those reported in Ref. [8] are also due to a different approach to the analysis. In Ref. [8] the intensities for each probe are fitted as a function of the concentration $c$ using the three adjustable parameters $A$, $K = \exp(-\beta\Delta G)$ and the background level $N$. Although a three parameters fit appeared to reproduce experimental data for different probes very well [8] one of the problems with this analysis is that $A$ was found to vary over one order of magnitude from probe to probe. As this is an unphysical feature, $A$ was then kept constant for all probes while the other two parameters were allowed to vary [8]. Ref. [8] reports as best fitted value $A \approx 9\ 500$, which compares favorably to our choice $A = 10^4$. The problem is that the fits of Intensities vs. concentration obtained by binning over different free energies are less convincing (see Fig. 3 of Ref. [8]). Analyzing then the decay of $1/K$ as function of the DNA/DNA hybridization free energy an effective temperature of $T \approx 2\ 100°$ K is found. This is roughly three times higher of what we find in this paper, from a direct analysis of plots of intensities vs. RNA/DNA hybridization free energies. The use of an accurate free energy parametrization is very important: as $\Delta G$ is related to the intensity thorough an exponential factor, small variations of $\Delta G$ estimates may have profound influences on the values and robustness of the fitting parameters.


[1] P. O. Brown and D. Botstein, Nature **21**, 33 (1999).
[2] R. J. Lipshutz et al., Nature **21**, 20 (1999).
[3] C. Li and W. H. Wong, Proc. Natl. Acad. Sci. **98**, 31 (2001).
[4] B. M. Bolstad et al., Bioinformatics **19**, 185 (2003).
[5] L. Zhang, M. F. Miles, and K. D. Aldape, Nature Biotech. **21**, 818 (2003).
[6] D. Hekstra et al., Nucleic Acids Res. **100**, 1962 (2003).
[7] F. Naef and M. O. Magnasco, Phys. Rev. E **68**, 011906 (2003).
[8] G. A. Held, G. Grinstein, and Y. Tu, Proc. Natl. Acad. Sci. **100**, 7575 (2003).
[9] J. M. Deutsch and S. Liang and O. Narayan, preprint q-bio.BM/0406039.
[10] V. A. Bloomfield, D. M. Crothers, and I. Tinoco, Jr., *Nucleic Acids Structures, Properties and Functions* (University Science Books, Mill Valley, 2000).
[11] http://www.affymetrix.com/analysis/download_ center2.affx.
[12] H. Binder and S. Preibisch, "Specific and non-specific hybridization of oligonucleotide probes on microarrays", preprint q-bio.BM/0410028.
[13] C. J. Burden and Y. Pittelkow and S. R. Wilson, "An adsorption model of hybridization behaviour on oligonucleotide microarrays", preprint q-bio.BM/0411005.
[14] N. Sugimoto et al., Biochemistry **34**, 11211 (1995).
[15] N. Sugimoto, M. Nakano, and S. Nakano, Biochemistry **39**, 11270 (2000).
[16] J. E. Forman et al., ACS Symposium Series **682**, 206 (1998).
[17] http://iri.ibl.fr/bn/papers/affy_paper.html.
[18] See http://www.affymetrix.com/analysis/index.affx.
[19] F. Naef, D. A. Lim, N. Patil, and M. Magnasco, Phys. Rev. E **65**, 040902(R) (2002).
[20] J. SantaLucia Jr., Proc. Natl. Acad. Sci. **95**, 1460 (1998).
[21] A. Vainrub and B. M. Pettitt, Phys. Rev. E **66**, 041905 (2002).
[22] A. Halperin, A. Buhot, and E. B. Zhulina, Biophys. J. **86**, 718 (2004).
[23] L.-C. Tranchevent, T. Heim and E. Carlon, unpublished.