THE JOURNAL OF PHYSICAL CHEMISTRY B

ARTICLE

pubs.acs.org/JPCB

Nonequilibrium Effects in DNA Microarrays: A Multiplatform Study

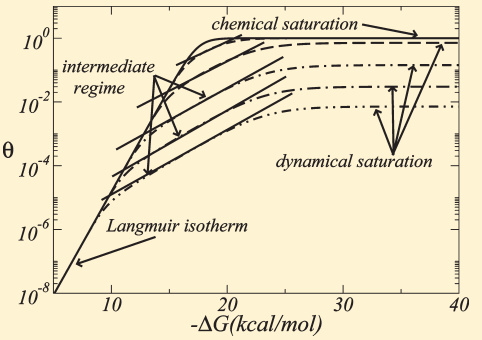
J.-C. Walter,[*,†] K. M. Kroll,[†] J. Hooyberghs,[†,‡,§] and E. Carlon[†]

[†]Institute for Theoretical Physics, KULeuven, Celestijnenlaan 200D, B-3001 Leuven, Belgium

[‡]Flemish Institute for Technological Research (VITO), Boeretang 200, B-2400 Mol, Belgium

[§]Hasselt University, Campus Diepenbeek, B-3590 Diepenbeek, Belgium

ABSTRACT: It has recently been shown that in some DNA microarrays the time needed to reach thermal equilibrium may largely exceed the typical experimental time, which is about 15 h in standard protocols (Hooyberghs et al. *Phys. Rev. E* **2010**, *81*, 012901). In this paper we discuss how this breakdown of thermodynamic equilibrium could be detected in microarray experiments without resorting to real time hybridization data, which are difficult to implement in standard experimental conditions. The method is based on the analysis of the distribution of fluorescence intensities I from different spots for probes carrying base mismatches. In thermal equilibrium and at sufficiently low concentrations, $\log I$ is expected to be linearly related to the hybridization free energy ΔG with a slope equal to $1/RT_{exp}$, where T_{exp} is the experimental temperature and R is the gas constant. The breakdown of equilibrium results in the deviation from this law. A model for hybridization kinetics explaining the observed experimental behavior is discussed, the so-called 3-state model. It predicts that deviations from equilibrium yield a proportionality of $\log I$ to $\Delta G/RT_{eff}$. Here, T_{eff} is an "effective" temperature, higher than the experimental one. This behavior is indeed observed in some experiments on Agilent arrays [Hooyberghs et al. *Phys. Rev. E* **2010**, *81*, 012901 and Hooyberghs et al. *Nucleic Acids Res.* **2009**, *37*, e53]. We analyze experimental data from two other microarray platforms and discuss, on the basis of the results, the attainment of equilibrium in these cases. Interestingly, the same 3-state model predicts a (dynamical) saturation of the signal at values below the expected one at equilibrium.



INTRODUCTION

While nucleic acid hybridization has been quantitatively characterized for strands binding in a bulk solution,[1] a microscopic-based understanding of hybridization in microarrays is still lacking. Compared to hybridization in solution, the characterization and understanding of physicochemical properties arising from hybridization in DNA microarrays is more challenging, and several groups dedicated their research to this subject[2−12] (see also the reviews in refs 13−15).

There are two main reasons which make this task difficult. The first one arises due to differences between different microarray platforms; for example the fabrication method, surface chemistry, the length of the surface-tethered probe strands, and their density might significantly differ from platform to platform. Since it is a priori rather difficult to assess the influence of the above-mentioned factors (for a theoretical analysis of factors influencing hybridization in microarrays see refs 13 and 15), it is necessary to investigate the properties and characteristics of each platform separately. The second reason arises due to the complexity of the samples (i.e., cell extract) which are used in most hybridization experiments. In these experiments, next to the intended hybridization between a probe and its complementary target in solution, a wide range of other reactions may occur: cross hybridization between partially complementary targets in solution, target folding, and cross hybridization between probes and partially complementary targets. Therefore, although a large number of biological data is available, it is difficult to disentangle from these experiments the properties of hybridization between targets and surface-bound probes from other spurious effects. A better understanding of the microscopic properties of hybridization in DNA microarrays involves the analysis of well-controlled experiments with simple target solutions.

In this context, we recall the outcome of recent experiments,[16] in which a solution containing a single target sequence at different concentrations was hybridized to a large number of mismatched probe sequences. It was shown that the general assumption of thermodynamic equilibrium, commonly used when discussing experimental data, does not hold. Apart from using a simple hybridization target solution, these experiments were carried out at standard conditions with respect to the buffer, the temperature, and the hybridization times. The lack of thermal equilibrium was shown to cause a decrease in specificity of microarrays;[16] therefore, it implies very practical consequences: it limits the performance of the microarrays below their maximal attainable level.

The aim of this paper is to discuss the microscopic origins of the breakdown of thermal equilibrium and to investigate experiments on different microarray platforms where such effects might

Received: February 12, 2011
Revised: March 30, 2011
Published: May 04, 2011

be of relevance. We start by reviewing the two-state Langmuir model of DNA hybridization and discuss a recent extension. The latter is termed three-state model and includes a long-lived partially hybridized state. This model offers a plausible explanation of the observed nonequilibrium behavior. We then turn our attention to three different platforms for which controlled hybridization experiments are available, and discuss evidence of the breakdown of equilibrium in these cases.

TWO- VS THREE-STATE KINETIC MODEL

Hybridization in solution as well as in DNA microarrays is commonly described by a two-state process using the so-called Langmuir model (see, e.g., ref 15 for a review). According to the Langmuir model the fraction of hybridized probes, θ, is given by

$$\frac{d\theta}{dt} = ck_1(1-\theta) - k_{-1}\theta \quad (1)$$

where k_1 and k_{-1} are the on- and off-rates and c is the target concentration in solution. The rates are linked to the equilibrium free energy ΔG via $k_1/k_{-1} = \exp(-\Delta G/RT)$, where T is the temperature and R the gas constant. Note that by convention $\Delta G < 0$, and strong target-probe affinities thus imply larger negative values for ΔG. The two-state hybridization model is assumed to be valid if the hybridizing sequences are sufficiently short such that intermediate states can be ignored. In solution, the zippering of two hybridizing sequences is sufficiently rapid, and the two-state assumption is expected to be applicable to oligomers containing up to 30 nucleotides.[1]

By imposing $d\theta/dt = 0$ on eq 1, we find the equilibrium value

$$\theta_{eq} = \frac{ck_1/k_{-1}}{1 + ck_1/k_{-1}} \quad (2)$$

which becomes

$$\theta_{eq} \approx ck_1/k_{-1} = c\exp(-\Delta G/RT) \quad (3)$$

in the limit $ck_1 \ll k_{-1}$ (i.e., low concentration and/or weak binding).

In this paper we consider, for fixed target concentrations c, θ vs ΔG plots (in this section) or, equivalently, intensity I vs ΔG (in the Experiments section). Equation 3 shows that far from chemical saturation ($\theta \ll 1$) the equilibrium isotherm is characterized by the linear relationship $\log\theta \propto \Delta G$, with a slope equal to $1/RT$. The full solution of eq 1 is

$$\theta(t) = \theta_{eq}(1 - e^{-t/\tau}) \quad (4)$$

where τ is the characteristic time such that

$$\tau^{-1} = ck_1 + k_{-1} \quad (5)$$

In order to understand the relaxation to equilibrium of sequences with different ΔG, we make a hypothesis concerning the rate constants. Experiments for hybridization in solution and in microarrays show that k_1 only weakly depends on the sequence composition.[17,18] Typically, there is a much stronger sequence dependence of the off-rate k_{-1}. We thus approximate

$$k_1 = \alpha, \quad k_{-1} = \alpha\exp(\Delta G/RT) \quad (6)$$

where α is sequence independent.

Figure 1a shows θ vs ΔG for a fixed concentration c (see eq 4) on a semilogarithmic scale at five different hybridization times (dashed lines), where the rates are chosen according to eq 6.

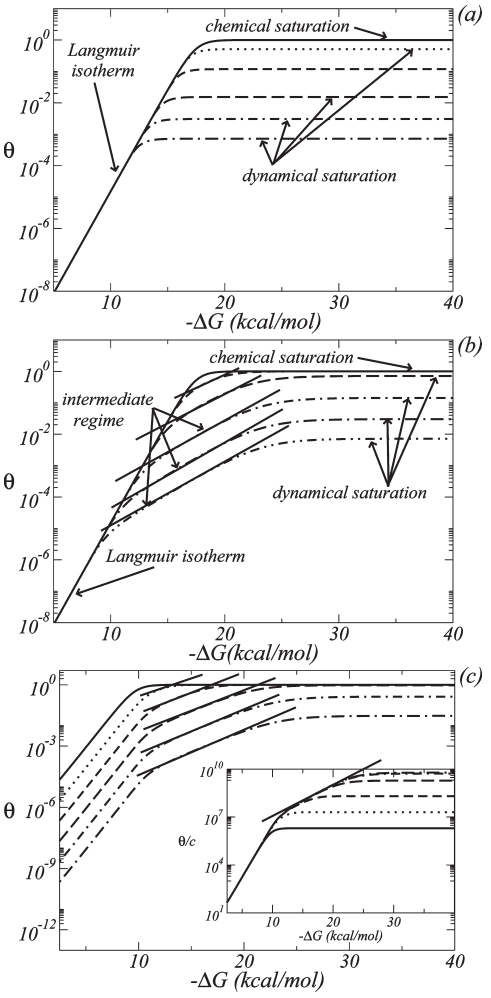


Figure 1. Plot of log θ vs ΔG for the two- and three-state models. (a) Two-state model with $c = 5$ pM, $\alpha = 10^4$ s^{-1} M^{-1}, the curve with solid line is the equilibrium isotherm (eq 2), the other lines correspond to plots of $\theta(t)$ (eq 4) for $t = 4$, 17, 86, 700, and 4000 h (bottom to top). (b) Three-state model for $c = 5$ pM, $\alpha = 10^5$ s^{-1} M^{-1}, $\omega = 1$ s^{-1} and $\gamma = 1/3$: Sum of the solutions of eqs 8a and 8b ($\theta = \theta_1 + \theta_2$) vs ΔG for different t (same as in (a)) with equilibrium value of θ (solid line). An effective temperature $T_{eff} > T_{exp}$ can be defined for the intermediate regime (solid straight lines). (c) Three-state model: Fixed hybridization time $t = 17$ h and different concentrations varying between 5 pM and 500 nM by steps of a factor 10 (from bottom to top). Remaining parameters identical to those in (b). System equilibrated for the largest concentration. The intermediate regime with an effective temperature T_{eff} can also be observed (solid straight lines). Inset: vertical axis shows fraction of hybridized probes divided by the concentration, i.e., θ/c.

The solid line coincides with the equilibrium regime (limiting case $t \to \infty$). The dashed lines and the equilibrium isotherm merge for small binding affinities. This indicates that weakly bound sequences equilibrate faster than more strongly bound ones. This behavior is a consequence of eqs 5 and 6: The relaxation time decreases for weak binding.

In the limiting case of strong binding (ΔG far from zero, i.e., large τ) and short times, the expansion of eq 4 yields

$$\theta(t) \approx \alpha ct \quad (7)$$

which no longer depends on ΔG.

dx.doi.org/10.1021/jp2014034 |J. Phys. Chem. B 2011, 115, 6732–6739

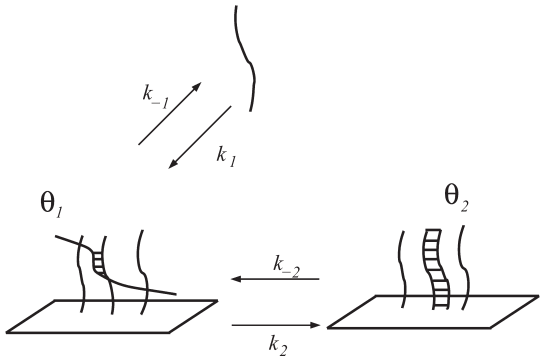


Figure 2. Sketch of the three-state model for hybridization in DNA microarrays. θ_1 and θ_2 are the fractions of partially and fully hybridized strands respectively. This model is specified by the four rate constants.

We refer to this as the dynamical saturation regime which appears as a constant limiting behavior in the θ vs ΔG plot as shown in Figure 1a. Instead, we refer to chemical saturation as the limit $\theta \rightarrow 1$, where all available probes are hybridized, and no further hybridization is possible. We emphasize that the existence of the dynamical saturation limit is linked to the choice of the rates k_1 and k_{-1} (see eqs 6), where we assumed that k_1 is sequence independent.

In view of the subsequent discussion of experimental data, we review a recently introduced extension of the two-state model.[16] This extension includes an intermediate long-lived configuration of partially hybridized molecules as sketched in Figure 2. The presence of the intermediate state slows down hybridization to a fully formed duplex due to the interaction of the target molecule with multiple probes. There are now four rate constants and the model is defined by two coupled linear equations

$$\frac{d\theta_1}{dt} = ck_1(1-\theta_1-\theta_2) + k_{-2}\theta_2 - (k_{-1}+k_2)\theta_1 \qquad (8a)$$

$$\frac{d\theta_2}{dt} = k_2\theta_1 - k_{-2}\theta_2 \qquad (8b)$$

where θ_1 and θ_2 denote respectively the fractions of partially and fully hybridized states. The ratios between the forward and backward rates are fixed by thermodynamics

$$\frac{k_1}{k_{-1}} = e^{-\Delta G'/RT} \qquad (9a)$$

$$\frac{k_2}{k_{-2}} = e^{-(\Delta G - \Delta G')/RT} \qquad (9b)$$

where ΔG and $\Delta G'$ are the hybridization free energies of the fully and partially hybridized states, respectively (both are free energy differences with respect to the unhybridized state). In general, θ_1 describes a distribution of partially hybridized states where only a fraction of the target-nucleotides is bound to a probe and the associated free-energy $\Delta G'$ can be considered as an effective ΔG. We expect the fully hybridized state to be energetically more favorable than the partially hybridized state, i.e., $\Delta G < \Delta G'$. The forward reaction rates $k_1 = \alpha$ and $k_2 = \omega$ are assumed to be sequence independent while the reverse rates are fixed by the relations in eqs 9. We then assume a monotonic link between ΔG and $\Delta G'$. Consider two sequences with different total hybridization free energies where the first one has higher binding affinity than the second $\Delta G_1 < \Delta G_2$ (Note that this may refer to the same probe with one being a perfect match and the other a mismatching sequence). For the partially hybridized states, we then expect $\Delta G'_1 < \Delta G'_2$. To establish a relationship between the two free energies, we make the simple assumption

$$\Delta G' \approx \gamma \Delta G \qquad (10)$$

where $\gamma < 1$ since $\Delta G < \Delta G'$. With this input we numerically solve eqs 8a and 8b. The solution can be expressed in terms of eigenvalues and eigenvectors of a 2×2 matrix. Figure 1b plots the solution for the fraction of hybridized probes $\theta = \theta_1 + \theta_2$ for five different times (dashed lines). The solid line corresponds to the equilibrium isotherm, which is identical to that of the two-state model in Figure 1a. In addition, we find a regime of dynamical saturation for very strongly bound sequences, which is analogous to that of the two-state model. The difference between the two models is the emergence of a new nonequilibrium regime. There, the logarithm of the fraction of hybridized probes varies linearly with ΔG with a slope equal to γ/RT, where the parameter γ is defined in eq 10. We note that a regime characterized by a slope smaller than expected from the equilibrium isotherm, is equivalent to introducing an effective temperature $T_{eff} = T_{exp}/\gamma$ higher than the experimental one.

We conclude this section with a discussion of the dependence of θ on the target concentration. Figure 1, panels a and b, shows θ at a fixed concentration with varying hybridization times. Let us now consider a protocol in which the concentration is varied and the time fixed. In all cases far from the chemical saturation limit $\theta \ll 1$, the data scale linearly with the concentration in the two as well as the three-state model. This behavior is a consequence of the form of eqs 1 and 8a: far from chemical saturation we approximate $ck_1(1-\theta) \approx ck_1$ in eq 1, and $ck_1(1-\theta_1-\theta_2) \approx ck_1$ in eq 8a, such that the concentration c only enters in the source term of the differential equations in the two models. As a result θ is proportional to c far from chemical saturation. This is illustrated in Figure 1c where the hybridization time is fixed to $t = 17$ h. The concentration values range from $c = 5$ pM to $c = 500$ nM in increments of a factor 10 (from bottom to top in the main graph). Only at the highest concentration the system has reached equilibrium after $t = 17$ h (solid line), at all other concentrations the system is not equilibrated. This is a consequence of a dependence of the equilibration time on c (see eq 5 for the two-state model). The inset of Figure 1c shows the intensity divided by the concentration. For the two lowest concentrations the θ/c plots overlap for the whole range of ΔG, while there is no overlap if θ is close to 1.

EXPERIMENTS

In this Section we discuss the experimental results obtained by different groups on three different microarray platforms. The details of the experimental setups are summarized in Table 1. These experiments are performed on samples containing a controlled number of different sequences in solution at some fixed concentrations. Agilent arrays held one, Codelink 8 and Affymetrix 150 different target sequences. In all cases the target hybridizes to a perfect match probe and to a number of mismatching probes. The emphasis of the analysis lies on the behavior of the signal intensities as a function of hybridization free energies ΔG. To estimate these free energies we use the nearest-neighbor model,[1] according to which ΔG can be written as a sum of parameters depending on the identity and orientation of neighboring nucleotides. Nearest-neighbor parameters from experiments on hybridizing strands in bulk solutions have been intensively investigated in

Table 1. Overview of Experimental Details of the Three Platforms Analyzed in This Paper

	Agilent	Affymetrix	Codelink
temperature (°C)	55, 65	45	45
hybridization time (hours)	17, 86	16	12
target length (nucleotides)	25, 30	14−25	70
probe length (nucleotides)	25, 30	25	30
number of different targets	1	150	8
target concentration (picomolar)	50, 500, 5000	1.4×10^{-3}-1.4×10^{3}	3000

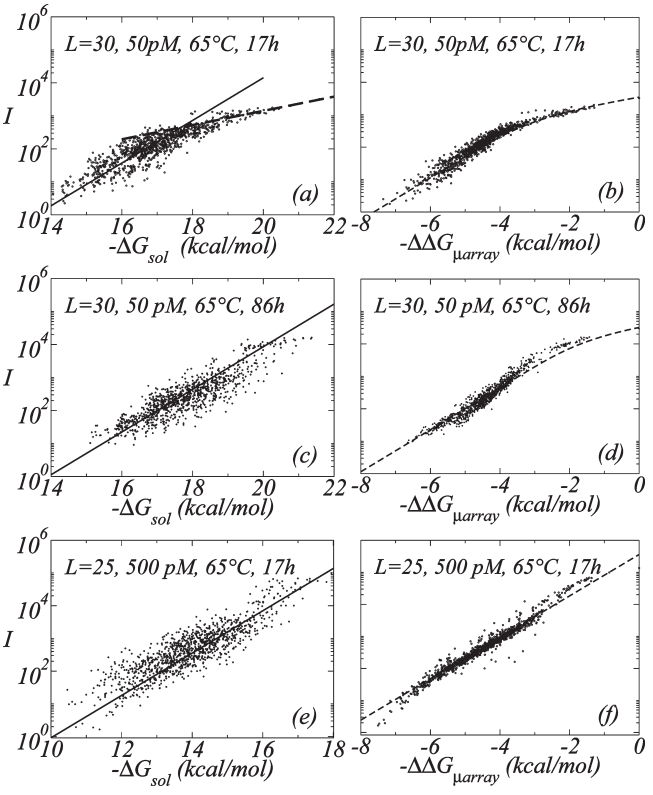


Figure 3. Left column: I vs ΔG_{sol}, right column: I vs $\Delta\Delta G_{\mu array}$ obtained under different experimental conditions in an Agilent platform. The full lines in (a, c, and e) are the Langmuir isotherm with T_{exp}, as expected at equilibrium. The dashed line in (a) is the Langmuir isotherm with $T_{eff} = 3T_{exp}$ (intermediate regime). The dashed lines (b, d, and f) are fitted using the three-state model with $\gamma = 1/3$, $k_1 = 5 \times 10^3\ M^{-1}\ s^{-1}$, $k_2 = 1\ s^{-1}$, and we introduce a shift of 7.5 kcal/mol on the ΔG axis. The amplitudes are slightly adjusted to match the intensity scale in the different experiments. (a and b): Data collapse using $\Delta\Delta G_{\mu array}$ of ref 20 (right) better than for parameters obtained in solution (left). (c and d) Intermediate regime almost vanished since equilibration is reached for nearly all sequences. The effect is more pronounced for $\Delta\Delta G_{\mu array}$ than for ΔG_{sol}. (e and f) Faster equilibration for shorter sequences: Compared to longer strands, the binding free energy here is smaller, resulting in a shorter characteristic time. (a and b) Data from ref 20. (c and d) Data from ref 16. (e and f) New experiment.

the past and are available in the literature (see, e.g., ref 19). As a first approximation, these parameters could also be used for hybridization in microarrays. In controlled experiments the parameters can also be extracted from microarray data, see for instance ref 20. In the following ΔG_{sol} refers to estimated free energies from experiments in bulk solution (data from ref 19) and $\Delta G_{\mu array}$ to estimated free energies from microarray data.

Custom Agilent Arrays. We review some experimental results recently obtained by two of us[16,20] and present novel experimental data of custom Agilent arrays. In this setup only one sequence is present in solution at various concentrations. The microarray is then designed to contain probes with either one, two, or no mismatches with respect to the target sequence in solution. In total there are about 10^3 different probe sequences, replicated 15 times on the array (for the details see ref 20). The experimental data are shown in Figure 3, which reproduces a pairwise comparison of the logarithm of the same intensities vs ΔG_{sol} (left column) and vs $\Delta G_{\mu\ array}$ (right column) for different target lengths and hybridization times.

Here, ΔG_{sol} is calculated based on nearest neighbor parameters by SantaLucia[19] at 1 M [Na+] concentration. The hybridization buffer in the experiment is the standard Agilent buffer which contains a buffering agent, monovalent cation and a ionic surfactant. It is known that the salt concentration influences the stability of the double helix and, thus also ΔG. However it usually affects all nearest-neighbor parameters by a constant salt-dependent constribution.[19] A change in salt concentration (and presumably of other chemicals) would result in a global shift of the horizontal axis in Figure 3. This, however, does not influence our analysis. More details about the effect of salt concentration and about electrostatic interactions are discussed in the Appendix 1. The $\Delta G_{\mu\ arrays}$ are obtained from a linear least-squares fit of the data, as explained in ref.[20] The left column of Figure 3 plots the data as function of $\Delta\Delta G_{\mu arrays} = \Delta G_{\mu array} - \Delta G_{PM\mu array}$, i.e., the subtraction from the hybridization free energy of the perfect match probe, which cannot be determined from the fitting procedure.[20] The similarities of the plots in both columns of Figure 3 show a correlation between ΔG_{sol} and $\Delta G_{\mu array}$.[20] However, when plotted as a function of $\Delta\Delta G_{\mu array}$ the data "collapse" better into single master curves, compared to the plots as a function of the nearest-neighbor model parameters in solution which are more spread.

Figure 3, panels a and b, shows the measured intensities for a target strand of length $L = 30$ hybridizing for 17 h, which is the typical hybridization time of the standard protocol used in biological experiments. In the cases (c and d) the time was increased to 86 h, while the cases (e and f) correspond to a shorter target sequence ($L = 25$) hybridized for 17 h. All data are obtained at a temperature of $T = 65$ °C. Experiments were also repeated under the same conditions at different target concentrations; a change in global concentration leads to an overall multiplicative factor in the measured intensities as explained in the Appendix 2.

The solid lines in Figure 3, panels a, c, and e, correspond to the slope expected in the case of thermodynamic equilibrium ($1/RT_{exp}$), the dashed line in Figure 3a corresponds to the Langmuir isotherm with an effective temperature $T_{eff} = 3T_{exp}$, whereas the dashed lines in (b, d, and f) are obtained from the solution of the three-state model. The measured intensity is assumed to be proportional to the fraction of hybridized probes $I = A\theta$. The scale factor A is slightly adjusted in the dashed lines of Figure 3, panels b, d, and f, to match for a global scale difference in intensities in the different experiments. Apart from the rescaling the dashed lines are obtained by using the same values for the adjustable parameters in the three-state model (the values are given in the caption). When increasing the hybridization time (from b−d) a larger fraction of the data tend to align along the expected equilibrium isotherm. Full equilibration is reached after 17 h when using a shorted target length (f), consistent with the three-state model

dx.doi.org/10.1021/jp2014034 |J. Phys. Chem. B 2011, 115, 6732–6739

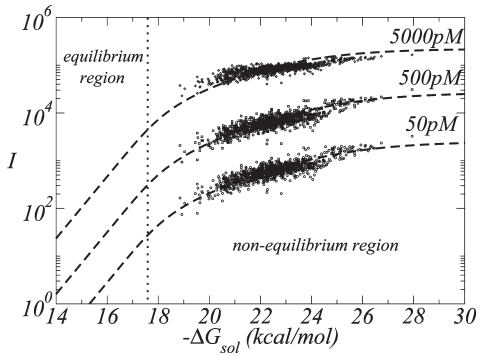


Figure 4. I vs ΔG for c = 50, 500, and 5000 pM, $L_{probe} = L_{target} = 30$, T_{exp} = 55 °C, and a hybridization time of 17 h. Lowering the temperature shifts the free energies to higher $-\Delta G$'s compared to Figure 3a. The dotted lines are a fit from the three-state model using the same parameters as Figure 3. Note that the shape for the largest concentration is flattened due to the saturation of the scanner around 10^5.

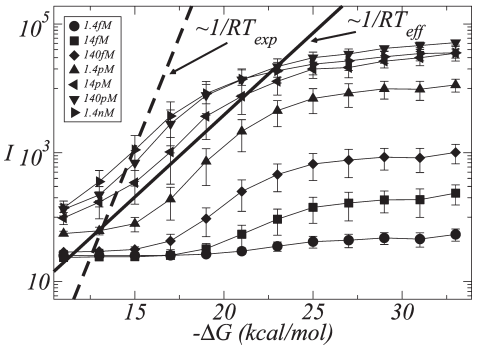


Figure 5. I vs ΔG curves for data set 1 obtained by Suzuki et al.[21] Binding free energies ΔG determined by means of nearest neighbor model using the values obtained in solution.[19] Saturation occurs at different values for different concentrations. The dotted and solid lines correspond to the Langmuir isotherm using respectively $T_{exp} = 318K$ and T_{eff} = 850 K $\approx 2.7T_{exp}$.

prediction: shorter sequences have lower binding affinity (lower $-\Delta G$), and are therefore expected to equilibrate faster.

Figure 4 shows experimental data for L = 30 and a hybridization time of 17 h at a temperature of T = 55 °C. Lowering the temperature leads to higher binding affinities (higher $-\Delta G$), which causes an increase in the relaxation time according to the three-state model. We therefore expect that these data would be even further away from thermodynamic equilibrium. This is indeed experimentally confirmed. In the present case none of the data, even for the lowest intensities attains the limiting behavior $I \propto e^{-\Delta G/RT}$, which suggests a breakdown of thermal equilibrium. Note that we did not plot the data as function of $\Delta\Delta G_{\mu array}$ since it is only possible to extract these parameters if equilibrium data are available, which is not the case at T = 55 °C. The dashed lines in Figure 4 are fits of the three-state model, with the same fitting parameters compared to the case T = 65 °C (see figure caption). The data at the highest affinities tend to bend toward a flat asymptotic limiting behavior, which suggests that they approach the dynamical saturation limit, discussed in the previous section. For the sake of completeness, we mention that the signal of the highest concentration is even more flattened due to the saturation of the scanner around 10^5. Note also that there is a linear scaling with the concentration and this is verified in a range of 2 orders of magnitude in c (from 50 to 5000 pM).

Custom Affymetrix Microarray. We discuss next the experiments by Suzuki et al.[21] In these experiments the arrays were custom-designed and synthesized on the Maskless Array Synthesizer platform with the Affymetrix NimbleExpress program. The target length was L = 25 nucleotides throughout the experiments. 150 artificial target sequences with a CG-content ranging from 32% to 72% were used such as to match typical distribution of probes of the Affymetrix *E. coli* array. The probe lengths varied from L = 14 to 25 nucleotides. The probes were shortened from the 5′-end. This implies that hybridization between probes with L < 25 nucleotides and targets produced duplexes with dangling ends afar from the microarray surface. For each target sequence three corresponding PM sequences and all possible mismatch combinations were present on the surface, i.e. three mismatches per position along the strand. In order to avoid terminal effects (it is known that mismatches close to the helix ends have different stability properties compared to mismatches in the center of the helix[22]), our analysis considers only those mismatches which are at least three nucleotides spaced from the ends. The concentration of each target sequence ranged from 1.4 fM to 1.4 nM incrementing in steps of factor ten; a concentration of a single target of 1.4 nM corresponds to a total concentration of 1.4 × 150 = 210 nM. The hybridization time was 16 h at T = 45 °C. The data cover a broad range of concentrations and ΔG, compared to the setup discussed in the previous section.

Figure 5 shows a log I vs ΔG_{sol} plot of the data for different values of the total concentration. A large number of data is available for each experiment (78 300 data points including the PM's and single mismatches of the 150 different targets). To help the visualization of the data were binned in intervals $[\Delta G - \delta/2, \Delta G + \delta/2]$ with δ = 2 kcal/mol. We then calculate the median value of the intensities within one ΔG interval of all probe sizes. The error bars in the graphs provide an estimate of the typical spreading of the data using the mean absolute deviation.

For very weak binding ($-\Delta G$ < 10 kcal/*mol*), the data of Figure 5 tend to a constant background value. The background level increases as the concentration is increased, due to aspecific binding, as expected. The slope of the dashed line in Figure 5 is $1/RT_{exp}$. The solid line (which provides a good extrapolation of the low $-\Delta G$ data) corresponds to an "effective" temperature of T_{eff} = 850 K (a value close to the one found in the analysis of the experiments in the Agilent platform). Thus, these experimental data cannot be reconciled with equilibrium thermodynamics. In addition, we note that for strong binding affinities the data reach a saturation value. This behavior would be consistent to the dynamical saturation predicted by the kinetic models discussed in the section describing the two- and three-state models.

In Figure 6 we plot the three-state model isotherm (thick solides lines) corresponding to an hybridization time of 16 h as in the experiment and using the values for the parameters k_1, k_2, and γ given in the caption. The model fits well the data within 2 orders of magnitude in concentration. To fit the data in the low $-\Delta G$ limit we added a background value (I_0): $I = I_0 + A\theta$ (where $\theta = \theta_1 + \theta_2$ is the total fraction of bound probes as calculated from the three-state model). Higher concentrations are also well fitted by the three-state model, provided the adjustment of the global fluorescence intensity level by rescaling the prefactor A. This rescaling is necessary since at high concentrations the intensities in Figure 5 no longer scale linearly with the concentration, differently from what observed in all the experiments on

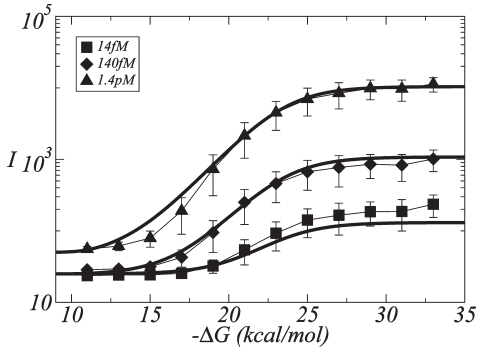


Figure 6. For three intermediate concentrations, the experimental data are well fit by the three-state model (solid thick lines). The parameters are: $\gamma = 0.374$, $k_1 = 10^5\ s^{-1}\ M^{-1}$ and $k_2 = 1\ s^{-1}$. The value are scaled by a factor $A = 2^{18}$ in order to match the experimental data. We have adjusted the value of the background I_0 for each concentration.

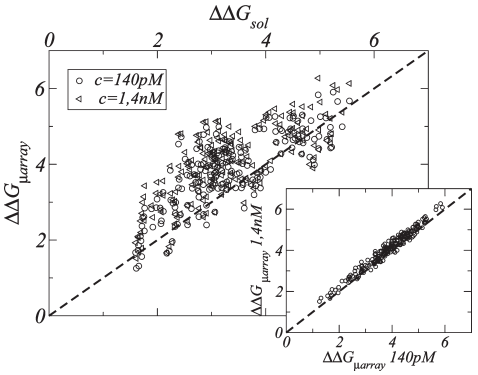


Figure 7. Main graph: $\Delta\Delta G_{\mu array}$ vs $\Delta\Delta G_{sol}$. The data of Suzuki et al[21] are fitted (probe size $L = 20$) using the linear least-squares method in the interval $-\Delta G$ between 13.5 and 19 kcal/mol for two concentrations 140 pM and 1.4 nM. We have used the effective temperatures $T_{eff} = 750$ K for 140 pM and $T_{eff} = 960$ K for 1.4 nM. The Pearson coefficient between the two axes is 0.784 for 140 pM and 0.735 for 1.4 nM, showing a moderate correlation. Inset: the results for both concentrations are plotted with respect to each other in order to check the cross-correlation (Pearson coefficient between the two concentrations: 0.983).

Agilent arrays (see Figure 4, appendix 2 and refs 16 and 20). This behavior is to be expected close to chemical saturation $\theta \rightarrow 1$, but not in the region 15 kcal/mol < $-\Delta G$ < 20 kcal/mol, where the measured intensities are very far from saturation. It suggests that other effects may play a role such as e.g. interactions (partial hybridization) between target molecules in solution. The experimental data of Figure 5 have also been analyzed in the recent literature[23,24] and were fitted with a model taking target depletion into account. Depletion arises when a relevant fraction of the target molecules in solution hybridizes to the microarray, so that during the experiment the concentration in solution decreases. Depletion would explain why the intensities at high binding affinity reach a saturation value below the chemical saturation,[23,24] without the need of invoking a nonequilibrium effect. However the depletion models of refs 23 and 24 did not addressed the differences between the experimental data and equilibrium thermodynamics in the low binding limit ($-\Delta G$ < 20 kcal/mol in Figure 5), whereas the nonequilibrium scenario developed within the three-state model could explain the experimental data in both regimes.

It remains to be proven that the observed behavior is due to breakdown of thermal equilibrium. One possibility would be to perform hybridization experiments at higher temperatures. According to the three-state model higher temperatures (recall that the experiments of Figure 5 are at T = 45 °C) would favor equilibrium due to a decrease in $-\Delta G$. This is indeed observed in Agilent arrays (see Figures 3, 4). We recall that other effects could also explain why the slope of the I vs ΔG data is smaller than what expected from the Langmuir model. A recent study[28] showed that synthesis errors (i.e., the synthesis of incorrect nucleotide) along the probe may also influence this slope. At present, further experiments are necessary to clarify the origin the discrepancy between the Affymetrix data and the equilibrium Langmuir model.

Finally, we used the experimental data to perform a linear least-squares fit to obtain estimates of the nearest neighbor parameters. The interval range was chosen such that $\log I \propto \Delta G$, far from any saturation behavior. We chose the two largest concentrations c = 140 pM and c = 1.4 nM, focusing on the hybridization of size $L = 20$ with a range of ΔG between 13.5 and 19 kcal/mol. In this range $\log I$ is approximately linear as a function of ΔG as seen in Figure 5. The fitting procedure is similar to the one described in ref 20, and it is obtained through singular value decomposition from the minimization of a quadratic cost function. The procedure provides 58 nearest neighbor parameters, of which 10 refer to perfect match parameters and 48 to single mismatch parameters. Due to degeneracies,[25] some of these parameters are not unique. In order to compare them with the corresponding parameters in solution, we consider the combination

$$\Delta\Delta G\begin{pmatrix} T & \underline{A} & C \\ A & \underline{C} & G \end{pmatrix} = \Delta G\begin{pmatrix} T & \underline{A} & C \\ A & \underline{C} & G \end{pmatrix} - \Delta G\begin{pmatrix} T & A & C \\ A & T & G \end{pmatrix}$$

which is instead unique.[20] Figure 7 shows a comparison between the parameters obtained from a fit of the microarray data compared to those obtained from hybridization/melting experiments in ref 19. The Pearson correlation coefficients are 0.784 for 140 pM and 0.735 for 1.4 nM, indicating a moderate correlation between the two sets. The inset shows a plot between parameters obtained from the 140 pM experiment vs those of the 1.4 nM experiment. There is a strong correlation between these two sets (Pearson coefficient 0.983) confirming the robustness of the extraction procedure of the parameters. However, note that this last analysis has to be considered with care. The validity of the estimation of the parameters in a regime where the use of an effective temperature is needed (and possibly a regime out-of-equilibrium) is still an open question.

Codelink Activated Slides. Last, we present another set of experiments on custom Codelink activated slides by Weckx et al.[9] In these experiments the target solution contained 8 different sequences. Note that only four were considered in the analysis (since the other four are richer in AT and bind very weakly to their probes, having a signal just above the background level). The target concentration, temperature and hybridization times were c = 3000 pM, T = 45 °C, and 12 h respectively. The target is a 70-mer, with a stretch of 30 nucleotides complementary to the probes and 40 nucleotide spacer terminating with a fluorophore. The spacer keeps the fluorophore away from the hybridization area and from the probe layer. Twelve probe sequences were associated to each target: one perfect match, three carry one mismatch, seven carry two mismatches, and one has three mismatches. This sums up to 48 data points. Each probe was

dx.doi.org/10.1021/jp2014034 |J. Phys. Chem. B 2011, 115, 6732–6739

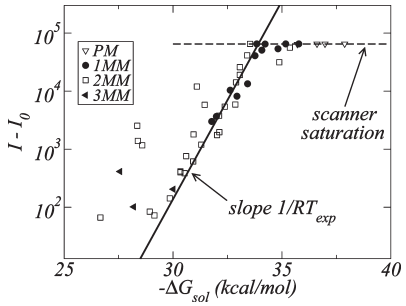


Figure 8. Background subtracted intensities vs hybridization free energies in solution for the experiments of Weckx et al.[9] on custom Codelink activated slides. The data are generally in agreement with equilibrium thermodynamics as seen from the alignement along a line with slope $1/RT_{exp}$. The experimental temperature is $T = 45$ °C for a hybridization time of 12 h.

replicated 12 times on the array, the analyzed signal is the median over the replicates.

The experimental results are shown in Figure 8. The data are plotted as a function of ΔG_{sol} calculated from the nearest-neighbor parameters for hybridization in solution by ref 26. The horizontal line corresponds to the maximum intensity detectable from the scanner (optical saturation). The steep line has a slope equal to $1/RT_{exp}$. Most probes with low ΔG are distributed in the close vicinity of the $1/RT_{exp}$ line. This suggests that the system had attained thermodynamic equilibrium. Although the number of data point is insufficient to extract free energy parameters, the alignment of the majority of the data points along a line with slope $1/RT_{exp}$ suggests a good degree of correlation with the hybridization free energy parameters measured in aqueous solution.[26]

In Codelink activated slides, the probes are not directly bound to the solid surface but to a three-dimensional polymeric coating which shifts them further away from the surface. This is to reduce the steric hindrance and to facilitate hybridization. The reduced surface probe density will also reduce electrostatic interactions from the negatively charged DNA probes. This is probably why these systems attain thermal equilibrium faster than those presented above.

CONCLUSION

Although several papers have discussed the application of physicochemical models to describe microarray data, so far most works focused on equilibrium thermodynamics.[2−12] The tacit assumption behind the use of these models is that hybridization in typical microarray experiments is sufficiently long (usually 15−17 h), so that thermal equilibrium can be considered to be attained. However, in standard microarrays, which are based on fluorescence detection, it is not possible to perform real time measurements of the hybridization signal. Then, it is difficult to test this hypothesis directly from experimental data.

In this paper we discussed how the attainment of equilibrium can be checked in experiments. We analyzed experiments from three different microarray platforms where the number of mismatches per probe varied from 0 to 3. This is an experimental setup which can offer very interesting insights about the physical properties of hybridization. First of all, given a sufficiently large number of mismatches, one can cover a broad range of intensities in an experiment even with a single hybridizing sequence, avoiding thus interactions between different targets in solution. Second, the attainment of thermodynamic equilibrium can be tested from an analysis of the slope of the experimental data in a log I vs $-\Delta G$ plot. We expect a slope equal to $1/RT_{exp}$ for equilibrium data, whereas a deviation to a different slope can be taken as an indication of the breakdown of equilibrium (sufficiently far from chemical saturation). We note also that in high density arrays used in biological experiments, the slope of log I vs $-\Delta G$ does not usually match the experimental temperature,[2,7] which could be a signature of nonequilibrium behavior, although different explanations exist.[28]

To explain the observed experimental results, we discussed a three-state model of hybridization. This is an intermediate between the unbound state and the fully hybridized state. For finite hybridization times, the three-state model predicts that sufficiently weakly bound targets are in equilibrium, while nonequilibrium effects are expected to occur for higher binding affinities. This feature is generally observed in experiments. In the limit of strong binding affinities the model predicts the existence of a dynamical saturation limit, in which the hybridization signal becomes independent of ΔG. These features are also observed in experiments on different platforms.

We expect that the absence of thermal equilibrium has important practical consequences for the functioning of the microarrays also in biological experiments with complex mixture of target sequences from mRNAs extracts. In that case different target sequences "compete" for hybridization to the same probe. Let us consider the case of two competing targets, one perfect matching to the probe and one carrying mismatches. The hybridization of the probe with the latter produces a so-called cross-hybridization signal. Assuming that the concentration of the targets is low, the following model is expected to hold

$$\theta = cf(\Delta G) + c'f(\Delta G') \tag{11}$$

where c is the concentration of the perfect matching target, ΔG is its hybridization free energy, and c' and $\Delta G'$ refer to the mismatched target. The function $f(x)$ describes the isotherm. In thermal equilibrium $f(x) = e^{-x/RT_{exp}}$. A high specificity, which is the desired working condition, corresponds to a perfect match signal dominating over the cross-hybridization contribution. This is obtained when the ratio $f(\Delta G)/f(\Delta G') \gg 1$, being c and c' fixed in an experiment. This ratio is maximal in thermodynamic equilibrium. If, for instance, $\Delta G' - \Delta G = 3$ kcal/mol (typically the free energy difference of one mismatch) at equilibrium one has $f(\Delta G)/f(\Delta G') = \exp((\Delta G' - \Delta G)/RT_{exp}) \approx 90$. In the nonequilibrium regime characterized by an effective temperature $T_{eff} = 3T_{exp}$ one has $f(\Delta G)/f(\Delta G') = \exp((\Delta G' - \Delta G)/RT_{eff}) \approx 4$. Thus the thermodynamic equilibrium corresponds to the highest specificity.

APPENDIX 1: EFFECT OF SALT CONCENTRATION

Isotherms different from the Langmuir model were discussed in the microarray literatures of the past decade.[27] Particular attention was devoted to the effect of electrostatic interactions, which arise during hybridization between a target molecule and a dense probe layer. As DNA is negatively charged, an additional electrostatic repulsive force may arise when a target molecule approaches the probes at the microarray surface.

In the context of a mean-field approximation Vainrub and Pettit[27] and Halperin et al.[29] derived the following equilibrium

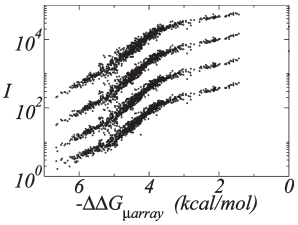


Figure 9. Data obtained with Agilent platforms for a target and probe length of 30 nucleotides at 65C with various concentrations c = 2, 10, 50, and 250 pM from bottom to top (data from ref 16). The linear scaling in concentration for the different curves means that the kink does not correspond to the approach toward chemical saturation where the linearity does not hold anymore.

model

$$\frac{\theta_{eq}}{c(1-\theta_{eq})} = e^{-\Delta G/RT - \Gamma(1+\theta_{eq})/RT} \qquad (12)$$

where, as in the main text, θ_{eq} is the fraction of hybridized probes, c is the target concentration, and ΔG is the hybridization free energy between a single probe and a single target. The term $\Gamma(1+\theta_{eq})/RT$ accounts for the electrostatic repulsion, where Γ is a constant independent of the sequence. Note that if $\Gamma = 0$ one recovers from eq 12 the Langmuir isotherm of eq 2. At high salt concentrations, electrostatic interactions are screened. Equation 12 shows that at larger coverages (θ_{eq} large) the electrostatic effects increase, reducing target-probe binding affinity.

The experiments discussed in this paper corresponds to the low concentration limit, as the measured intensities in all cases results proportional to the global target concentration c. In the limit $c \to 0$, eq 12 becomes

$$\theta_{eq} = ce^{-\Delta G/RT - \Gamma/RT} + \ldots \qquad (13)$$

where the dots indicate higher orders in c. Compared to the limit obtained from the Langmuir model, the electrostatic effects provides an additional contribution to the hybridization free energy (Γ). In the mean-field model discussed in the literature, Γ is proportional to the charge density of the unhybridized layer, to its thickness and to the length of the target DNA. It is however independent from the sequence composition. Γ is therefore the same for hybridization to a perfect matching probe or to a probe with one or two mismatches. In conclusion, in the low concentration limit, according to the above model, electrostatic effects can cause a uniform shift of the free energy scale, compared to hybridization free energies in solution.

APPENDIX 2: CONCENTRATION SCALING

As discussed throughout the manuscript all experiments shown are performed in a regime far from the chemical saturation limit which corresponds to $\theta \to 1$. In the limit $\theta \ll 1$, θ and thus the measured intensity is proportional to the target concentration c. We show this explicitly here for one set of experiments on Custom Agilent arrays. Figure 9 shows the same data as Figure 3, panels a and b, ($L_{probe} = L_{target} = 30$, $T = 65$ °C) for different concentrations c = 2, 10, 50, and 250 pM from bottom to top. The different curves clearly display an overall linear scaling relationship with respect to the concentration, both in the weak (equilibrium) and strong (nonequilibrium) binding regimes. The 5-fold increase in the concentration correspond to a 5-fold increase of the fluorescence intensity.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jean-charles.walter@fys.kuleuven.be.

ACKNOWLEDGMENT

We are grateful to S. Weckx for providing us the data for Figure 8. We thank Karen Hollanders (VITO) for help with the experiments. We acknowledge financial support from Research Foundation-Flanders (FWO) Grant No. G.0311.08 and from KULeuven Grant No. OT/07/034A.

REFERENCES

(1) Bloomfield, V. A.; Crothers, D. M.; Tinoco, I., Jr. *Nucleic Acids Structures, Properties and Functions*; University Science Books: Mill Valley, 2000.
(2) Held, G. A.; Grinstein, G.; Tu, Y. *Proc. Natl. Acad. Sci.* **2003**, *100*, 7575.
(3) Hekstra, D.; Taussig, A. R.; Magnasco, M.; Naef, F. *Nucleic Acids Res.* **2003**, *31*, 1962.
(4) Zhang, L.; Miles, M. F.; Aldape, K. D. *Nat. Biotechnol.* **2003**, *21*, 818.
(5) Hagan, M. F.; Chakraborty, A. K. *J. Chem. Phys.* **2004**, *120*, 4958.
(6) Binder, H.; Preibisch, S.; Kirsten, T. *Langmuir* **2005**, *21*, 9287.
(7) Carlon, E.; Heim, T. *Physica A* **2006**, *362*, 433.
(8) Fish, D. J.; *Nucleic Acids Res.* **2007**, *35*, 7197.
(9) Weckx, S.; Carlon, E.; De Vuyst, L.; Van Hummelen, P. *J. Phys. Chem. B* **2007**, *111*, 13583.
(10) Zhang, L.; Wu, C.; Carta, R.; Zhao, H. *Nucleic Acids Res.* **2007**, *35*, e18.
(11) Burden, C. J. *Phys. Biol.* **2008**, *5*, 16004.
(12) Naiser, T.; Kayser, J.; Mai, T.; Michel, W.; Ott, A. *BMC Bioinformat.* **2008**, *9*, 509.
(13) Halperin, A.; Buhot, A.; Zhulina, E. B. *J. Phys. Condens. Matter* **2006**, *18*, S463.
(14) Levicky, R.; Horgan, A. *Trends Biotechnol.* **2005**, *23*, 143.
(15) Binder, H. *J. Phys.: Condens. Matter* **2006**, *18*, S491.
(16) Hooyberghs, J.; Baiesi, M.; Ferrantini, A.; Carlon, E. *Phys. Rev. E* **2010**, *81*, 012901.
(17) Glazer, M.; *Anal. Biochem.* **2006**, *358*, 225.
(18) Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry, part III: The behavior of biological Macromolecules*; W. H. Freeman: New York, 1980.
(19) SantaLucia, J., Jr. *Proc. Natl. Acad. Sci.* **1998**, *95*, 1460.
(20) Hooyberghs, J.; Van Hummelen, P.; Carlon, E. *Nucleic Acids Res.* **2009**, *37*, e53.
(21) Suzuki, S.; Ono, N.; Furusawa, C.; Kashiwagi, A.; Yomo, T. *BMC Genomics* **2007**, *8*, 373.
(22) SantaLucia, J., Jr.; Hicks, D. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 415.
(23) Ono, N.; *Bioinformatics* **2008**, *24*, 1278.
(24) Burden, C. J.; Binder, H. *Phys Biol* **2010**, *7*, 016004.
(25) Gray, D. M. *Biopolymers* **1997**, *42*, 783.
(26) Peyret, N.; *Biochemistry* **1999**, *38*, 3468.
(27) Vainrub, A.; Pettitt, M. *Phys. Rev. E* **2002**, *66*, 041905.
(28) Naiser, T.; Kayser, J.; Mai, T.; Michel, W.; Ott, A. *Phys. Rev. Lett.* **2009**, *102*, 218301.
(29) Halperin, A.; Buhot, A.; Zhulina, E. *Clin. Chem.* **2004**, *50*, 2254.