Stability of a Surface-Bound Oligonucleotide Duplex Inferred from Molecular Dynamics: A Study of Single Nucleotide Defects Using DNA Microarrays

Thomas Naiser,[1] Jona Kayser,[1] Timo Mai,[1] Wolfgang Michel,[1] and Albrecht Ott[1,2,*]

[1]*Experimentalphysik I, Universität Bayreuth, D-95440 Bayreuth, Germany*

[2]*FR 7.2 Experimentalphysik, Universität des Saarlandes, D-66041 Saarbrücken, Germany*

(Received 4 April 2008; published 26 May 2009)

Microarray technology uses the sequence dependent hybridization (binding) affinity of surface-bound oligonucleotide strands for the quantification of complex nucleic acid mixtures. In spite of its huge potential in life science and medicine, microarray oligonucleotide hybridization remains far from being understood. Taking advantage of microarray combinatorial possibilities we show that, although surface bound, the hybridization affinities of single-base mismatched oligonucleotides can be derived from first principles using parameters from bulk.

DOI: 10.1103/PhysRevLett.102.218301 PACS numbers: 82.39.Pj, 87.15.H−, 87.85.fk

DNA is a biopolymer that carries genetic information. It is composed of four types of nucleotides or bases A, C, G, and T (for our purposes we can neglect noncanonical oligonucleotides here). Under appropriate conditions two simple DNA polymers, so-called single strands, can pair to form a helicoidal double strand (duplex). If only A · T or C · G base pairs occur in the duplex (in other words if the single strands are of complementary nucleotide sequence), the stability of the duplex is much higher than if only a few other base pairs occur. When the temperature of a solution containing double stranded DNA is raised above the melting temperature, it reversibly separates into two single strands (of complementary sequence). The reverse reaction, termed hybridization [1], is reduced if noncomplementary (mismatched) bases are present [2]. The stability of the DNA duplex is due to hydrogen bonding of the complementary pairs, as well as base stacking interactions between adjacent base pairs, which include van der Waals, electrostatic, and hydrophobic interactions. The widely used nearest-neighbor model predicts duplex stability [3–5]. It includes free-energy parameters for 10 nearest-neighbor doublets [6] as well as other parameters accounting for duplex initiation, A · T terminal pairs, and a symmetry penalty in case of self-complementary sequences.

Microarrays are used to determine and quantify the composition of complex mixtures of oligonucleotides. Microarrays consist of a surface with a regular array of spots (also called ''features''), each spot consisting of a large number of surface-bound single stranded DNA of only one particular sequence (the probe), which specifically binds to its complementary target sequence. With up to 10^6 spots of different sequence on a single surface, the measurement turns massively parallel. Possible applications include the determination of single nucleotide point mutations of genomic DNA and the quantitative measure of the different RNA strands transcribed from tissue or cells. Such a measure, termed ''expression profile,'' can be understood as a function of state of the genetic network of the tissue. Because of its high potential not only in medicine, pharmacology, and biology but also as a foundation for abstract modeling of biological processes, the microarray technique has inspired a major interest from physicists.

DNA hybridization on microarrays has recently been investigated [7–17]. In spite of the good knowledge about DNA stability in solution, predictions of probe-target hybridization affinities on microarrays remain empirical [8,10,16]. Data analysis from microarray experiments reveals a great deal of noise [7]. Studies [9,11–13] report that even the influence of a single-base defect on hybridization signal intensity cannot be predicted. In order to account for reduced binding on arrays, effective temperatures of 700 K [14] or 2130 K [10] were suggested. Often, such difficulties are attributed to secondary structure formation, excluded volume effects, or surface effects.

Here, we investigate the influence of single-base mismatches and base insertions or deletions (leading to base bulges upon duplex formation) on microarray hybridization theoretically and experimentally. Using *in situ* synthesized probe sets on microarray, we consider a simple system in which each hybridization assay has a single target sequence. This avoids intertarget binding as well as the competition between different target sequences for the same probe sequence. We choose the target length to be of the same order as that of the probes (about 20 base pairs), thus limiting excluded volume interactions or secondary structure effects. For a given target we produce arrays with feature sets compassing all possible single-base modifications. Using thermodynamic parameters from bulk we show that a double-ended molecular zipper model (where the strands need to open progressively from the ends in order to separate, like a zipper—see Ref. [18] for details) reproduces the experimentally observed hybridization signal intensities well. We take into account the heterogeneity of probe binding affinities caused by the unavoidable *in situ* synthesis errors (insertion, deletions, and base substitutions). Our results show that the noise of array measurements as it occurs in life science experiments

is due to the complexity of the applied DNA mixtures rather than the array based measurement itself.

Oligonucleotide microarrays are homemade by light-directed *in situ* synthesis [19,20] using the synthesis apparatus described by Naiser *et al.* [13]. (Further details on materials and methods used can be found in Ref. [18])

The measured intensities from surface-hybridized, fluorescently labeled target sequences (the ''hybridization signal'') are the lowest when defects are located in the middle of the probes. This feature generates a trough-shaped ''hybridization profile'' (Fig. 1). This result does not depend on whether the defect consists of mismatched or lacking bases (leading to bulges upon duplex formation) except for the case of positional degeneracy of the bulged base [17,21]. The result does not vary when the time left for hybridization is prolonged. Deviations from the average trough shape depend, however, on individual sequences. The characteristic length of these deviations exceeds a base pair. These observations taken together indicate that molecular zipping may play a role.

The double-ended zipper model [15,22,23] assumes that a DNA duplex can separate (unzip) from the ends only (as illustrated in Fig. 1 in Ref. [18]). It neglects the occurrence of bubbles (transient strand openings bounded by closed duplexes on both sides). Because of the large bubble initiation barrier (owing to stacking interactions towards both sides of a nucleotide) and the relatively short length of the duplexes, we expect bubble formation to be negligible. The zipper model accounts for a distribution of partially denatured duplex states. With N (the number of base pairs) and l and k (the positions of the zipper fork along the sequence of the oligonucleotide duplex from both ends), the partition function Z_D of the duplex [Eq. (1)] is the sum of the statistical weights $e^{\Delta G^0_{k,l}/RT}$ of all partially hybridized duplex states:

$$Z_D = \sum_{k=0}^{N-1} \sum_{l=k+1}^{N} e^{\Delta G^0_{k,l}/RT}, \tag{1}$$

where the $\Delta G^0_{k,l}$ represent the free-energy levels of partially denatured duplexes,

$$\Delta G^0_{k,l} = \sum_{i=1}^{k} \Delta g^0_i + \sum_{i=l+1}^{N} \Delta g^0_i, \qquad \Delta G^0_{0,l} = \sum_{i=l+1}^{N} \Delta g^0_i,$$
$$\Delta G^0_{k,N} = \sum_{i=1}^{k} \Delta g^0_i, \tag{2}$$

with Δg^0_i the free-energy-changes resulting from base-pair association. Their numerical values are known as unified nearest-neighbor parameters [6]. Secondary structures can be neglected since target sequences were chosen to prevent them. For simplicity, duplex initiation free energies and other possible corrective terms (which may arise from surface anchoring) have also been neglected. These terms change the duplex binding constant K by a constant factor only, which does not modify the shape of the defect profile. Upon dissociation, each single strand accounts for half of the duplex dissociation free energy ΔG^0_D. The duplex binding constant K is given by the ratio of the statistical weights of the bound to the unbound states:

$$K = \frac{Z_D}{Z_P Z_T} = \frac{Z_D}{e^{\Delta G^0_D/RT}}, \tag{3}$$

where Z_P and Z_T are the partition functions of the probes and targets (see Ref. [18] for details). For an analytic derivation of the positional influence, we consider a homopolymer replacing the canonical nearest-neighbor parameters Δg^0_i by an average Δg^0. We account for point defects at base position x by defect nearest-neighbor parameters $\Delta g^0_{\mathrm{def}}$. Using (1) and (2), with $Z_{D_{PM}}$ the partition function of the defect free duplex and $\delta\Delta g^0_{\mathrm{def}} = \Delta g^0_{\mathrm{def}} - \Delta g^0$, one easily finds to a good approximation [18]

$$Z_D(x) = Z_{D_{PM}} + (e^{(N-x)\Delta g^0/RT} + e^{x\Delta g^0/RT})(e^{\delta\Delta g^0_{\mathrm{def}}/RT} - 1). \tag{4}$$

The defect position dependent partition function $Z_D(x)$ [and the binding constant $K(x)$] indeed has a trough-shaped form (Ref. [18], Fig. 2). However, the predicted position dependent binding affinity from Eq. (4) varies by several orders of magnitude, which is much more pronounced than the variation in hybridization signal in our experimental observations. Thus Eq. (4) cannot explain our experimental observations.

In order to clarify the connection between calculated binding constants and experimentally observed hybridization signals, we need to find out how the hybridization signal intensity depends on duplex stability. To address this question experimentally we perform a hybridization assay with increasing probe length. Probe length is roughly

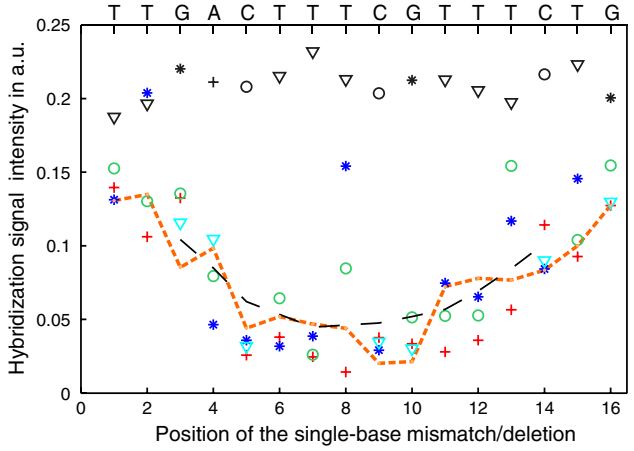

FIG. 1 (color online). Impact of single-base mismatches and deletions on the hybridization affinity for the probe sequence motif 3'-TTGACTTTCGTTTCTG-5'. The influence of defects depends on position and leads to a trough-shaped ''hybridization profile'' along the DNA sequence. Symbols: perfect match probe signal replicates (upper row of black symbols); MM probes with substituent bases A [gray (red) crosses], C [gray (green) circles], G [gray (blue) stars], T [gray (cyan) triangles]; moving average of all mismatch intensities (dashed black line); single-base deletions [gray (orange) dashed line].

proportional to the duplex free energy. Experimental results (Fig. 2) show a sigmoid relation between hybridization signal and probe length. The transition region over which an increase of the hybridization signal is observed extends over at least 13 base pairs, or $\delta\Delta G^0_{D_{37}} \approx 20$ kcal/mol.

The equilibrium between single stranded targets, probes, and hybridized duplexes $T + P_0 \rightleftarrows D$ is expected to follow a Langmuir-type adsorption isotherm [Eq. (5)], where the hybridization signal is given by the fraction of hybridized probes $\theta = [D]/[P_0]$. In our experiments targets are about 10- to 100-fold in excess, implying $[T] \approx [T_0]$.

$$\theta = \frac{[D]}{[P_0]} = \frac{K[T_0]}{1 + K[T_0]}. \quad (5)$$

With $T = 310\,\mathrm{K}$ and $[T_0] = 1\,\mathrm{nM}$, we see, however, that Eq. (5) does not agree with the experimental results (Fig. 2) since the slope of the isotherm is more pronounced.

Light-directed microarray synthesis unavoidably introduces single-base substitutions, insertions, deletions, and truncations. We consider this heterogeneity of the probe population to be around 10% per base [13]. We calculate the binding constants of the individual, mutated probe sequences using the zipper model for different error rates. The total hybridization signal is obtained by summing over the distribution of probes, where the contribution of each probe follows the Langmuir equation (5) and the errors are randomly distributed. This leads to a heterogeneous distribution of binding affinities, which on average causes an isotherm with a significantly broadened transition region from low to high binding affinity (broken lines in Fig. 2) as compared to the Langmuir isotherm. The result is similar to a so-called Sips isotherm [15], a generalization of the Langmuir isotherm in which the Langmuir single binding energy is replaced by a distribution. In order to meet the absolute experimental values, a free-energy penalty of 4.5 kcal/mol is added to the hybridized state. This free energy is larger than that of duplex initiation free energy often used in solution (2 kcal/mol) [24]. This discrepancy could be due to a high probe density on the surface, charge interactions with the surface, or other surface-based effects. Figure 2 shows that by taking into account probe heterogeneity due to synthesis defects, we can predict the experimentally observed hybridization signals by averaging over the binding constants K (which are obtained from the equilibrium free-energy values of the mutated probes).

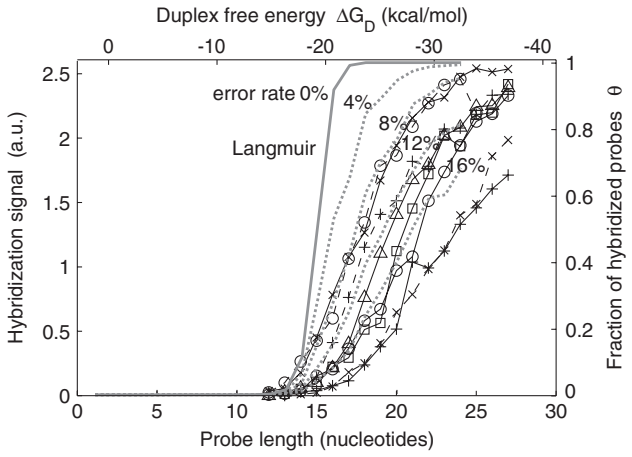

FIG. 2. Microarray hybridization signal as a function of probe lengths (probe length is about proportional to the duplex free energy). Experimentally, we observe an approximately linear increase of the hybridization signal over at least 13 base pairs ($\delta\Delta G^0_{D37} \approx 20$ kcal/mol). Different symbols indicate different sequence motifs. Error-free probes are described by the Langmuir isotherm (solid gray line), Eq. (5). Note the narrow transition region $\delta\Delta G^0_D \approx 6$ kcal/mol in that case (prediction for $T = 310$ K and a target concentration of 1 nM as in the experiment). When the heterogeneous distribution of binding affinities caused by synthesis defects (4%, 8%, 12%, and 16% of random base substitutions per nucleotide coupling step) is taken into account, the transition region of the Langmuir prediction $\theta(\Delta G^0_D)$ (broken lines) broadens and agrees well with the experimental results.

In order to calculate the binding constants of the individual sequences used in the experiments, which may deviate from the behavior of the homopolymer considered above, we evaluate the partition functions [Eqs. (1)–(3)] numerically. As in the previous paragraph, we obtain the predicted hybridization signal using Eq. (5), however, now averaging over the heterogeneous distribution of binding affinities due to synthesis defects. For simplicity we consider the defect nearest-neighbor free energies $\Delta g^0_{\mathrm{def}}$ as independent of the defect type. However, since we substitute a perfect matching nearest-neighbor pair, the free-energy difference $\delta\Delta g^0_{\mathrm{def}} = \Delta g^0_{\mathrm{def}} - \Delta g^0_{PM}$ still depends on the duplex sequence. A comparison between the experimental data and our numerical model allows us to determine an average value of $\Delta g^0_{\mathrm{def}}$ as a fit parameter.

In Fig. 3 we show that the experimental "hybridization profile" is numerically well reproduced by our mean field approximation. The best fit is achieved with $\Delta g_{\mathrm{def}} = \mathrm{const.} = 0.5$ to 1 kcal/mol (at $T = 325$ K), which agrees well with the prediction of the nearest-neighbor parameters in bulk solution reported in [25].

A more detailed assessment would require a precise knowledge about the distribution of synthesis defects. Our description is based on relative changes in hybridization intensity because absolute measurements are most difficult to perform with microarrays. Our results agree well with current empirical models [26]. The present study of defects shows that, in spite of the presence of a surface, hybridization of oligonucleotides of DNA microarrays can be predicted from first principles at thermodynamic equilibrium, while considering a molecular zipper and solution-based nearest-neighbor free-energy parameters.

Solution-based studies do not report a dominant influence of defect position on hybridization affinity as observed in microarrays. It is the probe binding heterogeneity (resulting from synthesis effects) which

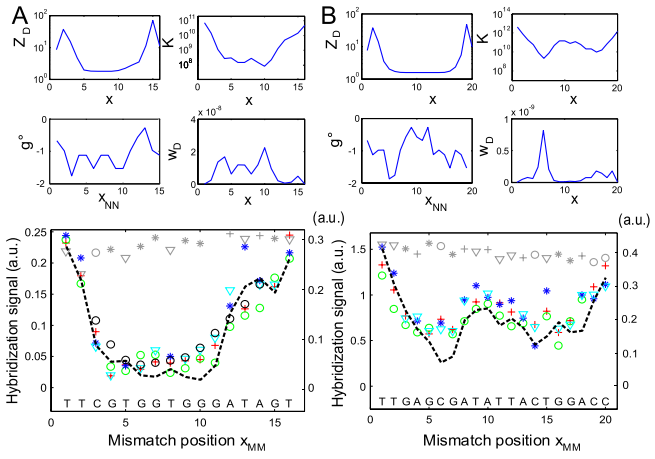

FIG. 3 (color online). Comparison of calculated vs experimentally determined hybridization affinities for two probe sequence motifs (a) and (b). Top left to bottom right: the partition function Z_D and the duplex binding constant K as a function of defect position x, the nearest-neighbor free energies Δg^0 of particular nearest-neighbor pairs as a function of nearest-neighbor pair position x_{NN}, and the statistical weight $w_D = e^{\Delta G_D^0/RT}$ of the completely dissociated duplex as a function of defect position. The decrease in $Z_D(x)$ at the duplex ends is due to the fact that only a single next neighbor pair is affected by a mismatch-base pair at the duplex end. The bottom subfigure shows the experimentally determined mismatch defect profile with A (red crosses), C (green circles), G (blue stars), T (cyan triangles). Perfect match control probes (gray in online color figure) have signals above 0.2 a.u. (left) and 1.25 a.u. (right). The calculated fraction θ of hybridized probes given as a function of defect position (dashed line) agrees well with the experimental data. $\Delta g^0{}_{\mathrm{def}} = 1$ kcal/mol at the simulation temperature of 325 K.

makes the positional influence of the mismatch visible in our experiments. This is similar to an increase of effective temperature [10,14,26]. The relation between the microarray hybridization signals and the duplex binding free energies shows that synthesis defects make the hybridization isotherm deviate from the expected Langmuir behavior; it is broadened like a so-called Sips isotherm. Different melting temperatures associated with different sequences can lead to important variations in target-probe binding affinity when the entire microarray is hybridized at a single temperature. As a consequence of the broadened isotherm, synthesis defects smooth these variations. Thus, these defects are useful in some respect.

Our work is an important step towards quantitative measurement with microarray technology. It encourages other quantitative, surface- or array-based applications aimed at studying oligonucleotide conformation, dynamics, or interaction. Rather than the physics at a solid-liquid interface, it is the inherent physical limits of DNA hybridization that make microarray data from biological samples noisy. Therefore, future work intending to improve array technology needs to address the fundamental but difficult question of the interactions of complex mixtures of DNA strands and their signal-to-noise ratio.

We thank M. Magnasco for fruitful discussions and J. Lehmann and D. Lee for critical reading. This work was funded by Universität Bayreuth.

*albrecht.ott@physik.uni-saarland.de

[1] W. Michel, T. Mai, T. Naiser, and A. Ott, Biophys. J. **92**, 999 (2007).
[2] J. W. Nelson, F. H. Martin, and I. Tinoco, Biopolymers **20**, 2509 (1981).
[3] H. Devoe and I. Tinoco, J. Mol. Biol. **4**, 500 (1962).
[4] P. N. Borer, B. Dengler, I. Tinoco, and O. C. Uhlenbeck, J. Mol. Biol. **86**, 843 (1974).
[5] K. J. Breslauer, R. Frank, H. Blocker, and L. A. Marky, Proc. Natl. Acad. Sci. U.S.A. **83**, 3746 (1986).
[6] J. SantaLucia, Proc. Natl. Acad. Sci. U.S.A. **95**, 1460 (1998).
[7] F. Naef, D. A. Lim, N. Patil, and M. Magnasco, Phys. Rev. E **65**, 040902(R) (2002).
[8] R. Mei *et al.*, Proc. Natl. Acad. Sci. U.S.A. **100**, 11 237 (2003).
[9] D. R. Dorris, A. Nguyen, L. Gieser, R. Lockner, A. Lublinsky, M. Patterson, E. Touma, T. J. Sendera, R. Elghanian, and A. Mazumder, BMC Biotechnology **3**, 6 (2003).
[10] G. A. Held, G. Grinstein, and Y. Tu, Nucleic Acids Res. **34**, e70 (2006).
[11] L. M. Wick, J. M. Rouillard, T. S. Whittam, E. Gulari, J. M. Tiedje, and S. A. Hashsham, Nucleic Acids Res. **34**, e26 (2006).
[12] A. Pozhitkov, P. A. Noble, T. Domazet-Loso, A. W. Nolte, R. Sonnenberg, P. Staehler, M. Beier, and D. Tautz, Nucleic Acids Res. **34**, e66 (2006).
[13] T. Naiser, T. Mai, W. Michel, and A. Ott, Rev. Sci. Instrum. **77**, 063 711 (2006).
[14] E. Carlon and T. Heim, Physica (Amsterdam) **362A**, 433 (2006).
[15] H. Binder, J. Phys. Condens. Matter **18**, S491 (2006).
[16] L. Zhang, C. L. Wu, R. Carta, and H. T. Zhao, Nucleic Acids Res. **35**, e18 (2007).
[17] T. Naiser, O. Ehler, J. Kayser, T. Mai, W. Michel, and A. Ott, BMC Biotechnology **8**, 48 (2008).
[18] See EPAPS Document No. E-PRLTAO-102-018924 for supplementary material on the derivation of the positional influence. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[19] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, Science **251**, 767 (1991).
[20] G. H. McGall, A. D. Barone, M. Diggelmann, S. A. Fodor, E. Gentalen, and N. Ngo, J. Am. Chem. Soc. **119**, 5081 (1997).
[21] J. Zhu and R. M. Wartell, Biochemistry **38**, 15 986 (1999).
[22] J. H. Gibbs and E. A. Dimarzio, J. Chem. Phys. **30**, 271 (1959).
[23] C. Kittel, Am. J. Phys. **37**, 917 (1969).
[24] J. SantaLucia and D. Hicks, Annu. Rev. Biophys. Biomol. Struct. **33**, 415 (2004).
[25] H. T. Allawi and J. SantaLucia, Jr., Biochemistry **36**, 10 581 (1997).
[26] T. Naiser, J. Kayser, T. Mai, W. Michel, and A. Ott, BMC Bioinformatics **9**, 509 (2008).