Published online 31 May 2012

Nucleic Acids Research, 2012, Vol. 40, No. 18 ***e138***
doi:10.1093/nar/gks475

Probing hybridization parameters from microarray experiments: nearest-neighbor model and beyond

W. W. Hadiwikarta[1,2], J.-C. Walter[2], J. Hooyberghs[1,3] and E. Carlon[2,*]

[1]Flemish Institute for Technological Research, VITO, Boeretang 200, B-2400 Mol, [2]Institute for Theoretical Physics, KULeuven, Celestijnenlaan 200D, B-3001 Leuven and [3]Department WNI, Hasselt University, Campus Diepenbeek, Agoralaan - Building D, B-3590, Diepenbeek, Belgium

Received March 15, 2012; Revised May 3, 2012; Accepted May 7, 2012

ABSTRACT

In this article, it is shown how optimized and dedicated microarray experiments can be used to study the thermodynamics of DNA hybridization for a large number of different conformations in a highly parallel fashion. In particular, free energy penalties for mismatches are obtained in two independent ways and are shown to be correlated with values from melting experiments in solution reported in the literature. The additivity principle, which is at the basis of the nearest-neighbor model, and according to which the penalty for two isolated mismatches is equal to the sum of the independent penalties, is thoroughly tested. Additivity is shown to break down for a mismatch distance below 5 nt. The behavior of mismatches in the vicinity of the helix edges, and the behavior of tandem mismatches are also investigated. Finally, some thermodynamic outlying sequences are observed and highlighted. These sequences contain combinations of GA mismatches. The analysis of the microarray data reported in this article provides new insights on the DNA hybridization parameters and can help to increase the accuracy of hybridization-based technologies.

INTRODUCTION

Hybridization of single-stranded nucleic acids to form a duplex is a reversible chemical reaction, which is at the basis of many processess and techniques currently used in biotechnology, as for instance PCR (1). Due to its central importance, hybridization has been intensively studied in experiments (focusing on the thermodynamics (2,3) or kinetics of the process) and also in computer simulations (4).

The thermodynamics of DNA hybridization is usually described by the nearest-neighbor (NN) model (5). This model assumes that the free energy of a duplex can be expressed as a sum of dinucleotide stability parameters; it is therefore based on the principle of additivity. From the NN parameters one can, for instance, estimate melting temperatures, compute melting curves and predict secondary structures in which RNA molecules fold (6,7). In the folding problem, many different local conformations arise as single nucleotide mismatches, bulges, stem–loop structures, etc. Describing these conformations in the framework of the NN model is very challenging and requires a large number of parameters (6). However, only a limited number of them have been measured directly in experiments (8). In addition, one may also wonder whether additivity holds in such cases. To investigate a large number of different conformations, it would be very advantageous to have access to high-throughput measurements, provided that they are sufficiently accurate.

In this article, we quantitatively determine free energy penalties for mismatches using microarray data obtained from a set of optimized and dedicated experiments. In DNA microarrays, several thousand of different sequences can be spotted at a surface, hence a large number of hybridization reactions takes place simultaneously. We use two different approaches: the first one is based on a linear regression of a large set of experimental data points (≈ 1000) to fit 58 NN dinucleotide parameters. The second method relies on the computation of the logarithm of the ratios of fluorescent intensities measured from different spots of the arrays. We show that both methods provide highly correlated set of NN parameters. In addition, the second approach allows to probe the limitations of the NN model. It is found that when two mismatches are closer than 5 nt, additivity breaks down and the free energy of the duplex is not equal to the sum of the two separate contributions of isolated mismatches. We also quantify the influence of mismatches close to the edge of the double helix and show that the free energy penalty is

*To whom correspondence should be addressed. Tel: +32 16 32 72 39; Fax: +32 16 32 79 86; Email: enrico.carlon@fys.kuleuven.be

much weaker in those cases. Overall, this work provides new insights on DNA hybridization thermodynamics and can help to increase the accuracy of hybridization-based technologies.

MATERIALS AND METHODS

The experiments were performed on custom Agilent arrays, following a standard protocol, which is discussed in (9). In each experiment, a single target sequence in solution was hybridized at concentrations ranging typically from ~10 picoM to ~2 nanoM. In total, three different sets of experiments were performed using the target sequences shown in Table 1. These sequences were selected from 25-mers human DNAs using Optimal Design methods (10). The theory of Optimal Design provides some criteria of selecting an optimal set of measurements, which minimize the uncertainties in the parameters of a statistical model (see Supplementary Data).

From the targets of Table 1, three different microarrays were designed and used for hybridization to either t_1, t_2 or t_3. Each microarray contains probes with either zero, one or two mismatches with respect to the given target, covering all possible mismatch combinations. In a stretch of N nucleotides, there can be $3N$ single mismatch probes and $9N(N-1)/2$ double mismatch probes. For $N = 25$, this gives in total 2776 different sequences, which were spotted in the microarray. The sequences were replicated 15 times to fill up completely a 44K custom Agilent array. Another design was also used in which mismatches have a minimal distance of 4 nt from the border and a minimal relative distance of 5 nt. In this case, the total number of sequences is 646. These sequences were replicated 23 times to fill a 15K custom arrays. We considered hybridizing sequences of 25 nt. This is because in the previous studies (11), these sequences were found to attain thermodynamic equilibrium after ~3 h of hybridization (in the experiments, the hybridization time is of 17 h, hence thermodynamic equilibrium is guaranteed). A hybridization experiment provides a large number of fluorescence intensities: the highest intensity is from spots containing perfect match sequence, whereas the intensity decreases with the number and type of mismatches. The reduction of the intensity provides an estimate of the hybridization free energy. We use two different methods to obtain the NN parameters, as discussed in the next sections.

Table 1. Target sequences used in the experiments

t_1:	5′-CTGGTCTTAGATGC**AGC**GACTGTTT-poly(A)-3′-Cy3
t_2:	5′-CTGCACAATTCCGG**AGC**TATGAATT-poly(A)-3′-Cy3
t_3:	5′-AATAATGCTCATTAGGCACCGGGAA-poly(A)-3′-Cy3

At the 3′ side of each sequence a 20-mer poly(A) is attached, terminating with a Cy3 fluorophore. The targets were selected from Optimal Design criteria (10) (Supplementary Data). Each target is hybridized separately on specific microarrays containing mismatched probes with up to two mismatches with respect to the target. Note that t_1 and t_2 share a common triplet of nucleotides AGC at the same sequence position (in bold characters). The mismatches centered around this triplet will be discussed in some details in the 'Results' Section.

RESULTS

Nearest-neighbor parameters from linear regression

Equilibrium thermodynamics predicts that the measured fluorescence intensity from a spot i equals to:

$$I_i = I_0 + \mathrm{Ace}^{-\Delta G_i/\mathrm{RT}} \quad (1)$$

where ΔG_i is the hybridization free energy between the target sequence and a probe sequence in i, A is a parameter, which sets the intensity scale, c the target concentration, R the gas constant and T the temperature (experiments are performed at $T = 65°\mathrm{C} = 338\mathrm{K}$, which is the value of the temperature used in the rest of the analysis). Although the data analyzed are background-subtracted from the Agilent scanner, there remains always some small aspecific signal, which we denote by I_0 in Equation (1). In the experiments, I_i is obtained from the average over typically approximately 15 replicated spots. One should note that Equation (1) is valid at sufficiently low target concentrations, i.e. when only a limited fraction of probes is hybridized in a spot, hence far from chemical saturation. On the other hand, at very low concentrations, the specific signal, i.e. the second term in Equation (1), can become comparable to I_0. Therefore, for the analysis of the data we restricted ourselves to intermediate concentrations and intensities for which we explicitly verified that the intensities scale linearly with concentrations, as predicted by Equation (1) (more details can be found in the Supplementary Data). In the intensity scale of the experiments $I_0 \approx 1$, whereas the values used in the analysis are $I_i \gtrsim 20$. In practice, the large majority of the intensities in experiments with target concentration $c = 100$ pM or higher are above this threshold value.

In the following, we will consider the logarithm of the intensities measured with respect to the perfect match (PM) intensity. Using Equation (1), for $I_i \gg I_0$ we get:

$$y_i \equiv \ln I_i - \ln I_{PM} = -\frac{\Delta G_i - \Delta G_{PM}}{RT} \quad (2)$$

which defines the free energy penalty of probe i with respect to the perfectly matching probe. This penalty can be expressed as a sum of NN dinucleotide parameters. Consider, for instance, the example of a probe i with a single mismatch of type A with respect to the target nucleotide G and with neighboring nucleotides G and T. We have:

$$\begin{aligned}\Delta G_i\begin{pmatrix}\ldots G\underline{A}T\ldots\\ \ldots C\underline{G}A\ldots\end{pmatrix} - \Delta G_{PM}\begin{pmatrix}\ldots GCT\ldots\\ \ldots CGA\ldots\end{pmatrix}\\ = \Delta G\begin{pmatrix}G\underline{A}\\ C\underline{G}\end{pmatrix} + \Delta G\begin{pmatrix}\underline{A}T\\ \underline{G}A\end{pmatrix} - \Delta G\begin{pmatrix}GC\\ CG\end{pmatrix} - \Delta G\begin{pmatrix}CT\\ GA\end{pmatrix}\\ \equiv \Delta\Delta G\begin{pmatrix}G\underline{A}T\\ C\underline{G}A\end{pmatrix}\end{aligned} \quad (3)$$

We use the following notation: the target sequence is the bottom strand and the probe sequence, which is oriented from 5′ – 3′, is the top strand. This example corresponds to

target t_1 or t_2 at position 10, counting from 3′ end (the triplet of nucleotides are indicated in bold in Table 1). In Equation (3) $\Delta\Delta G$ is defined as the free energy penalty of an isolated mismatch in a DNA duplex. This penalty is expected to be a local effect. In the NN model, this locality is inherent: the dots in Equation (3) indicate identical nucleotides in the two sequences, their contribution cancels out and leaves per isolated mismatch only four dinucleotide parameters around the mismatch position. There are in total only 58 such dinucleotide parameters: 10 perfect match parameters and 48 single mismatch parameters (taking into account symmetries). The dinucleotide parameters are not directly experimentally accessible and are not unique (12), e.g. they can be shifted by some constant value such that the physically accessible $\Delta\Delta G$ remains unchanged (see Supplementary Data).

Equations (2) and (3) define a linear problem: each measured y_i can be expressed by a linear combination of dinucleotide parameters. In order to extract the parameters from the data, we combined the results of the three experiments and performed a least square minimization of Equation (2). Mismatches closer than five sites from the helix edges were excluded from the analysis, as well as pairs of mismatches with a distance smaller than 5 nt.

The 58 adjustable parameters were fitted on a set of about a thousand of experimental data points above the intensity threshold. The fitted parameters then applied to produce the plot as shown in Figure 1 for all available intensities of the experiments in which either sequence t_1, t_2 or t_3 was hybridized on its corresponding microarray at a concentration of $c = 100$ pM. The data are plotted as a function of the unique $\Delta\Delta G$ for triplets defined as in Equation (3). We note that there is very good agreement between the data and the thermodynamic model of Equation (1). The experiments follow the equilibrium isotherm (a straight line with a slope equal to 1/RT) for a range of intensities of more than four orders of magnitude. A previous study (9) in which hybridizing strands were 30-mers did not provide a single straight line in a ln I versus $\Delta\Delta G$ plot. Deviations due to lack of thermodynamic equilibrium were observed in the high-intensity range, as discussed in (11,13).

Further, it is important to note that we do not only find internally consistent results, but that our microarray-derived free energy parameters also correlate to a fair degree with those reported in literature for hybridization in solution (8). Figure 2 shows a correlation plot of the free energy penalties [i.e. the $\Delta\Delta G$ defined as in the example of Equation (3)] obtained from the microarray data analysis and those from SantaLucia *et al.* from (8). The Spearman correlation coefficient is equal to 0.855. This clearly shows that free energy parameters for DNA features measured by the presented microarray approach also apply for thermodynamic properties in solution. This opens the highly parallelled microarray toolbox for the study of thermodynamics of DNA structures. An example is discussed in the next section.

Nearest-neighbor parameters from ratios of intensities: probing additivity

The crucial assumption of the NN model is additivity of local free energy contributions. We probe here the limits of additivity of free energy penalties as a function of the distance between two mismatches. We will access the free

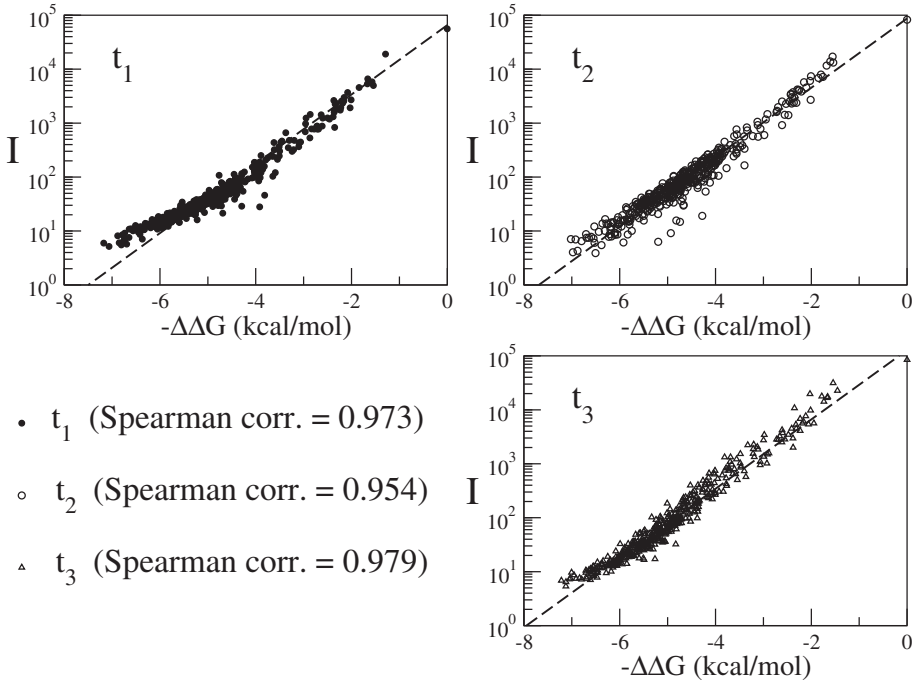

Figure 1. Plot of the Intensities for concentrations $c = 100$ pM from the experiments using hybridization of targets t_1, t_2 or t_3, as a function of the $\Delta\Delta G$ parameters obtained from least-squared minimization. The data agree well with hybridization isotherm given in Equation (2), shown as a straight line in the linear-log scale.

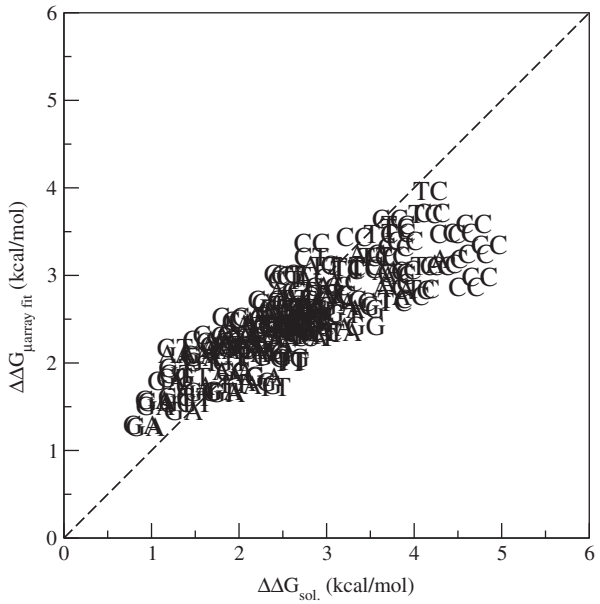

Figure 2. Plot of free energy penalties $\Delta\Delta G$ for triplets obtained from the microarray fit versus those from hybridization in solution (8). The central mismatching nucleotides of the triplet [underlined in Equation (3)] are indicated in the plot.

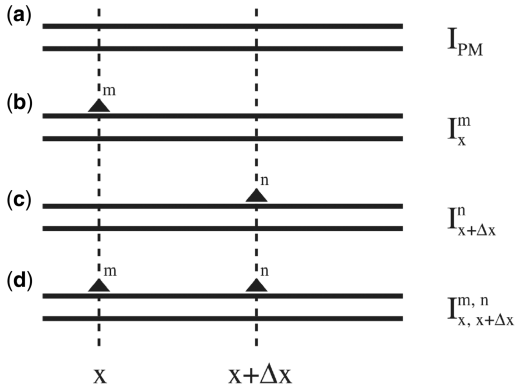

Figure 3. Schematic representation of hybridizing strands in the microarray experiment. From the appropriate ratios of intensities measured from these spots, the free energy parameters can be determined and the additivity principle can be tested. As in the rest of the article, the lower strand is the fixed target sequence. The upper strand is the probe sequence. The filled triangles denote mismatching nucleotides. In the four examples from the top we show: (**a**) hybridization with a PM probe, (**b** and **c**) hybridization with a single mismatch probe where the mismatching nucleotides are m and n at positions x and $x+\Delta x$, respectively, (**d**) hybridization with a probe carrying two mismatches. We use the notations I^m_x, $I^n_{x+\Delta x}$ and $I^{m,n}_{x,x+\Delta x}$ to denote the corresponding intensities measured in the experiment.

energy parameters by comparing ratios of intensities measured from different spots in the microarray.

Hereto, we combine microarray spots that contain probes with zero, one or two mismatches with respect to the target and we denote the location of the mismatch by x or $x+\Delta x$, as illustrated in Figure 3. The associated free energy penalties can then be derived from the intensity measurements as follows

$$\Delta\Delta G^m_x = -\mathrm{RT}\ln\left(\frac{I^m_x}{I_{\mathrm{PM}}}\right) \quad (4)$$

$$\Delta\Delta G^n_{x+\Delta x} = -\mathrm{RT}\ln\left(\frac{I^n_{x+\Delta x}}{I_{\mathrm{PM}}}\right) \quad (5)$$

$$\Delta\Delta G^{m,n}_{x,x+\Delta x} = -\mathrm{RT}\ln\left(\frac{I^{m,n}_{x,x+\Delta x}}{I_{\mathrm{PM}}}\right) \quad (6)$$

in which the superscript m and n represent the three possible mismatching nucleotides at location x and $x+\Delta x$, respectively. If the additivity of the NN model holds, the free energy penalty of Equation (6) should equal the sum of the individual penalties of Equations (4) and (5). To test this, we introduce

$$\alpha = \frac{\Delta\Delta G^m_x + \Delta\Delta G^n_{x+\Delta x} - \Delta\Delta G^{m,n}_{x,x+\Delta x}}{\Delta\Delta G^m_x + \Delta\Delta G^n_{x+\Delta x}} \quad (7)$$

which measures the relative deviation from additivity. Figure 4 shows the experimental results for α in which we averaged over x, m and n, leaving α as a function of the distance $|\Delta x|$ between two mismatches. From this data, we notice that α has a value of about zero when the mismatches are separated by $\geq$5 nt, but a clear positive value for smaller Δx. Apparently, the free

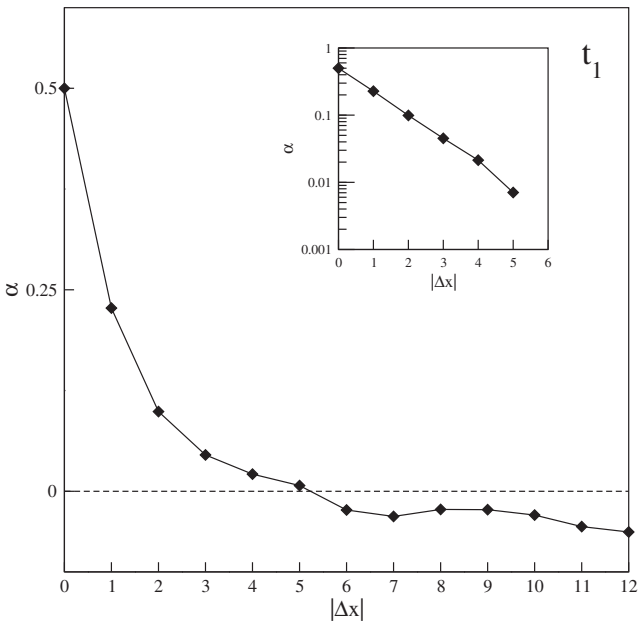

Figure 4. Parameter α, the relative deviation from additivity, from the experiment of target t_1, averaged over x, m and n as a function of the distance $|\Delta x|$ between two mismatches. The inset shows the plot with α in log scale.

energy penalty of two nearby mismatches is smaller than the sum of the two individual contributions, resulting in a positive α. Furthermore, the inset from Figure 4 shows that the relationship is linear in a semi logarithmic plot, hence α decays exponentially with $|\Delta x|$. Note that at $\Delta x = 0$ only one mismatch is present, hence $m = n$ and α will be identical to 1/2 according to Equation (7). All these observations result from direct measured values,

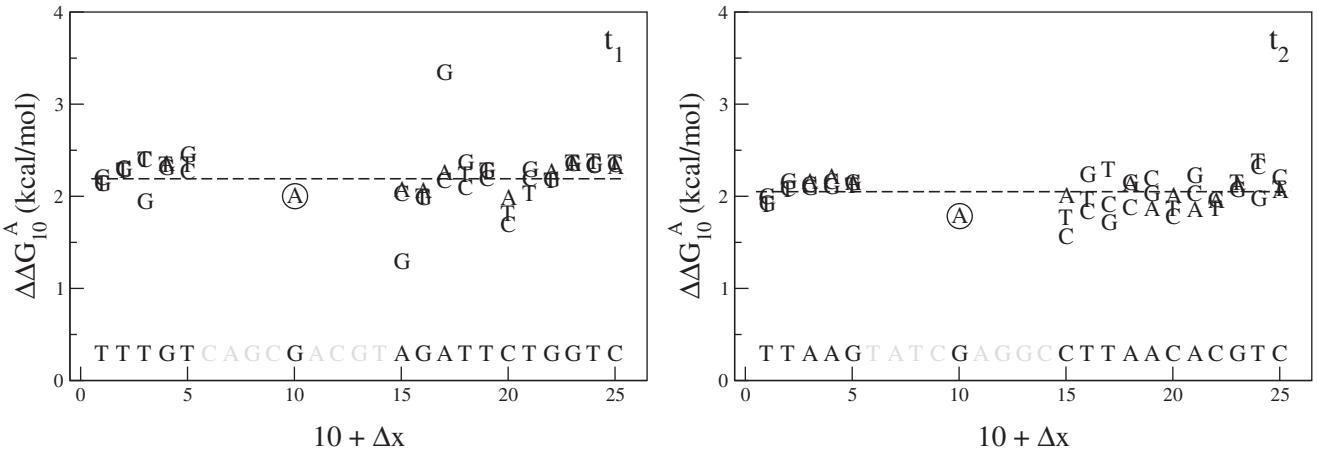

Figure 5. Free energy penalty $\Delta\Delta G^{A}_{10}$ for focus mismatch ($m = A$, $x = 10$) derived from experimental intensities according to Equation (8) as a function of the location $x + \Delta x$ of the second mismatch (n, $x + \Delta x$). For each $|\Delta x > 4|$ the three values, one per possible mismatch, are indicated by the letter representing the mismatching nucleotide n of the probe. The target sequence is written in top of the x-axis in $3' - 5'$ notation, t_1 in left pane, t_2 in right pane. The dotted line corresponds to the median value of the 48 estimates. The circled point is the estimate without second mismatch coming from Equation (4). For this particular mismatch, the free energy penalty for both t_1 and t_2 is identical and corresponds to $\Delta\Delta G\left(\begin{smallmatrix} GAT \\ C\underline{G}A \end{smallmatrix}\right)$.

containing no fitting parameters and strongly suggest that in double-stranded DNA mismatches have a physical interaction with each other, which decays exponentially to zero over a distance of about five nucleotides.

These results are setting some limitations on the additivity of the NN model. However, outside this interaction region of 4 nt, we expect the NN model to hold i.e. α should be zero and mismatches can be considered as isolated. This can be explicitly checked in a very direct way. When $\alpha = 0$, we get from Equation (7)

$$\Delta\Delta G^{m}_{x} = \Delta\Delta G^{m,n}_{x,x+\Delta x} - \Delta\Delta G^{n}_{x+\Delta x}. \quad (8)$$

The free energy penalty $\Delta\Delta G^{m}_{x}$ of a mismatch m at location x, which we will call the focus mismatch (m, x), can be estimated either directly using Equation (4) or via a second mismatch (n, $x + \Delta x$) using Equations (5) and (6) for any choice of n and $\Delta x > 4$. Hence, the free energy penalty of the focus mismatch can be estimated from measurements in many independent ways and they should provide the same answer if additivity holds. Note that, using Equations (5) and (6) I_{PM} drops out in the right hand side of Equation (8).

Figure 5 illustrates how Equation (8) can be used to estimate the $\Delta\Delta G$ using different combinations of n and Δx. In this specific example, we consider $\Delta\Delta G^{A}_{10}$, which corresponds both for target t_1 and t_2 to $\Delta\Delta G\left(\begin{smallmatrix} GAT \\ C\underline{G}A \end{smallmatrix}\right)$ (in the Supplementary Data, we show other examples featuring additivity for different focus mismatches).

In the pane for target t_2, all the estimates of the free energy penalty are close the each other, the $48 + 1$ estimates tightly lie around a median value, in this case, ~2.1 kcal/mol, is indicated by the dotted line. The picture in the right pane is a typical one, which we observe for any focus mismatch (m, x). This confirms that additivity holds in the regime $\Delta x > 4$, i.e. when mismatches are separated by >4 nt. Moreover, it shows that the microarray measurement is internally consistent. Second, the left pane, i.e. experiment t_1, provides the same median value for the free energy penalty, showing also the robustness of the microarray approach to estimate free energies of DNA structures. However, this figure was chosen because it is atypical in the sense that one notices two pronounced outlying values. They correspond to a sequence where both the focus mismatch and the second mismatch are of type AG. Since, they clearly deviate from an otherwise nicely consistent picture, we believe there must a physically underlying reason for it. We will come back to this point in the section where we discuss thermodynamic outliers.

Note that with this second method, we accessed values for the free energy penalties of isolated mismatches without using any multiple regression or fitting procedure, but we simply compared the ratios of intensities, Equations (4–6), to get a consistent set of independent estimates. The free energy penalties are then obtained from the median over all data points. We compared the free energy penalties obtained from this method (median) with those obtained from linear regression as discussed in the previous section. The two sets of data are well correlated with a Pearson correlation coefficient equal to 0.966 (see Supplementary Data). This correlation shows the equivalence of the two approaches. In this analysis, we restricted ourselves to mismatches in the bulk of the sequence, i.e. x is >5 nt from the border. Closer to the border, we observe boundary effects, which are covered in the next section.

Boundary effects

The previous section ended by showing the equivalence of both approaches to access free energy penalties of an isolated mismatch, provided the data are restricted to bulk mismatches. The direct median method of the previous section can also assess penalties of mismatches

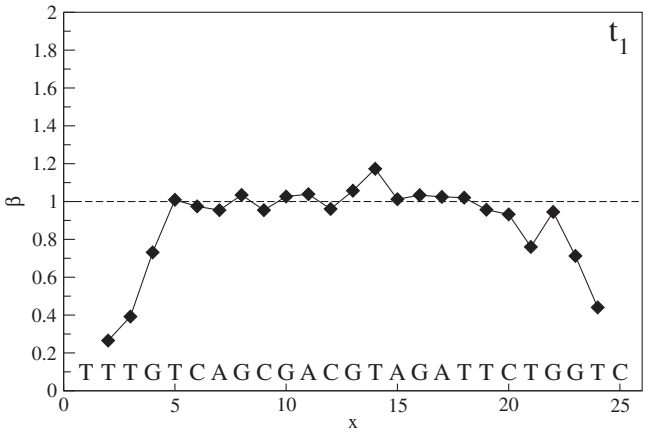

Figure 6. Boundary effect: β, the relative reduction of mismatch free energy penalty, as a function of location for experiment with target t_1. Each point is the average of three estimates, one per possible mismatch. Data are absent for the extremal locations $x = 1$ and $x = 25$, since no value can be calculated by the NN model.

close to the boundary, whereas on the contrary, the fitting method cannot by construction. The latter, however, has the advantage of fitting a full parameter set of the NN model and as such can easily provide bulk values for the free energy penalty of any isolated mismatch. The combination of both methods now provides an elegant way to assess the effect of boundary proximity on an isolated mismatch. Hereto, we introduce the parameter β as the relative reduction of free energy penalty of a mismatch when compared to its bulk value.

$$\beta = \frac{\Delta\Delta G^m_x}{\Delta\Delta G^m_{bulk}} \quad (9)$$

In Figure 6, the parameter β is shown as a function of x after averaging over m. It is clear that, as expected, β is approximately equal to one in the bulk, whereas when approaching the boundary, a reduction of free energy penalty occurs which reaches up to 80%. Note that for mismatches at the boundary, $x = 1$ and $x = 25$, the NN model is not applicable and no data are presented. Figure 6 shows that the range of boundary effect is about 4 nt.

Thermodynamic outliers

As a final result of this article, we come back to the two outliers observed in Figure 5(a); the same deviations are found in replicated experiments at different concentrations: therefore, they are unlikely due to experimental errors. For these two cases, we find $\Delta\Delta G^{A,G}_{10,15} - \Delta\Delta G^{G}_{15} \approx 1.2$ kcal/mol and $\Delta\Delta G^{A,G}_{10,17} - \Delta\Delta G^{G}_{17} \approx 3.1$ kcal/mol, strongly deviating from the median value (≈ 2.1 kcal/mol). The common feature of these two sequences is that they involve GA mismatches. The two set of mismatches are arranged in an antiparallel way i.e. one G and one A are on the same strand. Mismatches of GA type in DNA and RNA helices have been the subject of several studies in the past (14–21). In the RNA folding, it is known that GA pairs contribute substantially to the RNA helix stability. Their contribution is comparable to that of a canonical AT pair. As AT pairs, GA form two hydrogen bonds, but can also assume four different conformations (14). The microarray data suggests that the antiparallel combination of GA and AG pairs of mismatches have a long range interaction effect, which is probably a signature of some structural conformational change of a double helix containing these pairs. Next-nearest neighbor effects extending up to 4 nt distance for antiparallel GA mismatches have been reported in the case of RNA duplexes in (19) (longer distances were not considered that case). We investigated antiparallel GA and AG pairs of mismatches also in sequences t_2 and t_3, but found no anomalous behavior in those cases. This suggests that the nucleotide sequences between the two GA/AG pairs plays an important role in the overall stability of the duplex.

As a further proof of the outlying behavior of antiparallel GA/AG pairs, we show in Figure 7 a plot of free energy penalties for tandem mismatches (neighboring double mismatches). These are again obtained from Equation (6) for different m and n mismatching nucleotides, where in the case of tandem mismatches, Δx is equal to 1. On each location of the sequence our data set contains nine different types of tandem mismatch. A clear boundary effect is noticeable, but when looking at the bulk data points tandem mismatch of the type GA/AG are again outlying, they appear to be particularly stable with a free energy penalty ~2 kcal/mol below average.

DISCUSSION AND CONCLUSION

In this article, we have analyzed DNA hybridization reactions in microarrays and quantified free energy penalties of single and double mismatches. We have shown that the experimental data are very precise and reproducible. The microarray data follow an equilibrium isotherm over a range of four orders of magnitude in the fluorescence intensities and allow the extraction of accurate thermodynamic parameters. First, the analysis provides a database with a large number of NN parameters for isolated mismatches. These parameters correlate well with those reported in the literature from hybridization experiments in solution. Second, the experiments contain systematic measurements of hybridization with two mismatches, which allowed us to probe the validity limit of the NN approximation. We showed that when two mismatches are separated by a distance of ≥ 5 nt their effect is additive, allowing a standard approach with the NN model. However, for shorter distances, the additivity is no longer valid and we found that duplexes with neighboring mismatches are more stable than expected from additivity. This interaction was shown to decay exponentially as a function of the distance between mismatches. Further, we investigated the behavior of mismatches close to the helix edges, and showed that their free energy penalty is reduced up to 80% when compared to the bulk behavior. The boundary effect was observable up to 4 nt from the helix edge. Finally, we also found some thermodynamic outliers, sequences involving two

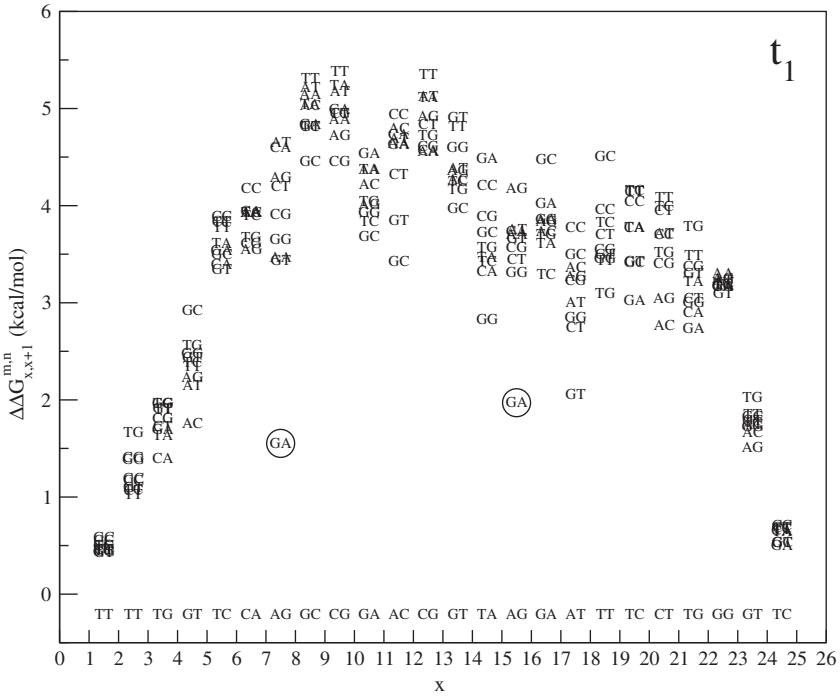

Figure 7. The free energy penalty of tandem mismatches from experiment with target t_1: $\Delta\Delta G\left(\frac{xmny}{x'aby'}\right)$, where x' and y' are complementary to x and y, respectively. *ab* denoted above the *x*-axis are the fixed nucleotides in the target. mn is a tandem mismatch in the probe and the vertical position of these letters in the plot give the associated free energy penalty. Note the low free energy penalty for $\frac{5'-\underline{GA}-3'}{3'-\underline{AG}-5'}$ mismatches (encircled).

antiparallel GA mismatches, in which the mismatch interaction appears to persist beyond 5 nt. These outliers were not related to experimental error indicating a signature of some structural conformational change of a double helix containing these mismatch pairs.

Overall, the analysis of the microarray data reported in this article provides new quantitative insights on the DNA hybridization parameters, on the NN model and its present limitations. Our study is in line with a number of recent papers, which have been dedicated to the investigations of fundamental physico-chemical properties of DNA arrays [(22–31)]. Due to the relevance of hybridization in many technologies, going from PCR (1) to recent developments in biosensors, e.g. (32), a good thermodynamic model is also important from the application point of view. A precise quantification of interaction free energies involved in the hybridization will help to increase the accuracy of microarrays and other hybridization-based technologies, so that these devise could realize their full potential, for instance, for clinical applications (33). For these applications, an increase in specificity and sensitivity is very important and can be achieved through better understanding of fundamental properties of hybridization in these devices.

There has been considerable attention in recent years (9,22,24–34) in understanding the fundamentals of hybridization in DNA microarrays and its impact in data analysis. Here, we have shown that microarrays are a reliable and high-throughput tool to gain insight on DNA hybridization thermodynamics. The same method could be used to screen other types of defects, as bulges. Indeed, it was recently used for understanding loop conformations (22).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–4.

ACKNOWLEDGEMENTS

We thank Karen Hollanders for expert technical assistance.

FUNDING

FWO Vlaanderen [G.0311.08]; KULeuven [STRT1/09/042]; VITO [ZL39010200-401]. Funding for open access charge: KULeuven [STRT1/09/042].

Conflict of interest statement. None declared.

REFERENCES

1. Alberts,B., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (2002) *Molecular Biology of the Cell*.. Garland Science, New York.
2. Breslauer,K.J., Frank,R., Blöcker,H. and Marky,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
3. SantaLucia,J. Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
4. Sambriski,E.J., Schwartz,D.C. and de Pablo,J.J. (2009) A Mesoscale Model of DNA and Its Renaturation. *Biophys. J.*, **96**, 1675–1690.
5. Bloomfield,V.A., Crothers,D.M. and Tinoco,I. Jr (2000) *Nucleic Acids Structures, Properties and Functions*.. University Science Books, Mill Valley.

6. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.
7. Andronescu,M., Condon,A., Hoos,H.H., Mathews,D.H. and Murphy,K.P. (2010) Computational approaches for RNA energy parameter estimation. *RNA*, **16**, 2304–2318.
8. SantaLucia,J. Jr and Hicks,D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 415–440.
9. Hooyberghs,J., Van Hummelen,P. and Carlon,E. (2009) The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters. *Nucleic Acids Res.*, **37**, e53.
10. Atkinson,A.C. and Donev,A.N. (1992) *Optimum Experimental Designs*. Clarendon Press, Oxford.
11. Hooyberghs,J., Baiesi,M., Ferrantini,A. and Carlon,E. (2010) Breakdown of thermodynamic equilibrium for DNA hybridization in microarrays. *Phys. Rev. E*, **81**, 012901.
12. Gray,D.M. (1997) Derivation of nearest-neighbor properties from data on nucleic acid oligomers. I. Simple sets of independent sequences and the influence of absent nearest neighbors. *Biopoly.*, **42**, 783–793.
13. Walter,J.-C., Kroll,K.M., Hooyberghs,J. and Carlon,E. (2011) Nonequilibrium Effects in DNA Microarrays: A Multiplatform Study. *J. Phys. Chem. B*, **115**, 6732–6739.
14. Li,Y., Zon,G. and Wilson,W.D. (1991) Thermodynamics of DNA duplexes with adjacent G.cntdot.A mismatches. *Biochem.*, **30**, 7566–7572.
15. Li,Y., Zon,G. and Wilson,W.D. (1991) NMR and molecular modeling evidence for a G.A mismatch base pair in a purine-rich DNA duplex. *Proc. Natl Acad. Sci. USA*, **88**, 26–30.
16. SantaLucia,J. Jr and Turner,D.H. (1993) Structure of (rGGCGA GCC)2 in solution from NMR and restrained molecular dynamics. *Biochem.*, **32**, 12612–12623.
17. Lane,A., Ebel,S. and Brown,T. (1994) Properties of multiple G · A mismatches in stable oligonucleotide duplexes. *Eur. J. Biochem.*, **220**, 717–727.
18. Ebel,S., Brown,T. and Lane,A.N. (1994) Thermodynamic stability and solution conformation of tandem G · A mismatches in RNA and RNA · DNA hybrid duplexes. *Eur. J. Biochem.*, **220**, 703–715.
19. Morse,S.E. and Draper,D.E. (1995) Purine-purine mismatches in RNA helices: evidence for protonated G.A pairs and next-nearest neighbor effects. *Nucleic Acids Res.*, **23**, 302–306.
20. Seela,F. and Budow,S. (2008) Mismatch formation in solution and on DNA microarrays: how modified nucleosides can overcome shortcomings of imperfect hybridization caused by oligonucleotide composition and base pairing. *Mol. BioSyst.*, **4**, 232–245.
21. Chang,C.Y.(J.) and Stellwagen,N.C. (2011) Tandem GA Residues on Opposite Sides of the Loop in Molecular Beacon-like DNA Hairpins Compact the Loop and Increase Hairpin Stability. *Biochem.*, **50**, 9148–9157.
22. Trapp,C., Schenkelberger,M. and Ott,A. (2011) Stability of double-stranded oligonucleotide DNA with a bulged loop: a microarray study. *BMC Biophys.*, **4**, 20.
23. Golovkin,M.V., Matveeva,O.V. and Nechipurenko,Iu.D. (2009) Properties of hybridization isotherms upon binding of ligands on microchips. *Biofizika*, **54**, 820–823.
24. Pozhitkov,A.E., Boube,I., Brouwer,M.H. and Noble,P.A. (2010) Beyond Affymetrix arrays: expanding the set of known hybridization isotherms and observing pre-wash signal intensities. *Nucleic Acids Res.*, **38**, e28.
25. Irving,D., Gong,P. and Levicky,R. (2010) DNA Surface Hybridization: Comparison of Theory and Experiment. *J. Phys. Chem. B*, **114**, 7631–7640.
26. Fuchs,J., Fiche,J.B., Buhot,A., Calemczuk,R. and Livache,T. (2010) Salt Concentration Effects on Equilibrium Melting Curves from DNA Microarrays. *Biophys. J.*, **99**, 1886–1895.
27. Burden,C.J. and Binder,H. (2010) Physico-chemical modelling of target depletion during hybridization on oligonulceotide microarrays. *Phys. Biol.*, **7**, 016004.
28. Qamhieh,K., Wong,K.-Y., Lynch,G.C. and Pettitt,B.M. (2009) The melting mechanism of DNA tethered to a surface. *Int. J. Numer. Anal. Model.*, **6**, 474–488.
29. Gharaibeh,R.Z., Fodor,A.A. and Gibas,C.J. (2010) Accurate estimates of microarray target concentration from a simple sequence-independent Langmuir model. *PLoS One*, **5**, e14464.
30. Lee,C.-Y., Harbers,G.M., Grainger,D.W., Gamble,L.J. and Castner,D.G. (2007) Fluorescence, XPS, and TOF-SIMS surface chemical state image analysis of DNA microarrays. *J. Am. Chem. Soc.*, **129**, 9429–9438.
31. Binder,H., Preibisch,S. and Kirsten,T. (2005) Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. *Langmuir*, **21**, 9287–9302.
32. van Grinsven,B., Vanden Bon,N., Grieten,L., Murib,M., Janssens,S.D., Haenen,K., Schneider,E., Ingebrandt,S., Schöning,M.J., Vermeeren,V. *et al.* (2011) Rapid assessment of the stability of DNA duplexes by impedimetric real-time monitoring of chemically induced denaturation. *Lab. Chip.*, **11**, 1656–1663.
33. Hooyberghs,J. and Carlon,E. (2010) Hybridisation thermodynamic parameters allow accurate detection of point mutations with DNA microarrays. *Biosens. Bioelectron.*, **26**, 1692–1695.
34. Shanahan,H.P., Memon,F.N., Upton,G.J.G. and Harrison,A.P. (2012) Normalized Affymetrix expression data are biased by G-quadruplex formation. *Nucleic Acids Res.*, **40**, 3307–3315.