Gene 422 (2008) 1–6

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

On the relationship between perfect matches and mismatches in Affymetrix Genechips

A. Ferrantini, E. Carlon *

Interdisciplinary Research Institute, Cité Scientifique BP 60069, F-59652, Villeneuve d'Ascq, France
Institute for Theoretical Physics, K.U.Leuven, Celestijnenlaan 200D, B-3001, Leuven, Belgium

ARTICLE INFO

Article history:
Received 14 September 2007
Accepted 29 November 2007
Available online 14 June 2008

Received by A. Bernardi

PACS:
87.15.-v
82.39.Pj

Keywords:
DNA microarrays
DNA hybridization
Mismatches
Langmuir isotherm

ABSTRACT

A relationship which links the fluorescence intensity histograms for perfect match (PM) and mismatch (MM) probes in Affymetrix Genechips is derived using inputs from physical-chemistry as the Langmuir and the Nearest Neighbor models. This relationship is in good agreement with experimental data from few dozens of chips belonging to about 10 different organisms. Principles of physical-chemistry impose some constant value for the average ratios of PM and MM intensities. Experimental data, however, show quite some variations in these parameters, although they follow the same inequalities as expected from hybridization free energies for oligonucleotides melting in solution. It is suggested that the anomalous experiment to experiment differences are due to 1) problems with chip design and 2) excessive fragmentation of the target in solution. The histogram analysis developed may be a useful preprocessing step to evaluate the global quality of the experimental data, prior to the calculation of the gene expression level.

1. Introduction

DNA microarrays have found, in recent years, an increasing number of applications beyond the gene expression studies for which they were originally introduced Schena et al. (1995). As examples we mention: the analysis of DNA mutations, of single nucleotide polymorphisms, of alternative splicing events, and of transcribed regions of whole chromosomes (for a recent review see Stoughton, 2005). All microarrays rely on the principle of hybridization of DNA or RNA strands in solution, the *targets*, with surface-bound strands, the *probes*. The latter may be spotted, or grown in situ through, for instance, photolitography as for Affymetrix Genechips (Lipshutz et al., 1999).

There is a vast literature reporting studies of the physico-chemical properties of nucleic acids hybridization in which both strands are free in solution (see Bloomfield et al., 2000 and references therein). Unfortunately these inputs have hardly been used for the analysis of DNA microarray data, which is mostly done with algorithms which are purely statistical (see Gentleman et al., 2003). Only a few DNA microarrays papers (Held et al., 2003; Naef and Magnasco, 2003; Binder and Preibisch, 2005; Halperin et al., 2006; Carlon and Heim, 2006) have been dedicated to the development of models and concepts relying on the underlying physical-chemistry of DNA hybridization.

In this paper physico-chemical models of nucleic acid hybridization are exploited for data analysis of Affymetrix microarrays. These chips are characterized by the presence of surface-bound probes which come in pairs: a probe (known as Perfect Match) whose sequence is perfectly complementary to the sequence in solution and a probe with a single internal mismatch. The behavior of MM probes has raised quite some debate in the literature (Naef and Magnasco, 2003; Carlon et al., 2006). They were introduced by Affymetrix with the purpose of estimating the contribution of cross-hybridization, i.e. the binding to the given probe of other sequences, which are only partially complementary to it. Affymetrix suggested to analyze the intensity difference $I_{PM}-I_{MM}$, assuming that I_{MM} is a measure of non-specific hybridization. Further analysis of test experiments showed that this is not the case: a single mismatch is not sufficient to destabilize hybridization with the target complementary to the PM sequence. For this reason most of the current algorithms for data analysis do not use the difference $I_{PM}-I_{MM}$. They instead employ different methods for estimating the non-specific hybridization using only PMs. This is somehow unfortunate as half of the experimental data from Affymetrix chips are then ignored.

We show here that MMs provide still some valuable information and can be exploited to test the global performance on a hybridization experiment. From principles of physical-chemistry we derive a rescaling expression relating histograms for PM and MM probes. We verify this scaling for about 50 chips on 10 different organisms and extract parameters which are found to be consistent with hybridization free energies in solution. We also briefly use some physical ideas to justify variations from experiment to experiments observed in

Abbreviations: MM, mismatch; PM, perfect match.

* Corresponding author. Institute for Theoretical Physics, K.U.Leuven, Celestijnenlaan 200D, B-3001 Leuven, Belgium. Tel.: +32 16 327239; fax: +32 16 327986.
E-mail address: enrico.carlon@fys.kuleuven.be (E. Carlon).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.06.013

some parameters. A general discussion of the results can be found in the final section of this paper.

2. Materials and methods

The affinity between two strands forming a double helix can be quantified by the so-called hybridization free energy ΔG. Given a sequence, the hybridization free energy between two nucleic acid strands in solution $\Delta G=\Delta H-T\Delta S$ (here ΔH and ΔS denote as usual enthalpies and entropies) can be calculated with the Nearest Neighbor model (Bloomfield et al., 2000). This model assumes additivity, i.e. both ΔH and ΔS are given as the sum of stacking terms depending on pairs of nucleotides. For instance, the enthalpy of hybridization of a stretch of helix can be calculated as:

$$\Delta H\begin{pmatrix}\mathrm{ATCC}\\ \mathrm{TAGG}\end{pmatrix}=\Delta H\begin{pmatrix}\mathrm{AT}\\ \mathrm{TA}\end{pmatrix}+\Delta H\begin{pmatrix}\mathrm{TC}\\ \mathrm{AG}\end{pmatrix}+\Delta H\begin{pmatrix}\mathrm{CC}\\ \mathrm{GG}\end{pmatrix} \quad (1)$$

(by convention the upper strand has orientation 5′–3′) where the parameters $\Delta H\binom{\mathrm{AT}}{\mathrm{TA}}$ … are tabulated. In Affymetrix expression chips, which we study here, the target strands are RNAs hence hybridization produces RNA/DNA duplexes. The enthalpy and entropy parameters of these duplexes were measured by Sugimoto et al. (1995). The same group considered also the case of RNA/DNA helices with one internal mismatch (Sugimoto et al., 2000). In this case the interaction was found to depend on the nature of the mismatch and on the identity of the two flanking nucleotides, i.e. one deals with triplets. For instance the parameters referring to GG mismatches

$$\Delta H\begin{pmatrix}\mathrm{dC\underline{G}G}\\ \mathrm{rG\underline{G}C}\end{pmatrix},\quad \Delta H\begin{pmatrix}\mathrm{dG\underline{G}G}\\ \mathrm{rC\underline{G}C}\end{pmatrix}\quad\text{and}\quad \Delta H\begin{pmatrix}\mathrm{dC\underline{G}C}\\ \mathrm{rG\underline{G}G}\end{pmatrix} \quad (2)$$

are all different, because of the differences in flanking nucleotides (here the labels *d* and *r* identify the DNA and the RNA strands). In Affymetrix Genechips probes are 25 nucleotides long. The MM probe sequence is obtained by swapping the 13th nucleotide of the PM probe sequence in the following way: C↔G and A↔T, which implies that there are 4 types of mismatches at the MM site dGrG, dCrC, dArA and dTrU. Hence, counting the two flanking nucleotides, there are in total 64 mismatch free energies. Unfortunately, only 11 of these 64 parameters have been measured in hybridization experiments (Sugimoto et al., 2000), and thus the comparison of our microarray analysis with the experimental free energies in solution can only be partial.

The physical-chemistry of hybridization in microarrays is described by the so-called Langmuir model (for a recent review see Halperin et al., 2006). In this model the intensity of the signal measured from a given probe is:

$$I=I_0+\frac{Ace^{\Delta G/RT}}{1+ce^{\Delta G/RT}} \quad (3)$$

where I_0 is a background term, A is a constant which sets the scale of intensities, c is the target concentration in solution, ΔG is the hybridization free energy discussed above, R is the universal gas constant and T the temperature.

Fig. 1 shows histograms in log–log scale of PM and MM intensities for two different experiments on *C. elegans* and *X. laevis* arrays. Note that PM and MM histograms overlap at the lowest intensities, while there is a higher count of MM probes for intensities in the range $I\lesssim 150$ and a higher number of PM probes with high intensities, as expected. The PM histogram shows saturation with a drop at $I\approx 10{,}000$. This is due to the effect of the denominator in Eq. (3): if a gene is highly expressed, therefore it is at high concentration in solution (high c), or if a probe is rich in CG content (high ΔG), one may have $ce^{\Delta G/RT}\gg 1$. In this limit, Eq. (3) reaches a saturating regime for which $I\approx I_0+A\approx A$ (as typically $I_0\ll$ A).

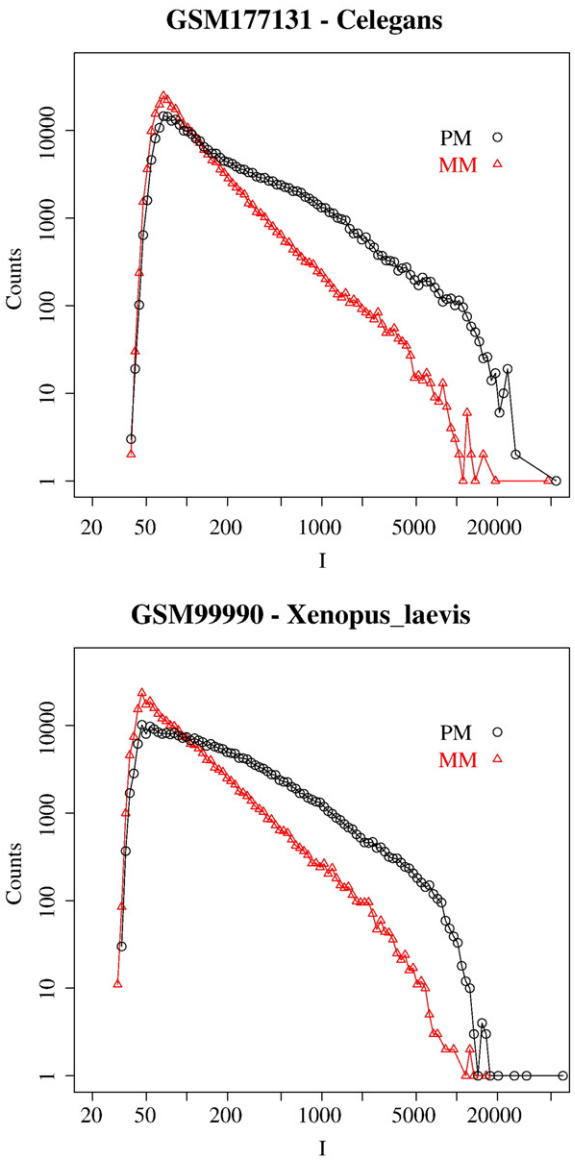

Fig. 1. Histograms of perfect match (PM) and Mismatch (MM) intensities in two Affymetrix Microarrays belonging to *C. elegans* and *X. laevis*.

From Fig. 1 we can estimate that only few hundred probes have intensities beyond the threshold $I=10{,}000$. We can therefore neglect the saturation regime and approximate the Langmuir isotherm to:

$$I\approx I_0+Ace^{\Delta G/RT} \quad (4)$$

As second assumption we take the background I_0 constant in the whole chip, but we allow it to vary from experiment to experiment. It is known however that I_0 is probe dependent and indeed several background subtraction schemes for Affymetrix Genechips have been developed (see e.g. Gentleman et al., 2003). A good background estimator is particularly important in the case one analyzes low intensity signals, i.e. low expressed genes. However, we will show that in our analysis of histograms the most important part is the intermediate intensity region. Hence our results are weakly sensitive to a precise background value and a constant background value suffices.

For a given PM/MM pair there is a difference in hybridization free energies. If we assign a free energy ΔG_{PM} for the PM, then the MM has

supposedly a smaller one, i.e. $\Delta G_{MM}=\Delta G_{PM}-\Delta G_0$. As each PM/MM pair differs only for a central nucleotide and as RNA/DNA mismatches in the Nearest Neighbor model are described by triplet free energies (see e.g. Eq. (2)), it is natural to expect that ΔG_0 depends only on the central triplet sequence. From Eq. (4) one has for a PM/MM pair of intensities:

$$I_{PM}-I_0 = e^{\Delta G_0/RT}(I_{MM}-I_0). \quad (5)$$

As ΔG_0 depends only on the three central nucleotides of the probe sequence, the previous relation can be generalized to histograms of PM and MM intensities having a fixed central triplet. We label with α=AAA, CAA, GAA … and indicate with $P^{\alpha}_{MM}(x)$ the histogram of background subtracted intensities for PM probes with a central triplet α. One has thus

$$P^{\alpha}_{MM}(I-I_0) = P^{\alpha}_{PM}(a_\alpha(I-I_0)), \quad (6)$$

where we have introduced the scaling factor $a_\alpha=e^{\Delta G_0/RT}$.

3. Results

3.1. Histogram rescaling for different central triplets

In order to test the validity of Eq. (6) we need an estimate of the background level I_0. In what follows we take for I_0 the intersection point of the global PM and MM histograms of Fig. 1. For each central triplet α the scaling parameter is estimated as:

$$a_\alpha = \left\langle \frac{I_{PM}-I_0}{I_{MM}-I_0} \right\rangle_m, \quad (7)$$

where as estimator we use the median $\langle . \rangle_m$, which was calculated for every probe pair that satisfies the inequalities $I_{PM}>I_0$ and $I_{MM}>I_0$. Note that the median (and not the mean) value is taken, since the distribution of the scaling factor a, deduced from Eq. (7) has an asymptotic long tail, generated by the intensities for which $I_{MM}\approx I_0$; these large a values make the mean value of the distribution unreliable.

Fig. 2 shows an example of four histograms with central triplets TAA, CTA, CCA and ATG. The triplets sequence refers to the DNA

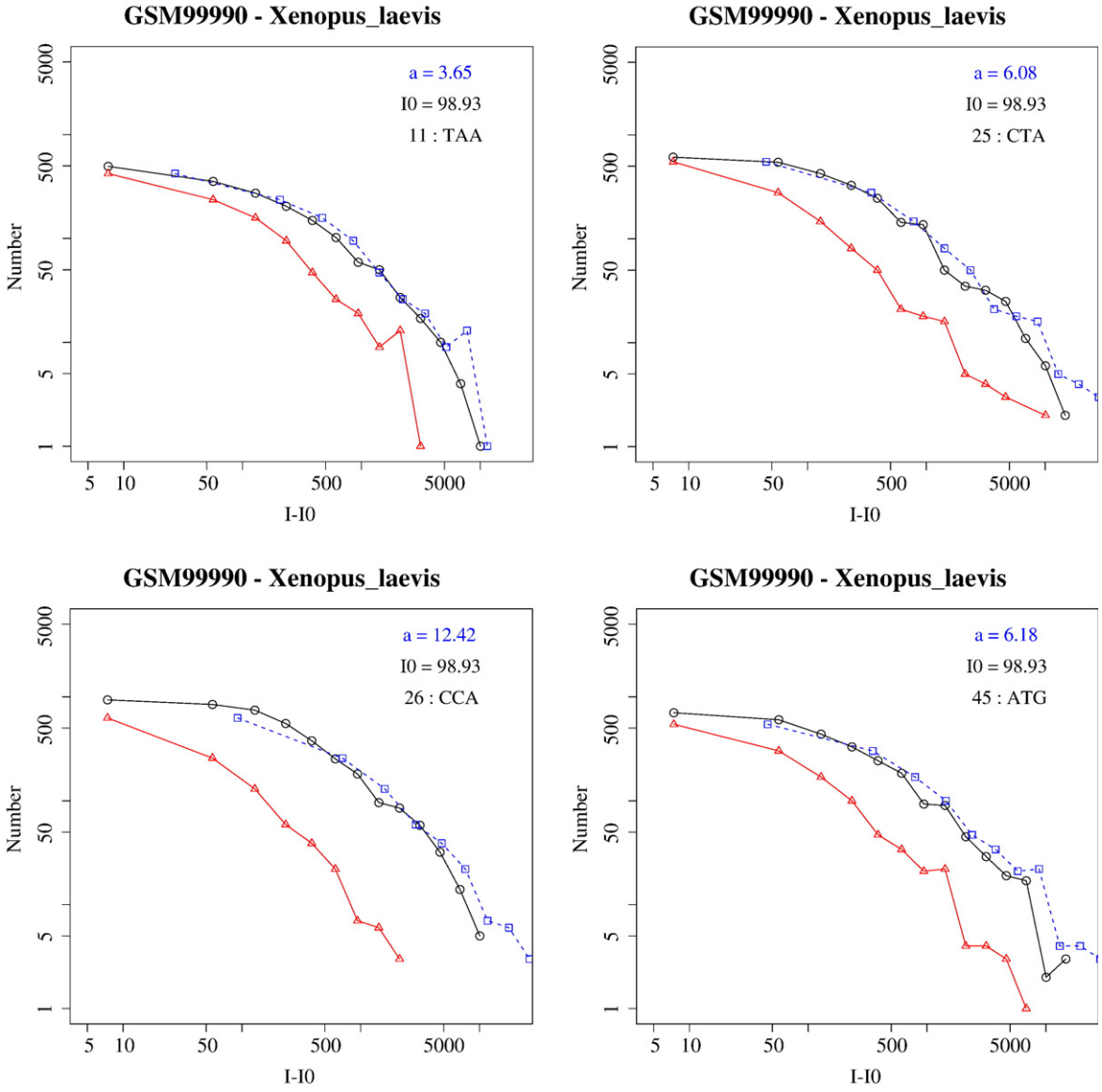

Fig. 2. Example of four histograms for probes with different central triplets. Circles and triangles refer to PM and MM histogram. The dashed line (squares) is obtained from the red one by multiplying the horizontal axis by a scaling factor a. The overlap between the solid lines (circles) and the dashed lines (squares) is a proof of the validity of Eq. (6).

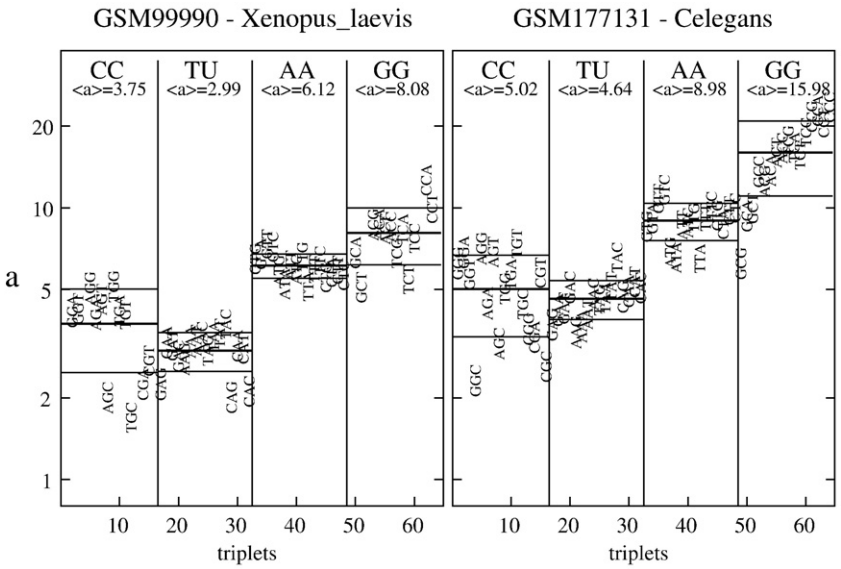

Fig. 3. Summary of the scaling factors for the 64 triplets in two different experiments. The data are organized in 4 subgroups of equal mismatch type: CC, TU, AA and GG. The central line in each subgroup is the average value within the subgroup (shown as $\langle a \rangle$ in the graphs). The two other lines cover one standard deviation. In almost all experiments analyzed, the average values and standard deviations within each subgroup satisfy the inequalities given in Eqs. (8) and (9).

sequences of the PM probe. Hence a, say, CCC triplet corresponds to a CGC triplet at the MM probe and thus, to a mismatch of GG type. In Fig. 2 circles and triangles are the original background subtracted histograms for PM and MM. The squares (and the dashed line) are obtained by multiplying the $I-I_0$ for the MM by the scaling factor a_α calculated from Eq. (7). In the log–log scale this amounts to a horizontal shift of the MM histograms. The overlap between the solid lines (circles) and the dashed lines (squares) is a direct verification of the validity of Eq. (6). Note that the calculated scaling factor a_α varies from triplet to triplet: for instance in Fig. 2 for the triplet TAA we find a_α=3.3 and for a triplet CCA a_α=13.1.

3.2. Global relations among the scaling factors

In Fig. 3 we summarize the results of the 64 different triplets (for each experiment) by plotting the 64 scaling factors obtained from Eq. (7). We verified that, in each case, the computed a_α produce satisfactory overlaps of PM and MM histograms as in Fig. 2 (in a few cases the estimate of I_0 was adjusted in order to produce better overlapping PM/MM histogram in the low intensity region). In Fig. 3 the scaling factors a_α are organized in four different subgroups according to the mismatch, i.e. CC, TU, AA and GG. This allows an easier comparison of the results obtained by different arrays. Note that, as expected, the scaling factor depends on the triplet identity. Note also that purine–purine mismatches (AA and GG) have higher scaling factors compared to pyrimidyne–pyrimidyne mismatches (CC and TU) as expected from the geometry of the bases: purines have double rings and hence are expected to cause a stronger steric hindrance when brought close to each other in the double helix (Naef and Magnasco, 2003). In general the scaling factors for CC and TU mismatches are always found to be lower than those for AA and GG mismatches. Globally we find the following relation:

$$a_{CC} \approx a_{TU} < a_{AA} < a_{GG} \tag{8}$$

with a_{CC} … being the scaling factor average over all CC … mismatches.

Another global feature is that the scaling factors for AA and TU mismatches are weakly dependent on the nature of the two flanking nucleotides. On the contrary the scaling factors for CC and GG mismatches strongly depend on the identity of the two flaking nucleotides (see Fig. 3). Our findings about the spreading in the scaling factors can be summarized in the following relation:

$$\sigma_{a_{TU}} \approx \sigma_{a_{AA}} < \sigma_{a_{GG}} \approx \sigma_{a_{CC}} \tag{9}$$

valid in most of the cases studied (here σ is the standard deviation).

3.3. Comparison with free energies in solution

We can now compare these findings with experimental free energies for mismatch hybridization in solution. Our expectation is that $a_\alpha = \exp(\Delta G_0/RT)$, with ΔG_0 the difference between of hybridization free energies from a PM and a MM probe. As already mentioned, unfortunately only 11 of the triplets have a mismatch free energy that we know from experimental measures of hybridization in solution (Sugimoto et al., 2000). Fig. 4 shows a plot of ΔG_0 for the 11 triplets,

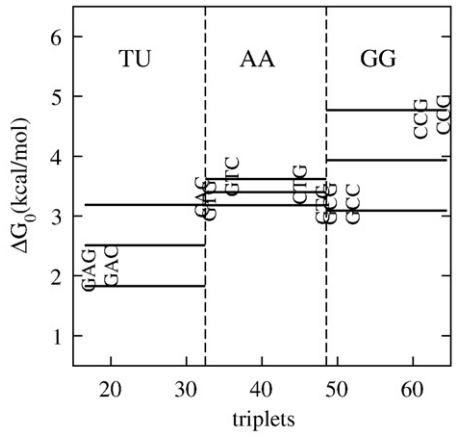

Fig. 4. Plot of free energy differences between MM and PM for the 11 triplets which have been experimentally studied at the experimental temperature T=318 K=45 °C. No data for CC mismatches are available. The other mismatches follow qualitatively the behavior of scaling factor determined from Eq. (7) and shown for two examples in Fig. 3.

where the horizontal labeling is as in Fig. 3. For the experimental hybridization free energies in solution we find:

$$\Delta G_0|_{TU} < \Delta G_0|_{AA} < \Delta G_0|_{GG} \tag{10}$$

in agreement with Eq. (8). No CC mismatch free energies have been determined experimentally so far. Note that the spreading of the GG free energies in Fig. 4 is high compared to the spreading in AA, in agreement with what found in the histogram analysis (Eq. (9)). The spreading in TU is however somewhat higher than that found in experiment for which $\sigma_{a_{TU}} \approx \sigma_{a_{AA}}$, but note that we calculate the spreading only from 3 points.

Early test experiments (Fotin et al., 1998) on some oligonucleotide microarrays suggested the following correlation between hybridization free energies in solution and in microarrays:

$$\Delta G(\text{chip}) = \gamma \Delta G(\text{solution}) + \delta \tag{11}$$

with $\gamma \approx 1$ and $\delta < 0$, implying smaller affinity for microarray hybridization compared to hybridization in solution. The $\gamma \approx 1$ has been confirmed in more recent experiments on another class of spotted microarrays (Weckx et al., 2007). No direct experimental measure of ΔG(chip) has been performed on Affymetrix arrays. Assuming that Eq. (11) with $\gamma \approx 1$ holds also on Affymetrix arrays and recalling that $\Delta G_0 = \Delta G_{PM} - \Delta G_{MM}$, one expects that ΔG_0 (chip) $\approx \Delta G_0$ (solution). We recall also that ΔG_0 depends on the nucleotides of the central triplet which is separated from the surface by a stretch of 12 nucleotides: it is plausible to expect that there is a weak surface effect and that ΔG_0 (chip) can be approximated using free energies in solution. Typical values for the latter are $2 \lesssim \Delta G_0(\text{solution}) \lesssim 4$ kcal/mol (see Fig. 4). Using the relation $a = \exp(\Delta G_0/RT)$ where we take the experimental temperature $T = 318$K and $R = 2$ cal mol/K we find $23 \lesssim a \lesssim 540$, values more than one order of magnitude higher than those found in experiments (see y-axis of Fig. 3). Another parameter that is to be taken into account when calculating the free energy is the salt concentration of the solution in Genechips experiments. Unfortunately we could not find the value of this concentration in Affymetrix experiments (the composition of the hybridization cocktail is kept secret, presumably there are some other molecules that may affect the stability of the double helix, rather than just salt). However according to the nearest-neighbor model analysis by SantaLucia and coworkers the salt concentration affects all ΔGs equally both for mismatches and perfect matches with a term proportional to the log [Na] (see for instance Eq. (6) in Peyret et al., 1999). If that is true also for RNA/DNA hybrids, then we expect that our PM/MM hybridization free energy difference ΔG_0 would not depend on salt concentration, or at least that the dependence would be very weak.

3.4. Comparison of scaling factors of different experiments

Another puzzling aspect of the analysis shows up when comparing scaling factors obtained in different experiments. Table 1 shows a summary of some experimental data analyzed on several organisms. The experiments reported have been downloaded from the Gene Express Omnibus (GEO) server at http://www.ncbi.nlm.nih.gov/geo/. The Table reports the background intensity (I_0), the average scaling factor $\langle a \rangle$ for all probes and for each of the four mismatch types. As the scaling factor measures the difference in intensities between PM and MM probes, for experiments in some given standard conditions (temperature etc …) one expects that they should be constant. We estimated the error bar on the scaling factors to be roughly of 10%. Even with these errors taken into account, the data in Table 1 still shows some large degree of variation from experiment to experiment. Note that, as mentioned before, although varying strongly the data still satisfy the inequality (Eq. (8)): the experiment to experiment variation affects all triplets in a way to conserve Eq. (8).

Table 1
Summary table of the analysis on several Affymetrix Genechips

Organism	Experiment	I_0	$\langle a \rangle$	a_{CC}	a_{TU}	a_{AA}	a_{GG}
A. thaliana	GSM134519	201	6.3	6.8	4.2	7.7	12.1
A. thaliana	GSM134783	94	6.8	8.1	4.2	9.1	12.4
A. thaliana	GSM170940	371	3.1	1.8	1.7	6.3	6.7
C. elegans	GSM177131	144	8.7	5.0	4.6	9.0	16.0
C. elegans	GSM151748	51	7.6	6.4	5.6	10.2	15.7
C. elegans	GSM135480	318	4.5	2.7	3.9	7.5	12.6
C. elegans	GSM106351	198	5.3	3.5	3.6	7.7	13.9
C. elegans	GSM40314	326	2.6	1.2	1.7	4.5	7.2
D. rerio	GSM112806	115	4.5	3.2	3.1	6.9	7.8
D. melanogaster	GSM124533	238	4.2	3.4	3.2	6.4	8.3
G. gallus	GSM158368	308	5.1	4.8	3.5	6.7	8.3
H. sapiens (HG-U95a)	2353p99av08	135	1.4	0.7	0.8	2.2	2.6
H. sapiens (HG-U133)	GSM132760	157	3.3	2.3	2.3	4.9	5.9
H. vulgaris	GSM71159	162	5.4	4.6	3.4	8.9	9.1
M. musculus	GSM115696	155	3.7	2.4	2.9	5.6	7.3
O. sativa	GSM154960	270	4.9	4.5	3.1	7.0	7.6
R. norvegicus	GSM130651	161	2.6	1.4	1.9	4.3	4.5
S. lycopersicum	GSM155103	170	5.8	7.6	3.4	6.7	9.5
X. laevis	GSM99991	100	5.8	5.1	4.1	7.9	11.1
X. laevis	GSM99990	89	4.5	3.8	3.0	6.1	8.1
X. laevis	GSM99988	88	2.1	0.8	1.8	2.6	6.2

I_0 is the estimated global background intensity, $\langle a \rangle$ is the average scaling factor over all triplets.

In total about 50 chips were analyzed and thus roughly twice as many as those reported in Table 1. We note also that some organism chips have "low" scaling factors in all experiments analyzed. This is the case for instance of the human chip HGU95 for which $a < 2$ in all cases studied (only the experiment 2353p99av08 is shown in Table 1, but several more HGU95 chips have been analyzed). The more recently designed human expression chip, the HGU133, has typically higher scaling factors $\langle a \rangle \approx 3$, which is higher than the HGU95 but still well below the highest values shown in Table 1. High scaling factors can be understood as a consequence of the improvement of the probes design on the chip as the PM/MM intensity ratio gets closer to the "ideal" solution value $a = \exp(\Delta G_0/RT)$. A non-optimal design can lead, for instance, to a high level of cross-hybridization which in turn may cause strong deviations from the behavior predicted by Eq. (6). According to our opinion high scaling factors and good overlaps between PM and rescaled MM intensity histograms, as those shown in Fig. 2 are to be taken as indications for a good quality for an experiment. Using these criteria, the experiments showing the best quality, among those shown in Table 1, are the *X. laevis* GSM99991 and GSM99990, as well as other *A. thaliana* and *C. elegans* chips.

In Table 1 we evidenced the data for which the scaling factors are smaller than 1. This happens in 2 of the 21 chips analyzed, and only for mismatches of type CC and TU. Note that for these cases also the average scaling factors for AA and GG are quite low compared to other cases. A value for $a < 1$ implies, on average, that the MM probes are brighter than the PM ones, which is clearly at odds with thermodynamic expectations. The problem of bright mismatches, i.e. of probes for which $I_{MM} > I_{PM}$ has been discussed in the literature (Naef and Magnasco, 2003). In the experiments shown in Table 1 there are on average 20% of bright MMs, which drop to 8% if one excludes from the analysis the bottom 30% probes with the lowest PM intensity. Hence it seems that this problem is likely to be related to background level fluctuations for probes with lowest intensity.

4. Discussion

The two main questions arising from the histogram analysis are 1) Why are the scaling factors only in qualitative agreement with the data in solution (Figs. 3 and 4), while they differ quantitatively of more than one order of magnitude? 2) Why is there so much experiment-to-experiment variation? Here we discuss some simple phenomena

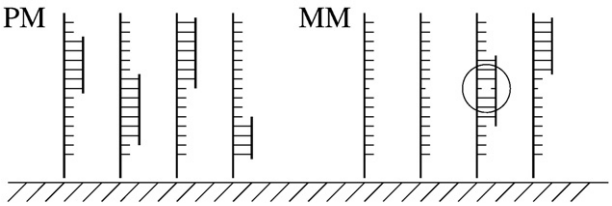

Fig. 5. Sufficiently short fragments can hybridize at the MM probe without binding to the central nucleotide. This yields a scaling factor $a<e\Delta^{G_0/RT}$. The fact that the distribution of fragment lengths may vary from experiment to experiment can explain the variability of the observed scaling factors. A higher degree of fragmentation would lead to an overall decrease of scaling factors for all the triplets.

that may explain the observed experimental data. In particular we focus on the effects of target fragmentation on the value of the scaling factor (see Fig. 5).

4.1. Fragmentation of RNA targets

Many biochemical steps are necessary to produce RNA for a hybridization experiment going from extraction from cells, to amplification and labeling. At the end of these processes the target molecules are typically several hundred base pair long. Such long molecules are impractical for experiments. In fact long targets have strong steric hindrance with the surface, and they tend to fold into stable secondary structures (this type of "self hybridization" is common especially for RNA molecules). To avoid these effects the target is treated in various ways so to produce shorter molecules. This process produces targets with a broad distribution of lengths.

A physical model of hybridization with fragments of variable lengths has been developed in Ferrantini (2007). Here we briefly outline the main ideas of that approach. We restrict ourselves to the simpler case of homogeneous system, neglecting the differences between AT and CG binding. A perfect matching fragment of length 1 will contribute to an average hybridization free energy $(l-1)\,\overline{\Delta G}$, where $\overline{\Delta G}$ is the stacking term averaged over all nucleotide pairs. The Langmuir model for PM hybridization is then modified as follows

$$I_{\mathrm{PM}} = I_0 + A\sum_{l\le 25} l c_l \exp\left[(l-1)\,\overline{\Delta G}/RT\right] \tag{12}$$

where c_l is the concentration in solution of a fragment of length l. The factor l inside the sum accounts for the number of fluorescent labels which is proportional to the length of the fragment. For the sake of simplicity we have restricted the sum in Eq. (12) to 25 nucleotides; in reality there could be dangling ends, but for the qualitative argument developed here this simplification is sufficient. A similar, slightly more complicated, expression can be written for the MM intensity. In that case one should distinguish between fragments covering the 13th nucleotide, and those not binding there. In the former case the Boltzmann factor contains an extra term $\exp(-\Delta G_0/RT)$.

A calculation of the scaling factor a_α (Eq. (7)) within this model shows that 1) $a \le \exp(\Delta G_0/RT)$ and 2) the results depend on c_l. A stronger degree of fragmentation leads to c_l "dominated" by fragments of shorter and shorter lengths, which yields to a lowering of the scaling factors. As fragmentation involves all target molecules in solution this effect is global, affecting all the scaling factors of different triplets at once. This is somewhat in agreement with what seen in experiments (see Table 1). Note that in two different experiments, in which the target samples are prepared independently, there can be different degree of fragmentation. Unfortunately there are not yet systematic measurements of fragment length distribution in Affymetrix experiments. Such measurements would indeed give a better insight on the effect of fragmentation on the scaling factors.

Summarizing, in this paper we derived a relationship between PM and MM histograms in Affymetrix arrays using as inputs the Langmuir and Nearest Neighbor models. This relationship has been tested on experimental data on chips of different organisms and in most of the cases is verified. The analysis is done separately for each central triplet probe sequence as suggested by the Nearest Neighbor model applied to RNA/DNA duplexes (Sugimoto et al., 2000). We verified that the scaling parameters extracted from our analysis are qualitatively consistent with the experimental stacking free energies for hybridization in solution (Sugimoto et al., 1995, 2000). As an explanation of the experiment to experiment variation we suggest that it may be due to a different degree of fragmentation of RNA molecules in solution. The analysis of histograms presented here can be used as a first test of the quality of the experimental data. It shows the importance of physical-chemistry in microarray data analysis.

References

Binder, H., Preibisch, S., 2005. Specific and nonspecific hybridization of oligonucleotide probes on microarrays. Biophys. J. 89 (1), 337 Jul.
Bloomfield, V.A., Crothers, D.M., Tinoco Jr., I., 2000. Nucleic Acids Structures, Properties and Functions. University Science Books, Mill Valley.
Carlon, E., Heim, T., 2006. Thermodynamics of RNA/DNA hybridization in high-density oligonucleotide microarrays. Physica A 362, 433.
Carlon, E., Heim, T., Klein Wolterink, J., Barkema, G.T., 2006. 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 73, 063901.
Ferrantini, A., 2007. Analysis of nucleic acid interaction on DNA microarrays. Master's thesis, University of Padua.
Fotin, A.V., Drobyshev, A.L., Proudnikov, D.Y., Perov, A.N., Mirzabekov, A.D., 1998. Parallel thermodynamic analysis of duplexes on oligodeoxyribonucleotide microchips. Nucleic Acids Res. 26, 1515.
Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, S., 2003. Bioinformatics and computational biology solutions using R and bioconductor. Statistics for Biology and Health. Springer.
Halperin, A., Buhot, A., Zhulina, E.B., 2006. On the hybridization isotherms of DNA microarrays: the Langmuir model and its extensions. J. Phys., Condens. Matter 18, S463.
Held, G.A., Grinstein, G., Tu, Y., 2003. Modeling of DNA microarray data by using physical properties of hybridization. Proc. Natl. Acad. Sci. 100, 7575.
Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., Lockhart, D.J., 1999. High density synthetic oligonucleotide arrays. Nat. Genet. 21, 20.
Naef, F., Magnasco, M.O., 2003. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. Phys. Rev. E 68, 011906.
Peyret, N., Seneviratne, P.A., Allawi, H.T., SantaLucia Jr., J., 1999. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal AA, CC, GG and TT mismatches. Biochemistry 38, 3468.
Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467.
Stoughton, R., 2005. Applications of DNA microarrays in biology. Annu. Rev. Biochem. 74, 53.
Sugimoto, N., et al., 1995. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry 34, 11211.
Sugimoto, N., Nakano, M., Nakano, S., 2000. Thermodynamics–structure relationship of single mismatches in RNA/DNA duplexes. Biochemistry 39, 11270.
Weckx, S., Carlon, E., De Vuyst, L., Van Hummelen, P., 2007. Thermodynamic behavior of short oligonucleotides in microarray hybridizations can be described using Gibbs free energy in a nearest-neighbor model. J. Phys. Chem., B 111, 13583.