Physical-Chemistry-Based Analysis of Affymetrix Microarray Data

T. Heim,[†] L.-C. Tranchevent,[†] E. Carlon,*[,†,‡] and G. T. Barkema[§]

Interdisciplinary Research Institute c/o IEMN, Cité Scientifique BP 60069, F-59652 Villeneuve d'Ascq, France, Ecole Polytechnique Universitaire de Lille, Cité Scientifique BP 60069, F-59652 Villeneuve d'Ascq, France, and Institute for Theoretical Physics, University of Utrecht, Leuvenlaan 4, 3584 CE Utrecht, The Netherlands

Received: May 11, 2006; In Final Form: July 11, 2006

We analyze publicly available data on Affymetrix microarray spike-in experiments on the human HGU133 chipset in which sequences are added in solution at known concentrations. The spike-in set contains sequences of bacterial, human, and artificial origin. Our analysis is based on a recently introduced molecular-based model (Carlon, E.; Heim, T. *Physica A* **2006**, *362*, 433) that takes into account both probe−target hybridization and target−target partial hybridization in solution. The hybridization free energies are obtained from the nearest-neighbor model with experimentally determined parameters. The molecular-based model suggests a rescaling that should result in a "collapse" of the data at different concentrations into a single universal curve. We indeed find such a collapse, with the same parameters as obtained previously for the older HGU95 chip set. The quality of the collapse varies according to the probe set considered. Artificial sequences, chosen by Affymetrix to be as different as possible from any other human genome sequence, generally show a much better collapse and thus a better agreement with the model than all other sequences. This suggests that the observed deviations from the predicted collapse are related to the choice of probes or have a biological origin rather than being a problem with the proposed model.

1. Introduction

DNA microarrays (see, e.g., refs 1 and 2) allow the measurement of the gene expression levels of thousands of genes simultaneously. This is a major step forward compared to traditional methods in molecular biology (such as Northern blots) that are applicable only to a limited set of genes at a time. The determination of gene expression levels is not the only application of DNA microarrays, which have been used also for the analysis of genetic variance between individuals (single nucleotide polymorphisms), as efficient tools for DNA sequencing, for the study of chromosomal defects, and for the determination of alternative splicing events.

Despite the increasing popularity that microarrays have known in the recent years there are still some problems with the technology. There has been, for instance, only a moderate effort in comparing different microarray platforms on the same biological system.[3] When this comparison was made, as in a recent study on expression analysis of stressed pancreas cells, it was found that different commercial platforms produced wildly incompatible data.[4] These problems call for a better fundamental understanding of the functioning of the microarrays. Such understanding will help researchers to design better algorithms for microarray data analysis based on the physical chemistry of the underlying hybridization process.

The basic mechanism underlying the functioning of DNA microarrays is that of hybridization (i.e., the binding between complementary single-stranded nucleic acids) between a strand anchored at the surface and a strand in solution, referred to as probe and target, respectively. Microarrays are produced in different ways: The DNA anchored at the surface is either deposited with a droplet[5] (spotting techniques) or synthesized in situ by photolitography as in the Affymetrix arrays.[6] In spotted arrays the deposited strands are not limited in length and typically vary from 30 to a few hundred bases, while in Affymetrix arrays the lengths of the surface strands are fixed at 25 bases.

In a previous paper[7] we have analyzed a series of publicly available data of experiments performed on Affymetrix microarrays, using a simple model of the hybridization process. In these experiments a set of selected genes are "spiked-in" at fixed concentrations into a solution containing other types of RNAs. This set of data has been widely used as a test ground for algorithms designed to extract gene expression levels from the raw data. Affymetrix is one of the major commercial producers of microarrays. Although in Affymetrix arrays the DNA anchored to the surface is limited to rather short oligos (25 nucleotides long) one of the advantages is that a high density of probe sequences per array can be obtained. In the latest generation 1 400 000 different probes have been placed in a single array. The large number of probes compensate for their limited length. Indeed Affymetrix uses multiple probes per gene, which define a so-called *probe set*: the number of probes per probe set varies typically between 10 and 20. Another peculiar feature of Affymetrix chips is that it uses as control a mismatch (MM) probe sequence, which differs from a perfect-matching (PM) sequence only at the base at position 13: A nucleotide A is interchanged with T, and a nucleotide C is interchanged with G.

In our previous work[7] we focused on the spike-in data set of the HGU95 human chipset, obtained with a background of RNA from the human pancreas. More recently this has been substituted by the HGU133 chipset, where spike-in experiments are carried out with a background of RNA from the HeLa (human

* Author to whom correspondence should be addressed. E-mail: enrico.carlon@polytech-lille.fr.
[†] Interdisciplinary Research Institute.
[‡] Ecole Polytechnique Universitaire de Lille.
[§] University of Utrecht.

10.1021/jp062889x CCC: $33.50
Published on Web 10/20/2006

adenocarcinoma) cell line. Probe sets have been completely redesigned in the HGU133 chipset; moreover there are typically only 11 probes per probe set compared to the 16 probes of the HGU95 array. In this paper we focus on the analysis of publicly available spike-in data on the HGU133 chip, building on our previous work[7] on HGU95. The first goal of this manuscript is to test the robustness of the model introduced in ref 7 to a new set of data.

The model of ref 7 features four fitting parameters: the effective inverse temperature β' and concentration $\tilde{c}$, used in the description of the hybridization in solution, and the effective temperature β and saturation intensity A, used in the description of the probe−target hybridization. The second goal of this paper is to investigate the physicochemical basis for the observed (fitted) values of these four fitting parameters. Third, we exploit an interesting feature of the spike-in data of the HGU133 chipset: Different from the HGU95 data where spikes correspond to human genes, the spikes in the HGU133 have been selected between human, bacterial, and "artificial" sequences. The latter were selected by Affymetrix to avoid cross-hybridization with any known human coding sequence. We will compare the quantitative agreement between spike-in data and the model of ref 7, distinguishing these three types of sequences.

Several papers[8−14] have been devoted to the modeling of physicochemical aspects of DNA hybridization to surface anchored strands using the Langmuir model, i.e., the basic model for surface adsorption/hybridization, and its extensions. The different approaches have been reviewed recently, for instance, in refs 15 and 16. In principle several different effects may play a role in the hybridization process. For instance, it has been pointed out[8,12] that electrostatic interactions between negatively charged strands in solutions and a negatively charged layer of DNA molecules may have the effect of impeding hybridization. Other issues that have been discussed in the recent literature include the effect of the surface on the hybridization dynamics,[11] the length of DNA binding to the surface,[16] factors influencing the limiting behavior of the hybridization at strong target concentrations,[14] and the role of marker molecules (as the biotin linker) on the hybridization affinity.[10,17−19] All these effects will most likely play a role in microarray experiments as well. However, since we could not incorporate these effects into our model without introducing new fitting parameters, we decided to use the basic model of ref 7.

2. A Simple Model for Hybridization in Affymetrix Arrays

In this section we briefly recall the model introduced in ref 7. Two basic processes are considered: (1) target−probe hybridization and (2) target−target hybridization in solution. According to the model the fluorescence signal measured from a given probe is

$$I = I_0 + \frac{A\alpha c\, e^{\beta\Delta G}}{1 + \alpha c\, e^{\beta\Delta G}} \qquad (1)$$

where I_0 indicates a background level due to nonspecific hybridization, A sets the scale of the intensities, c is the target concentration (a measure of the gene expression level), ΔG is the target−probe hybridization free energy, $\beta = 1/RT$ is the inverse temperature, and R is the universal gas constant. Here, α models the reduction in the concentration of available targets due to the target−target hybridization in solution: Only a fraction αc is available for the hybridization with probes as the remaining $(1 - \alpha)c$ form stable duplexes with other partners in solution (Figure 1a).

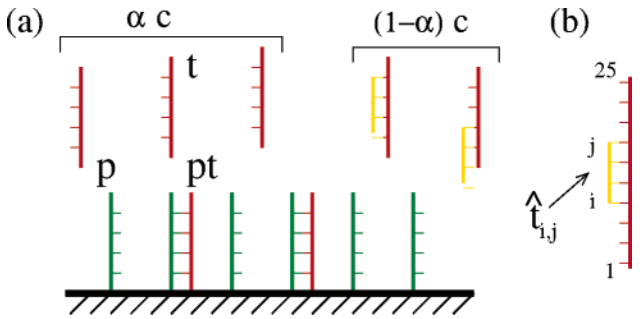

Figure 1. (a) Simple model of hybridization in Affymetrix microarrays used throughout this paper. It is defined by two basic reactions: (1) hybridization between target molecules (t) to surface anchored probes (p) leading to a duplex pt and (2) hybridization between target molecules in solution leading to the partial duplexes $t\hat{t}_{i,j}$. In the model, the effect of the hybridization in solution amounts to a reduction of the original target concentration c to a value αc. (b) Partial hybridization of a fragment in solution complementary to the target RNA sequence from base i to base j ($1 \le i < j \le 25$).

Many different scenarios of hybridization in solution have been discussed by other authors.[16,20] In the model introduced in ref 7, we approximate the target−target hybridization with the expression

$$\alpha \approx \frac{1}{1 + \tilde{c}\, \exp(\beta' \Delta G_R^{(37)})} \qquad (2)$$

with β' and $\tilde{c}$ fit parameters and $\Delta G_R^{(37)} \equiv \Delta G_R(1, 25)$, the (sequence-dependent) RNA/RNA free energy for duplex formation in solution at 37° C calculated over the whole 25-mer length; in close approximation, the binding free energies at 37 and 45 °C (the actual experimental temperature) are almost identical, apart from a small scaling factor, which is adsorbed into the rescaled temperature β'. In the next section, we will discuss the steps leading to eq 2 in more detail.

In the model defined in eqs 1 and 2 the hybridization free energies ΔG and ΔG_R are calculated from tabulated experimental data for DNA/RNA[21,22] and RNA/RNA[23] duplex formation in solution. The four parameters A, β, β', and $\tilde{c}$ were fitted against the spike-in data of the Affymetrix array HGU95a in ref 7. The parameters β', $\tilde{c}$, and A will be discussed in sections 3 and 4. The parameter β is the inverse temperature. Instead of fixing it to the experimental value we have kept it as a fitting parameter as explained in ref 7. The best fit yields a value $T \approx$ 700 K, which is twice as large as the true experimental temperature $T = 45$ °C. This suggests that the actual free energies involved in the chip hybridization are roughly 50% lower than the ΔG estimated from the experimental data in solution. The possible origin of this discrepancy has been discussed in ref 7. It has also been recently suggested that the origin of this difference may be due to the presence of a denaturant, dimethyl sulfoxide (DMSO), used in Affymetrix experiments.[16] It is likely that many other effects may play a role such as polydispersity in probe lengths, molecular crowding, partial target−probe hybridization, etc. In a recent work[24] we included part of these effects and found indeed a reduction of the effective temperature to $T \approx 500$ K, still somewhat higher than the true experimental value but significantly closer to it. This more refined model however does not substantially improve the fit to the experimental data; we therefore restricted ourselves to the original model introduced in ref 7, with the same set of fitting parameters.

We note that we fit mismatches and perfect matches with the same model. The difference between the two is that there is a different hybridization free energy ΔG: One expects a lower signal for mismatches compared to those of perfect matches, due to weaker binding. This is not always the case; as remarked in several studies for a substantial fraction of probes (30%, as reported in ref 10) one observes "bright mismatches" for which the mismatch intensity I_{MM} exceeds the intensity I_{PM} of the perfect match. However, it has been observed[13] that bright MM come predominantly from probes with low intensities, which suggests that bright mismatches are associated with weak specific hybridization when the signal I is dominated by I_0 in eq 1.

In recent work[24] we also compared the current model with the approach based on position-dependent effective affinities as, for instance, described in refs 10 and 13. The conclusion is that the two approaches are fully consistent with each other, provided that various effects are incorporated such as partial unzipping of the probe−target complex, less than 100% efficiency in the probe growth during lithography, and entropic repulsion between the target and the substrate. These additional effects are the main factors causing position dependence (and thus allowing for a comparison with position-dependent effective affinities); for a quantitative prediction of the intensities, their combined effect can be well approximated by a slight decrease of β in eq 1, and they are therefore not included in the current study.

3. Hybridization in Solution

We now discuss the approximations leading to the form of α. We denote the concentration of free 25-mer targets in solution as $[t]$, the concentration of free target strands that are complementary from nucleotide i up and including nucleotide j as $[\hat{t}_{i,j}]$, and the concentration of duplexes between these two as $[t\hat{t}_{i,j}]$. Chemical equilibrium (Figure 1b) yields for the equilibrium constant

$$K_{i,j} = \frac{[t][\hat{t}_{i,j}]}{[t\hat{t}_{i,j}]} = \mathrm{e}^{-\beta \Delta G_R(i,j)} \tag{3}$$

where $\Delta G_R(i,j)$ is the RNA/RNA hybridization free energy for target molecules in solution, which are complementary from nucleotide i up to and including j, and $\beta = 1.59$ mol/kcal (corresponding to the experimental temperature of 45 °C). For a given gene, the measure of the gene expression level that one wants to determine is the total target concentration c given by

$$c = [t] + \sum_{i,j} [t\hat{t}_{i,j}] \tag{4}$$

Solving eqs 4 and 3 we find for the fraction of single-stranded target in solution

$$\alpha_f = \frac{[t]}{c} = \frac{1}{1 + \sum_{i,j} [\hat{t}_{i,j}] \exp(\beta \Delta G_R(i,j))} \tag{5}$$

Note that the summation in the denominator of eq 5 was replaced in the approximate expression eq 2 by the single term $\tilde{c} \exp(\beta' \Delta G_R^{(37)})$, with fitting parameters $\tilde{c}$ and β'.

Equation 5 requires as input estimates of the concentration $[\hat{t}_{i,j}]$ of complementary sequences with lengths $l = j - i + 1$ present in solution. Assuming that all four nucleotides are roughly equally abundant and that there are no correlations along the sequence, the abundance of short sequences with length l will decrease as $[\hat{t}_{i,j}] \approx 4^{-l}$. This scaling breaks down beyond some length L; assuming for the human transscriptome a total length of 10^7 nucleotides, a random sequence longer than 12 is more likely not present at all, since $4^{12} > 10^7$. We therefore take as our approximation

$$[\hat{t}_{i,j}] = \begin{cases} c_0 \cdot 4^{-(j-i)} & \text{for } j - i < 12 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Here, c_0 is a measure of the RNA concentration. Using this approximation for the concentration of complementary strands, we can now compare eqs 2 and 5. Figure 2 shows the more elaborate model eq 5 as a function of the approximate form eq 2, with the values for the fitting parameters β' and $\tilde{c}$ taken from ref 7. There is a reasonable agreement between the two.

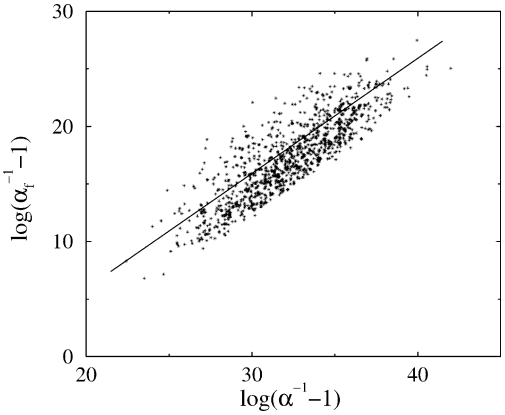

Figure 2. Comparison of the summation in eq 5, equal to $\alpha_f^{-1} - 1$, and its approximation in eq 2, equal to $\alpha^{-1} - 1$, for the first 1000 spike-in sequences of HGU133. Note that a change in c_0 corresponds to a vertical shift over $\log(c_0)$; in this figure, we used $c_0 = 1$. The straight line is a fit given by $y = x + b$ with $b = -14.1$.

Since eq 5 has a better microscopic foundation than eq 2, it should in principle allow for a better estimate of the hybridization in solution. There are however severe limitations to the use of eq 5. In the hybridization in solution, there is a competition between the contributions of short sequences, which are abundant but have a low affinity, and long sequences, for which the concentration is low but the affinity is high. The concentration drops on average approximately by a factor of 4 per added length (eq 6), but the affinity grows by approximately $\langle \Delta G \rangle \approx 2$ or 3 kcal/mol, the average value of RNA/RNA interaction parameters.[25] Since $\exp(\beta \langle \Delta G \rangle) > 4$, the longer sequences dominate the hybridization in solution. However, as discussed above, beyond length $L \approx 12$, there simply are no complementary strands. The accuracy of the more elaborate model eq 5 thus hinges crucially on knowing the longest complementary strand that is transcribed as well as its affinity and its concentration. Since the approximate model eq 2 is not expected to perform worse than the more elaborate model eq 5, we keep using the former.

The data points in Figure 2 can be fitted by a straight line with slope 1: The value of $\beta' = 0.67$ mol/kcal in ref 7, corresponding to 725 K, apparently is the appropriate value to describe the experiments at a temperature of 45 °C. The offset in the straight-line fit is equal to $\log(\tilde{c}) - \log(c_0)$. Since the straight-line fit has an offset of -14.1 and since we used the fitted value of $\tilde{c} = 2 \times 10^{-2}$ pM in ref 7, an estimate of the RNA concentration is $c_0 = \tilde{c} \exp(14.1) = 30$ nM. Even if we do not use the more elaborate model eq 5, it provides us with

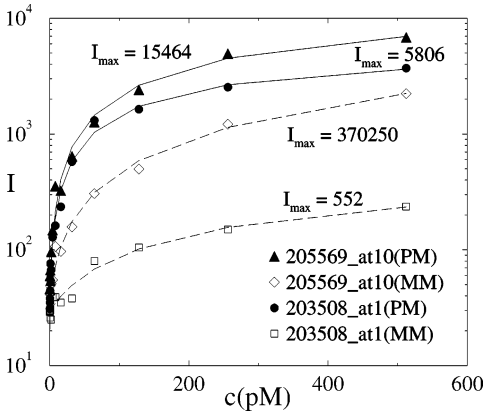

Figure 3. Plot of intensity vs concentration for three spike-in genes of the HGU133 chipset. I_{max} indicates the saturation value obtained from a nonlinear fit with three parameters (I_0, A, and K), based on eq 8.

a microscopic basis for the values of the parameters β' and $\tilde{c}$ in the approximate model eq 2.

4. Signal Saturation Level

If the target concentration c and the binding energy ΔG are sufficiently high, then the Langmuir isotherm saturates to a maximal value. From eq 1 we find for $c \exp(\beta\Delta G) \gg 1$

$$I_{max} = I_0 + A \approx A \tag{7}$$

where we have used the fact that typically the background level, I_0, is much lower than the value of A. The saturation intensity increases if targets are bound to almost all probes. Since the number of probes does not vary between the sequences being measured, this saturation intensity is also expected to be sequence-independent and, more specifically, should not distinguish between perfect matches and mismatches. A recent analysis of the Latin square set[9,14] reported widely different values for the saturation intensity. It is worth clarifying further this issue here.

The obvious procedure to determine the saturation intensity is to look at the intensity of a probe as a function of concentration. Assuming an effective affinity K_s for probe sequence s, the intensity $I_s(c)$ as a function of concentration c is given by

$$I_s(c) = I_{0,s} + \frac{A_s c K_s}{1 + cK_s} \tag{8}$$

in which $I_{0,s}$ is the (sequence-dependent) background intensity due to nonspecific binding. A plot of I_s versus c for two probes of the HGU133 spike-in set is shown in Figure 3. Taking I_0, A, and K in eq 8 as fitting parameters and extrapolating to high concentration then yield the saturation intensity.

Two research groups[9,14] followed this procedure, and both found saturation intensities that vary wildly between different sequences. A first effect that can cause deviations from the Langmuir fit in eq 8 is that the lithographic process, through which the probes are synthesized in situ in Affymetrix chips, is not 100% efficient. As estimated by Forman et al.,[26] only approximately 10% of the probes reach the full length of 25 nucleotides. At low intensities far from saturation, the incomplete probes can be safely ignored since their affinity is much lower than that of the fully grown probes. However, under conditions where the fully grown probes are saturated, clearly there will be contributions to the fluorescent intensity from the

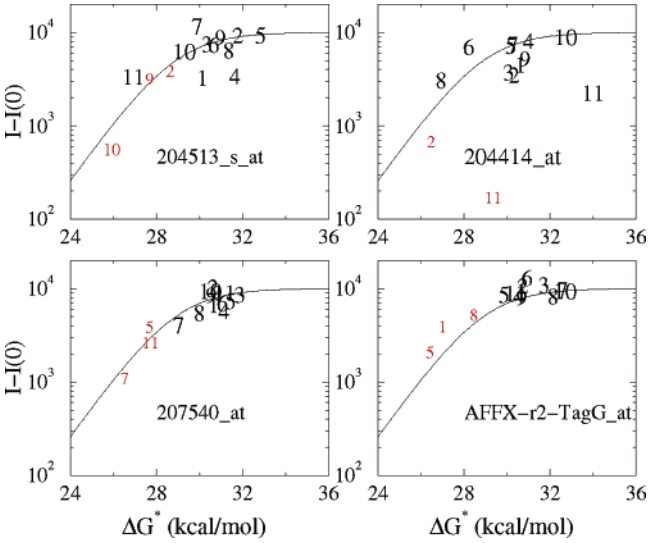

Figure 4. Plot of $I - I_0$ as a function of $\Delta G - RT \log \alpha$ for four sequences spiked-in at a concentration of $c = 512$ pM. The numbers indicate the probe set numbers. Smaller characters are used for the MM signals. Solid lines represent the Langmuir model as given by eq 2. The data are consistent, except a few outliers, with the Langmuir model with a roughly constant saturation level $A \approx 10^4$.

almost complete probes, and an even further increase in concentration will bring into play shorter and shorter incomplete probes. Consequently, the Langmuir fit in eq 8 breaks down near saturation; extrapolation to high concentration is an unreliable procedure.

A second cause of worry is that comparing fluorescent intensities from different chips is also potentially unreliable, since the microarrays might have undergone slightly different processing during the washing and staining. Since Affymetrix microarrays cannot be reused, the spike-in measurements used in refs 9 and 14 required a new chip for each concentration.

To avoid these two potential sources of error, we therefore consider the intensities for a given probe set at a specific concentration, i.e., constant c and variables ΔG and α in eq 1. The data belong to the same array. An example of this type of analysis is shown in Figure 4 for a concentration of $c = 512$ pM. On the horizontal axis we plot $\Delta G^* = \Delta G - RT \log \alpha$. The solid lines are given by the Langmuir curve in eq 1. Note that a large majority of the probes align along the expected curve, with a few exceptions as, for instance, for probe 11 (both PM and MM) for the probe set 204414_at. Therefore, the data are consistent with a value of A roughly constant in eq 1, which suggests indeed that the large variations in I_{max} obtained from the extrapolations of the data in the earlier analysis are more likely to be an artifact of the extrapolations. Note however that some variability of the saturation level can be seen in the data of Figure 4. Typically this variability is approximately 20%. To keep our model simple we will keep A constant in the rest of the paper. An interesting possible explanation of the variability of A has been given in ref 14, i.e., that this variation is due to the posthybridization washing of the array.

Yet another different way of addressing the issue of the saturation intensities is to analyze the histogram of the intensities on the whole chip, as in Figure 5, which shows both the intensities for the HGU95 and HGU133 spike-in data. To reveal the data at high intensities, they are plotted in a log−log scale. In the figure we note a decrease in the histogram around $I \approx 10\,000$, sharper in the HGU133 chipset, which is consistent with the estimate of the saturation intensity obtained from the fits of intensities versus $\Delta G - RT \log \alpha$, as given in Figure 4. Note

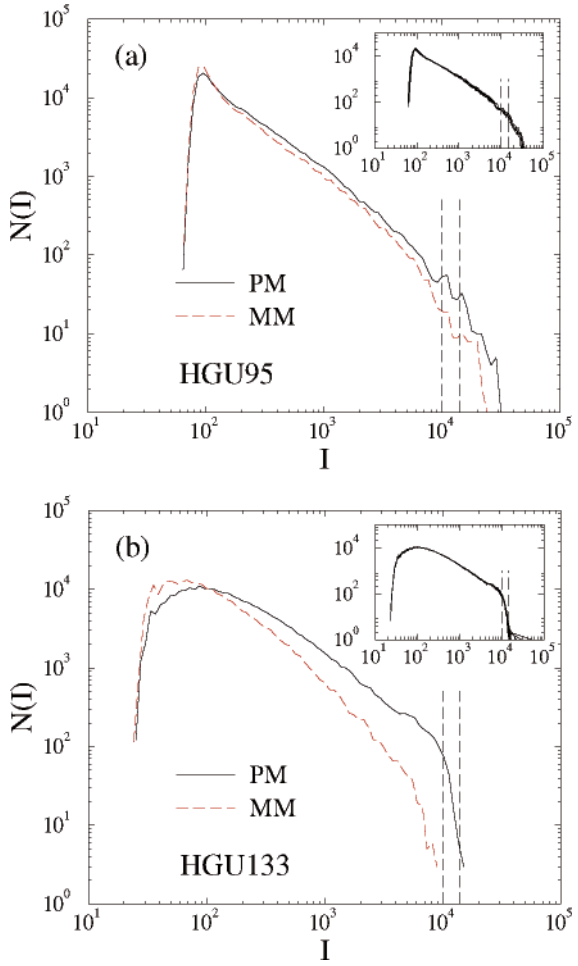

Figure 5. Histograms of the PM and MM intensities for the Latin square experiments in log−log scale for the chips (a) HGU95a and (b) HGU133. The plots contain (a) 19 and (b) 12 histograms, referring to different experiments. The dashed lines are positioned at $I = 10\,000$ and $I = 15\,000$. (Intensities are given on the Affymetrix scale.) Insets: Histograms of the total intensity of PM and MM together.

that in Figure 5b the decrease is 100-fold in the range $10\,000 < I < 15\,000$, which suggests that the data are consistent with a roughly constant value of the saturation. However, a closer inspection of the histogram of the HGU133 for PM and MM intensities separately reveals that the estimated saturation values for the two may be different. In the case of PM intensities alone the drop is rather sharp at $I \approx 10\,000$; however the MM intensities seem to saturate at lower intensities, a feature that is not seen in the HGU95 data (Figure 5a). The number of MM probes reaching an intensity close to the saturation level in the histogram of Figure 5b is quite small, so the fact that the MM and PM reach a different saturation level cannot be concluded for sure.

Also the low-intensity side of the histograms in Figure 5 contains interesting information. For both the HGU95 and the HGU133, the intensity drops steeply below a minimal intensity. For HGU95, this drop occurs at $I_{min} \approx 70$, while for HGU133 the drop occurs at $I_{min} \approx 30$. This increase of the dynamic intensity range by more than a factor of 2 is a clear demonstration of the fast rate of improvement in microarray technology.

5. Analysis of Data Collapses

As a test of the validity of the model we plotted[7] the data as a function of the rescaled variable

$$x' = \alpha c\, e^{\beta \Delta G} \tag{9}$$

If the model is to be trusted, then the data for different values of c and different probe sequences (i.e., different ΔG and α) ought to "collapse" onto a single master curve

$$I - I_0 = \frac{A x'}{1 + x'} \tag{10}$$

This collapse has indeed been observed in the large majority of the spike-in genes of the HGU95a chipset.[7] Interestingly, the very few outliers observed in that case could be explained as annotation errors or unbalance of free energies used for specific nucleotides, as discussed in ref 7.

We choose here the same fitting parameters used in ref 7 for the HGU95 chipset, that is: $A = 10\,000$, $\beta = 0.74$ mol/kcal, $\beta' = 0.67$ mol/kcal, and $\tilde{c} = 10^{-2}$ pM. These parameters fit equally well the HGU133 spike-in data.

In Figures 6−8 we show the collapse plots for all 42 genes of the spike-in data set HGU133. Each plot contains approximately 200 points, which all tend to cluster along the Langmuir curve $Ax'/(1 + x')$ (in some cases much better than others). All 13 concentrations, which range from 0.125 to 512 pM in the spike-in experiment, are shown. The intensities measured at $c = 0$ are taken as estimates of the background level I_0 in eq 10. In the collapse plots, only the MM sequences for which ΔG could be estimated are shown, as the mismatch free energies in RNA/DNA duplexes are known only for a limited set of mismatches.[22] (We could associate a free energy to approximately 30% of the mismatches, as discussed in ref 7.)

The HGU133 spike-in set contains 4 bacterial sequences and 8 artificial sequences (Figure 6) and 30 human sequences (Figures 7 and 8). A perfect agreement with the Langmuir theory would imply that all data align along the curve given by eq 10, which is shown as a solid line in Figures 6−8. In general the agreement is best for the artificial sequences. Occasionally, also some human sequences collapse well into a single curve in good agreement with the Langmuir model, but in general their behavior is worse than that of the artificial ones. To quantify the data dispersion we introduce the variable

$$w = \log\left(\frac{I}{I_{th}}\right) \tag{11}$$

where I is the measured intensity and I_{th} is the theoretical value as predicted from the Langmuir isotherm (eq 10) for x' corresponding to the measured I. For the definition of w in eq 11 we have kept only the values of I in the range $100 < I < 10000$. We determine its average $\langle w \rangle$ and standard deviation σ_w. If the data are well-centered around the expected behavior, then one has $\langle w \rangle = 0$, while σ_w is a measure of the spread in the data.

The values of $\langle w \rangle$ and σ_w for the bacterial, artificial, and human sequences are given in Tables 1 and 2, respectively. We note that σ_w is on average the lowest for the artificial sequences with a typical value of $\sigma_w \approx 1$. Only for two human probe sets (205790_at and 207540_s_at with $\sigma_w \approx 0.7$) the collapse is better than that of the artificial sequences. For three human probe sets (204205_at, 207641_at, and 212827_at) the collapse is very poor as indicated by a $\sigma_w > 2$. The collapses in the four bacterial sequences have a somewhat higher dispersion compared to the human sequences.

A very interesting feature of the whole analysis is that the quality of the collapses is much better for artificial sequences

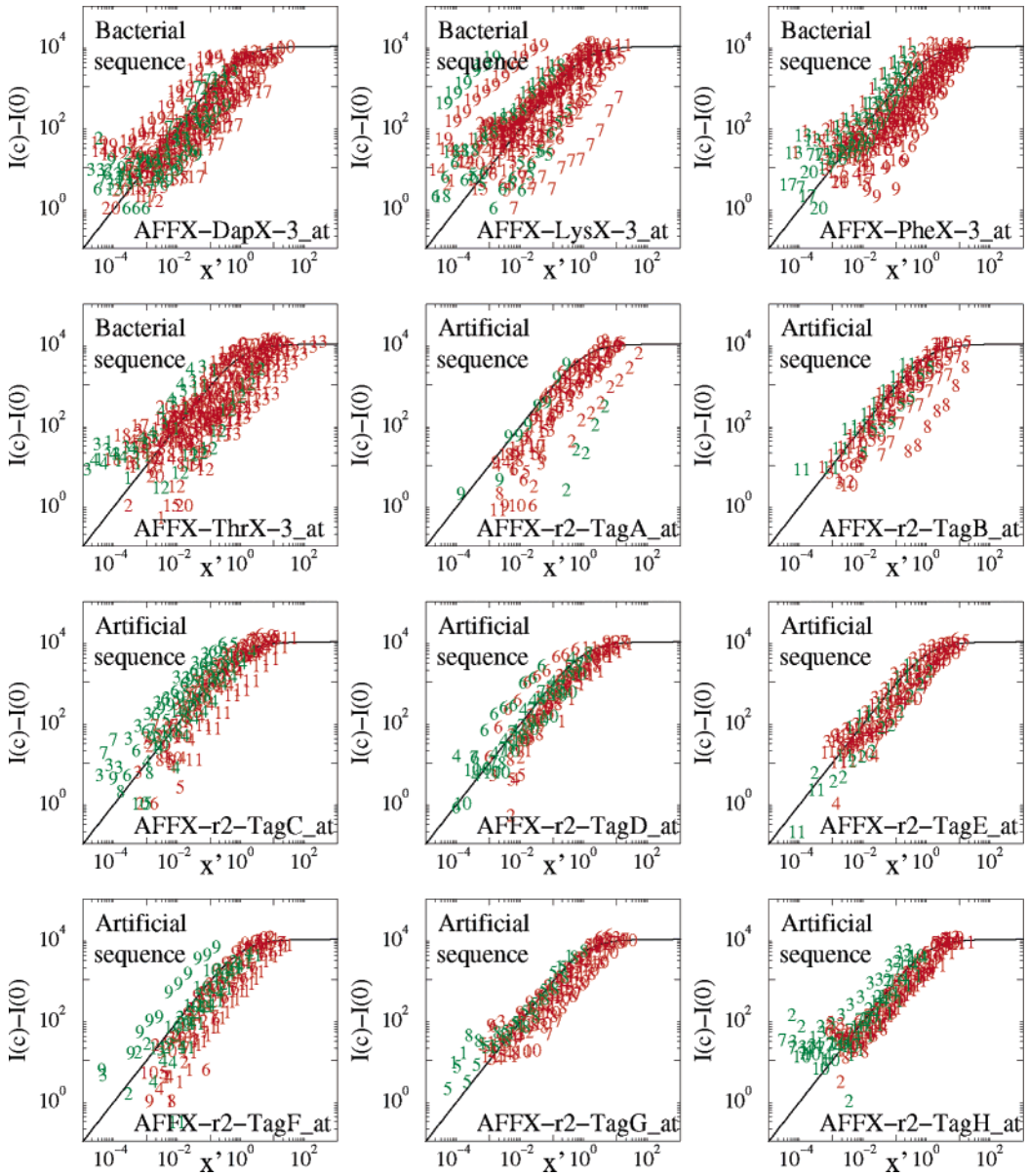

Figure 6. Collapse plots for the four bacterial and the eight artificial sequences of the HGU133 spike-in set. In these plots the background-subtracted intensities for a given probe set are plotted as functions of the rescaled variable x' given in eq 9. The data correspond to all spike-in concentrations for a given probe sets. Solid lines correspond to the Langmuir isotherm. In comparison with the human and bacterial sequences the artificial sequences are characterized by the best collapses.

than for any other sequence. Artificial sequences have been chosen by Affymetrix to be as different as possible from any human RNA, to minimize the effects of cross-hybridization. Their preparation, as far as labeling and target fragmentation are concerned, is the same as that for all other spikes.[27] As in all collapses the same set of parameters is used. Therefore the high σ_w for some probe sets is very likely an indication that the selected probes are not yet optimal. Possible deviations from the theory are due to cross-hybridization.

6. Determination of the Expression Level

The model defined by eqs 1 and 2, once all parameters have been fixed, can be used to fit the concentration c starting from the measured intensities. The target concentration in solution is a measurement of the gene expression level, and it is the quantity one wants to compute from the raw microarray data. As the concentrations in the spike-in experiments are known, we can compare the known values with the fitted ones. Figure 9 shows a plot of fitted concentration versus spike-in concentration for the artificial sequences. We limit ourselves here to show the data for these sequences, but the trend is quite general and valid for other genes as well. The solid line in Figure 9 corresponds to a line $y = x$, which means perfect agreement between spike-in and fitted values. The two other lines correspond to $y = 2x$ and $y = x/2$, drawn as guides to the eye.

As shown in Figure 9, most of the data fall in the range between the two lines, except for the spikes TagA and TagF, which give a much lower fitted concentration. All the points follow approximately straight lines with slope 1, except for the highest spike-in concentrations, corresponding to 256 and 512 pM. This is due to the fact that at high concentrations many probes are very close to saturation.

We note also that the fitted concentrations are all systematically lower than the spike-in values, as most of the concentrations fall in the interval $[c_{\text{spike-in}}/2, c_{\text{spike-in}}]$. This is a consequence of our choice to use the fitting parameters from a

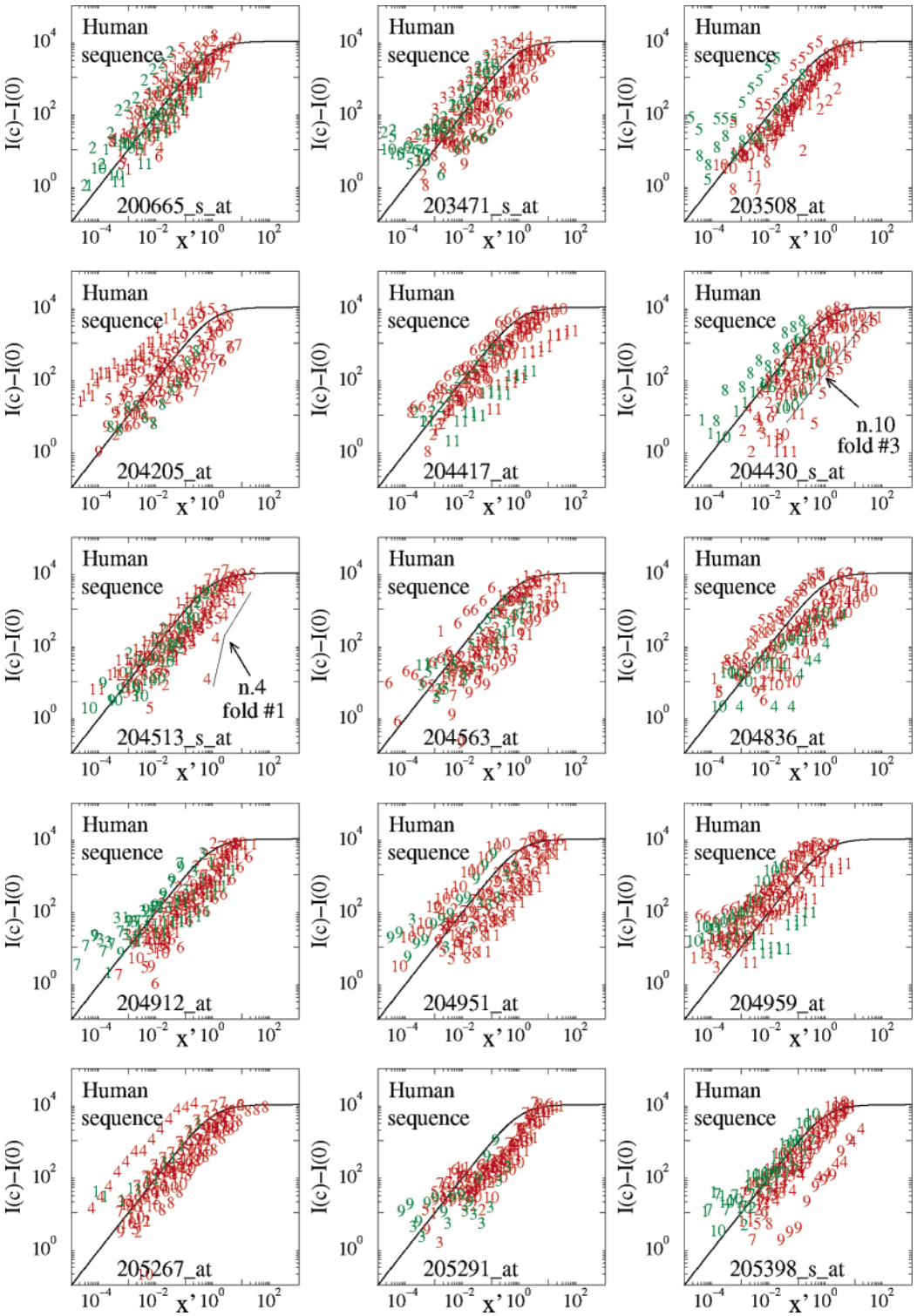

Figure 7. Collapse plots for human sequences of the HGU133 spike-in set (part 1). The probes that are complementary to targets with the largest folding free energies are emphasized (Table 3). They correspond to probes 204430_s_at10 and 204513_s_at4.

previous study[7] of spike-in experiments on HGU95. We have chosen not to refit these parameters here again for HGU133 to illustrate their universal validity. The slight underestimation of the absolute concentration is not a problem, since in gene expression measurements one is only interested in fold variations of expression levels between different experimental conditions. The fact that the data of Figure 9 follow lines with a slope of approximately 1 guarantees that the fold change in concentration in different experiments is correctly estimated.

7. One Cause of Outliers: Target Secondary Structures

It is well-known that single-stranded nucleic acids, particularly RNA, tend to form stable folded conformations by binding of complementary bases. Currently, algorithms that calculate RNA secondary structures are to be trusted for sufficiently short molecules, say less than 50 nucleotides, which is the situation for Affymetrix microarrays, where RNA targets are fragmented before hybridization. The average target length is 50 nucleotides, but probably only shorter fragments contribute to hybridization.

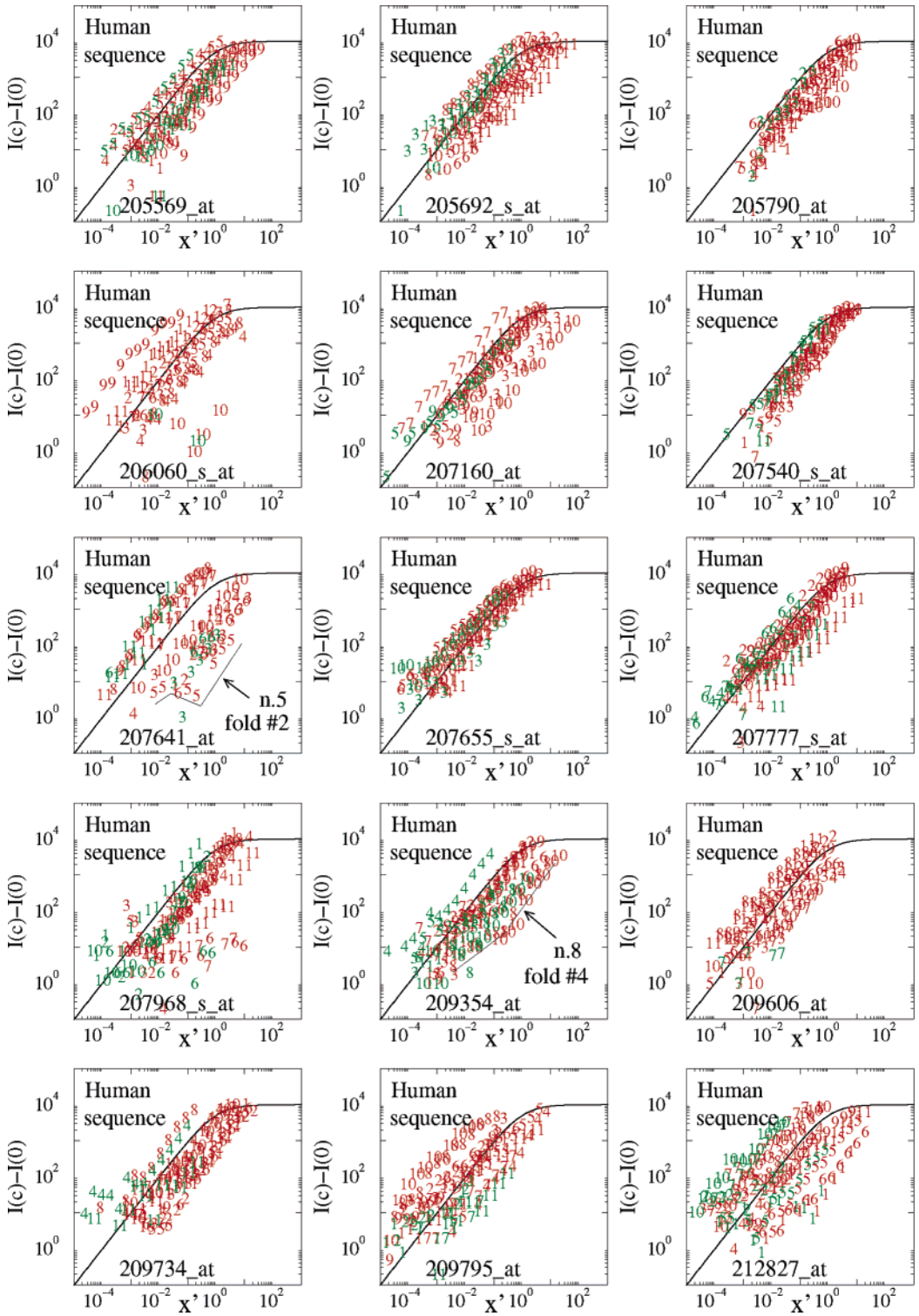

Figure 8. Collapse plots for human sequences of the HGU133 spike-in set (part 2). The probes that are complementary to targets with the largest folding free energies are emphasized (Table 3). They correspond to probes 207641_at5 and 209354_at8.

We used the Vienna package[28] for the calculation of folded RNA structures that may form in solution and impede hybridization. We considered first 25-mer targets in solution exactly complementary to the probes of the HGU133 spike-in data set. Table 3 shows a list of probes in this set, whose complementary target has the lowest folding free energy, i.e., that of the most stable conformation, calculated at the experimental temperature of 45 °C. Given a folding free energy ΔG_{fold}, one can use the

TABLE 1: List of Values of $\langle w \rangle$ and σ_w for the Bacterial and the Artificial Sequences in the Spike-In Set HGU133.

probe set	$\langle w \rangle$	σ_w	probe set	$\langle w \rangle$	σ_w
AFFX-DapX-3_at	0.08	1.49	AFFX-PheX-3_at	0.16	1.55
AFFX-LysX-3_at	0.89	2.46	AFFX-ThrX-3_at	0.22	1.59
AFFX-r2-TagA_at	−1.05	0.97	AFFX-r2-TagE_at	−0.32	0.82
AFFX-r2-TagB_at	−0.51	0.83	AFFX-r2-TagF_at	−0.46	1.09
AFFX-r2-TagC_at	0.43	1.08	AFFX-r2-TagG_at	−0.11	0.90
AFFX-r2-TagD_at	−0.03	0.90	AFFX-r2-TagH_at	0.11	1.22

TABLE 2: List of Values of $\langle w \rangle$ and σ_w for the Human Sequences in the Spike-In Set HGU133.

probe set	$\langle w \rangle$	σ_w	probe set	$\langle w \rangle$	σ_w
200665_s_at	0.54	1.26	205569_at	−0.28	1.12
203471_s_at	0.39	1.43	205692_s_at	0.24	1.27
203508_at	0.45	1.83	205790_at	−0.78	0.76
204205_at	0.86	2.11	206060_s_at	0.52	1.66
204417_at	−0.24	1.18	207160_at	−0.32	1.06
204430_s_at	−0.48	1.13	207540_s_at	−0.29	0.62
204513_s_at	−0.68	1.16	207641_at	0.24	2.72
204563_at	−0.57	1.44	207655_s_at	0.76	1.06
204836_at	−0.04	1.41	207777_s_at	−0.14	1.11
204912_at	−0.31	1.35	207968_s_at	−0.85	1.66
204951_at	−0.15	1.48	209354_at	0.04	1.41
204959_at	1.33	1.62	209606_at	0.77	1.44
205267_at	0.36	1.23	209734_at	−0.20	1.51
205291_at	−0.44	1.24	209795_at	0.63	1.71
205398_s_at	−0.15	1.37	212827_at	0.61	2.53

TABLE 3: Minimal Folding Free Energies for the Targets (Assumed to be 25-mers) Complementary to the Probes Forming the Spike-In HGU133 Data Set[a]

probe set	probe number	$-\Delta G_{\text{fold}}$ (kcal/mol)
204513_s_at	4	8.70
207641_at	5	8.16
204430_s_at	10	7.79
209354_at	8	7.67
207540_s_at	10	7.45
AFFX-r2-TagA_at	1	6.52
205398_s_at	1	6.43
AFFX-PheX-3_at	10	6.18
204836_at	10	6.17
203508_at	2	6.10
206060_s_at	3	6.05

[a] These free energies were calculated with the program RNAfold.

two-state model approximation to find p_{fold}, the probability that the sequence is folded into the most stable conformation

$$p_{\text{fold}} = \frac{e^{-\Delta G_{\text{fold}}/RT}}{1 + e^{-\Delta G_{\text{fold}}/RT}} \tag{12}$$

where we use $T = 45°$ C. According to this expression for a folding free energy $\Delta G_{\text{fold}} = -8$ kcal/mol, one finds $1 - p_{\text{fold}} \approx 4 \times 10^{-6}$, and for $\Delta G_{\text{fold}} = -6$ kcal/mol one finds $1 - p_{\text{fold}} \approx 10^{-4}$. The large majority of the targets complementary to the probes listed in Table 3 are thus folded and not expected to participate to hybridization.

Figure 10 shows the folding configurations for the four targets with the lowest free energies in Table 3. As shown in Figures 7 and 8 the corresponding probes have a signal that is a few orders of magnitude lower than that expected from the Langmuir model, although not as low as that derived from eq 12, using the ΔG_{fold} listed in Table 3. For instance, from the measured signals we find an intensity lower by a factor 10^3 for the probe 204513_s_at4 instead of a factor 10^6 as deduced from eq 12. This difference could have several origins. First, the hybridization in solution described by the term α in eq 2 may already take into account some secondary structure formation. Second, the RNA in solution is present with sequences of all lengths. The free energies listed in Table 3 refer to 25-mers, so shorter sequences will have a lower folding probability than that deduced from eq 12 on the basis of the free energies of 25-mers. Third, even if some secondary structure is present, hybridization with the surface-bound probes is still possible if the folded configuration has some dangling ends from which binding can initiate.

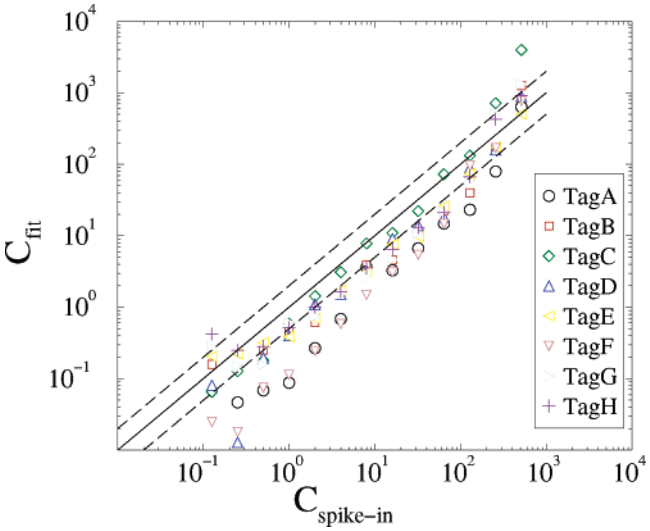

Figure 9. Plot of the fitted target concentration as a function of the spike-in concentration for the artificial sequences. The solid line corresponds to the diagonal $y = x$, while the two dotted lines are $y = x/2$ and $y = 2x$ and are drawn as guides to the eye. We note a systematic shift of the estimated absolute concentration compared to that of the spike-in one, although the fold variations of the concentrations are correctly estimated as the majority of the data follow lines parallel to the diagonal in the plot.

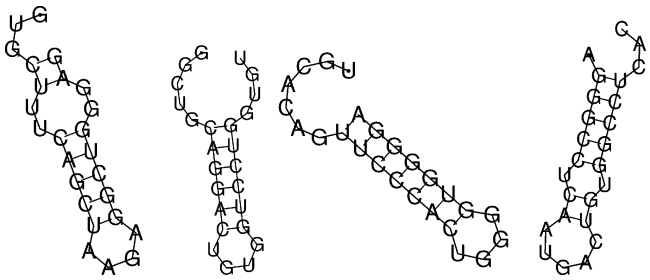

Figure 10. Folding configurations for the four targets with the lowest free energy, from left to right: 204513_s_at4, 207641_at5, 204430_s_at10, and 209354_at8.

We have analyzed the folding free energies of 25-mers complementary to all of the probes in the HGU spike-in set. We found that approximately 50% of the targets have a folding free energy lower than 1 kcal/mol, so secondary structure formation can be neglected safely. Approximately 10% of the targets have a folding free energy higher than 4 kcal/mol, so for this fraction the secondary structure formation may interfere with the target−probe hybridization.

The correct estimate of the folding probability involves a complex calculation over fragments of all lengths, possibly including sequences neighboring the 25-mer part complementary to the probe. However the folding is expected to have a relevant effect for at most 10% of the probes. A possible way out is that of excluding from the analysis of the gene expression levels those probes whose 25-mers folding free energy is above a certain threshold.

8. Conclusion

In this paper we have extended a previous study[7] of Affymetrix spike-in experiments on the chip HGU95 to a novel HGU133 chipset. We used the model introduced in ref 7 that takes into account both target−probe and target−target hybridization in solution. The hybridization free energies are calculated from the nearest-neighbor model[25] using the experimental parameters for RNA/DNA[21,22] and RNA/RNA.[23] There are four

global fitting parameters in the model that we took from ref 7. We found that these parameters also fit well the current data on the HGU133 chipset, apart from a systematic small shift of all the estimates of the absolute target concentrations.

There are several features that make the spike-in data of the more recent HGU133 chip interesting. First of all the spike-in set contains a larger number of sequences compared to the HGU95 experiments (42 instead of 14), and the chip has been entirely redesigned. Second, the spike-in sequences contain some of artificial origin, designed to avoid any cross-hybridization with human RNAs but prepared and labeled exactly as all other spikes. We find that these artificial sequences best fit the hybridization model, as they show the best collapses when the data are rescaled and plotted as a function of an appropriate thermodynamic variable. The good agreement suggests indeed that the simple model describes rather well the hybridization in Affymetrix arrays and that the deviations observed for some human sequences are probably related to the nonoptimal design of the sequences for a given probe.

When compared to the human sequences of the HGU95 spike-in experiments analyzed in ref 7, we find that the artificial spikes of the HGU133 set show definitely better collapses. However, when comparing the human sequences of the HGU133 with those in the HGU95 experiment, we find on average a better collapse for the latter. Only few probes out of the 32 human spikes of the HGU133 experiment have a better collapse than those of the HGU95.

Interestingly, the physics-based modeling developed here allows the assignment to each probe set of a quality score based on the level of agreement with the Langmuir model. This information may be used to reconsider and eventually redesign the low-quality probe sets.

Finally, we have discussed the physical basis of hybridization in solution and of RNA secondary structure formation. The latter effect, according to the statistics over the spike-in probes, will be relevant for approximately 10% of the probes only. The sequences with the highest folding probabilities correspond to probes whose measured fluorescent intensities are well below those predicted from the Langmuir model.

According to our current understanding of the system (see also refs 7 and 24), the hybridization in solution of partially complementary RNA molecules has a strong influence. One of the reasons for that is that RNA/RNA interaction parameters are, at given temperature and salt concentration, stronger than the DNA/DNA or RNA/DNA parameters. The simple approximation given in eq 2 captures the major features of the hybridization in solution. However, an improvement over this approach, as discussed above, remains an open challenge.

Acknowledgment. We acknowledge financial support from the Van Gogh Programme d'Actions Intégrées 08505PB of the French Ministry of Foreign Affairs and Grant No. 62403735 by the Netherlands Organization for Scientific Research (NWO).

References and Notes

(1) Lockhart, D. J.; Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **2000**, *405*, 827.

(2) Heller, M. J. DNA Microarray technology: Devices, systems, and applications. *Annu. Rev. Biomed. Eng.* **2002**, *4*, 129.

(3) Marshall, E. Getting the noise out of gene arrays. *Science* **2004**, *306*, 630.

(4) Tan, P. K.; Downey, T. J.; Spitznagel, E. L., Jr.; Xu, P.; Fu, D.; Dimitrov, D. S.; Lempicki, R. A.; Raaka, B. M.; Cam, M. C. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **2003**, *31*, 5676.

(5) Brown, P. O.; Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature* **1999**, *21*, 33.

(6) Lipshutz, R. J.; Fodor, S. P. A.; Gingeras, T. R.; Lockhart, D. J. High-density synthetic oligonucleotide arrays. *Nat. Genet.* **1999**, *21*, 20.

(7) Carlon, E.; Heim, T. Thermodynamics of RNA/DNA hybridization in high-density oligonucleotide microarrays. *Physica A* **2006**, *362*, 433.

(8) Vainrub, A.; Pettitt, B. M. Coulomb blockage of hybridization in two-dimensional arrays. *Phys. Rev. E* **2002**, *66*, 041905.

(9) Held, G. A.; Grinstein, G.; Tu, Y. Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7575.

(10) Naef, F.; Magnasco, M. O. Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E* **2003**, *68*, 011906.

(11) Hagan, M. F.; Chakraborty, A. K. Hybridization dynamics of surface immobilized DNA. *J. Chem. Phys.* **2004**, *120*, 4958.

(12) Halperin, A.; Buhot, A.; Zhulina, E. B. Sensitivity, specificity, and the hybridization isotherms of DNA chips. *Biophys. J.* **2004**, *86*, 718.

(13) Binder, H.; Preibisch, S. Specific and nonspecific hybridization of oligonucleotide probes on microarrays. *Biophys. J.* **2005**, *89*, 337.

(14) Burden, C. J.; Pittelkow, Y.; Wilson, S. R. An adsorption model of hybridization behaviour on oligonucleotide microarrays. *J. Phys.: Condens. Matter* **2006**, *18*, 5545.

(15) Levicky, R.; Hogan, A. Physicochemical perspectives on DNA microarray and biosensor technologies. *Trends Biotechnol.* **2005**, *23*, 143.

(16) Halperin, A.; Buhot, A.; Zhulina, E. B. On the hybridization isotherms of DNA microarrays: The Langmuir model and its extensions. *J. Phys.: Condens. Matter* **2006**, *18*, S463.

(17) Binder, H.; Kirsten, T.; Hofacker, I. L.; Stadler, P. F.; Loeffer, M. Interactions in oligonucleotide hybrid duplexes on microarrays. *J. Phys. Chem. B* **2004**, *108*, 18015.

(18) Carlon, E.; Heim, T.; Klein Wolterink, J.; Barkema, G. T. Comment on: "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays" by F. Naef and M. Magnasco. *Phys. Rev. E* **2006**, *73*, 063901.

(19) Naef, F.; Wijnen, H.; Magnasco, M. Reply to Comment on: "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays". *Phys. Rev. E* **2006**, *73*, 063902.

(20) Binder, H. Thermodynamics of competitive surface adsorption on DNA microarrays. *J. Phys.: Condens. Matter* **2006**, *18*, S491.

(21) Sugimoto, N.; Nakano, S.; Katoh, M.; Matsumura, A.; Nakamuta, H.; Ohmichi, T.; Yoneyama, M.; Sasaki, M. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **1995**, *34*, 11211.

(22) Sugimoto, N.; Nakano, M.; Nakano, S. Thermodynamics−structure relationship of single mismatches in RNA/DNA duplexes. *Biochemistry* **2000**, *39*, 11270.

(23) Xia, T.; SantaLucia, J., Jr.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson−Crick base pairs. *Biochemistry* **1998**, *37*, 14719.

(24) Heim, T.; Klein Wolterink, J.; Carlon, E.; Barkema, G. T. Effective affinities on microarray data. *J. Phys.: Condens. Matter* **2006**, *18*, S525.

(25) Bloomfield, V. A.; Crothers, D. M.; Tinoco, I., Jr. *Nucleic Acids Structures, Properties and Functions*; University Science Books: Sausalito, CA, 2000.

(26) Forman, J. E.; Walton, I. D.; Stern, D.; Rava, R. P.; Trulson, M. O. Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesized oligonucleotide arrays. In *Molecular Modeling of Nucleic Acids*; Leontis, N. B., SantaLucia, J., Jr., Eds.; ACS Symposium Series 682; American Chemical Society: Washington, DC, 1998; p 206.

(27) Affymetrix Europe. Private communication.

(28) Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **2003**, *31*, 3429.