Supporting Material for “Measuring single molecule DNA hybridization by active control of DNA in a nanopore”

1 Target DNA and binding oligodeoxynucleotide (ODN) purity

In our experimental studies of DNA hybridization as a function of fishing (exposure) time and the subsequent dissociation of bound oligomers, we compared the behaviors of two syntheses of the 59 mer target (labeled 59A and 59B) and 10 mer ODN (labeled 10A and 10B) in different experiments. The two syntheses yield significantly different results. For instance, the fraction of detectable hybridization at a long fishing time is near 100% for 59A whereas the fraction of detectable hybridization at a long fishing time is only about 80% for 59B (see Figure 2 in the text). Since the experimental conditions were set to be the same for the two syntheses, the cause of the different behaviors is likely to be the impurity in the syntheses.

Gels were run to establish the purity of the target DNA and ODN samples. 59A and 59B were run on 14% denaturing PAGE gels at 20 W for 5 hours. 10A and 10B were run on a 20% denaturing PAGE gel at 28 W for 1.5 hours. Gels were stained with SYBR gold (Invitrogen) according to manufacturer’s protocol, and imaged on a UVP gel documentation system. The relative intensities of the imaged bands in each lane were analyzed used ImageJ[1]. For 59A, the full length product comprises 81% of the relative intensity of the lane, whereas for 59B the full length product comprises only 29% of the relative intensity seen in the sample (Fig. S1, *I* and *II*). Both 10A and 10B showed comparable and high purity (Fig. S1 *III*). The results of gel experiments indicate i) target DNA 59B is significantly less pure than 59A, which is consistent with the results of hybridization experiments (see Fig. 2 in the main text); and ii) ODN 10A and 10B both have high purity.

2 Fitting the hybridization model to data: Estimation of $\{q, r\}$ and their uncertainty

Here we discuss a mathematical formulation and the associated numerical algorithm for fitting the hybridization model to the data presented. In the hybridization model, the fraction of observable hybridization as a function of fishing time has the form

$$p(t) = q[1 - \exp(-rt)]$$

where q is the fraction of observable hybridization at long fishing time and r is the rate of hybridization. In the hybridization model, q and r are two unknown parameters to be determined from the data. Specifically, the goal of the fitting is 1) to determine the values

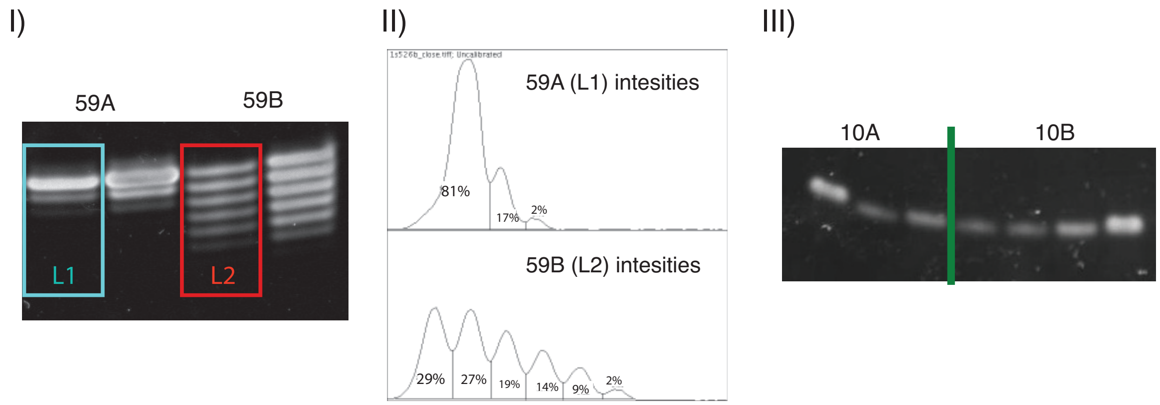

Figure S1: **Gel sample and relative lane intensity values for two syntheses of 59 mer target DNA (59A and 59B) on a common gel (I)-(II), and sample image of two syntheses of 10 mer binding ODN (10A and 10B) on a common gel**. (II) Lane intensity plots with percentage values for each peak computed using the ImageJ[1] software. Lanes L1 and L2 shown in main text Fig. 2 *B*. (III) ImageJ lane intensity plots showed ≥99% intensity with single peaks for both 10A and 10B.

of these two unknown parameters from data and 2) to determine the uncertainties in the estimated values of these two parameters.

2.1 Experimental Data

The experimental data consists of N entries. Each data entry is a vector of three components corresponding to the measurements at a given fishing time

$$\mathrm{D}_{\mathrm{exp}} = \{(t_j, n_j, p_j), \quad j = 1, 2, \ldots, N\}$$

where t_j is the fishing time, n_j is the number of fishing events at fishing time t_j, and p_j is the fraction of bound events at fishing time t_j.

2.2 Estimating the Values of q and r

We consider a measure of distance between the data and the fitting function

$$f(q, r) = \sum_{j=1}^{N} n_j \left(q(1 - \exp(-rt_j)) - p_j\right)^2$$

Mathematically, the fitting problem is a minimization problem of the form

$$\arg\min_{(q,r)} f(q, r)$$

This is a two-dimensional non-linear minimization problem. We notice that q appears linearly in the fitting function. Therefore, the one-dimensional minimization with respect to q while

r is fixed can be solved analytically. Let $q(r)$ denote the location of minimum of $f(q, r)$ with respect to q. Mathematically, $q(r)$ is

$$q(r) \equiv \arg\min_q f(q, r)$$

$q(r)$ satisfies

$$\begin{aligned}
& \left.\frac{\partial}{\partial q} f(q, r)\right|_{q=q(r)} = 0 \\
\Rightarrow \quad & \left. 2 \sum_{j=1}^{N} n_j \left(q \left(1 - \exp(-rt_j)\right) - p_j\right) \left(1 - \exp(-rt_j)\right) \right|_{q=q(r)} = 0 \\
\Rightarrow \quad & q(r) \sum_{j=1}^{N} n_j (1 - \exp(-rt_j))^2 = \sum_{j=1}^{N} n_j p_j (1 - \exp(-rt_j)) \\
\Rightarrow \quad & q(r) = \frac{\sum_{j=1}^{N} n_j p_j (1 - \exp(-rt_j))}{\sum_{j=1}^{N} n_j (1 - \exp(-rt_j))^2}
\end{aligned}$$

In function $f(q, r)$, we set $q = q(r)$ to write it as a function of r only.

$$g(r) \equiv f(q(r), r)$$

Thus, the two-dimensional minimization problem is reduced to the one-dimensional minimization problem

$$\arg\min_r g(r)$$

which, in turn, becomes the problem of solving the non-linear equation

$$g'(r) = 0$$

Let us derive an analytic expression for the derivative $g'(r)$.

$$g'(r) = \frac{d}{dr} f(q(r), r) = \left.\frac{\partial}{\partial q} f(q, r)\right|_{q=q(r)} \cdot q'(r) + \left.\frac{\partial}{\partial r} f(q, r)\right|_{q=q(r)}$$

Using $\frac{\partial}{\partial q} f(q,r)\Big|_{q=q(r)} = 0$, we get

$$
\begin{aligned}
g'(r) &= \frac{\partial}{\partial r} f(q,r)\Big|_{q=q(r)} \\
&= q(r)\frac{1}{N}\sum_{j=1}^{N} n_j t_j \exp(-rt_j)(q(r)(1-\exp(-rt_j)) - p_j) \\
&\equiv q(r)F(r)
\end{aligned}
$$

where function F(r) has the expression

$$F(r) = \sum_{j=1}^{N} n_j t_j \exp(-rt_j)(q(r)(1-\exp(-rt_j)) - p_j)$$

Notice that when $q = 0$, the fitting function is identically zero: $q(1-\exp(rt_j)) = 0$, which is not a meaningful case. For that reason, we exclude the case of $q(r) = 0$. Consequently, solving $g'(r) = 0$ is equivalent to solving

$$F(r) = 0$$

We use Newton's method with numerical differentiation to solve this non-linear equation.

2.3 Estimating the Uncertainty in the Determined Values of q and r

Let $q(\mathrm{D})$ and $r(\mathrm{D})$ denote the mapping from the data set D to the determined values of parameters q and r using the least square fitting described above.

To distinguish the experimental data set from the numerical data sets that we will consider below, we use $\mathrm{D_{exp}}$ to denote the experimental data set and use $\mathrm{D^{(i)}}$ to denote the i-th numerical data set. Let

$$q_{\mathrm{exp}} = q(\mathrm{D_{exp}}), \quad r_{\mathrm{exp}} = r(\mathrm{D_{exp}})$$

To estimate the uncertainty in q_{exp} and r_{exp}, we generate M independent numerical data sets:

$$\mathrm{D^{(i)}} = \left\{(t_j, n_j, p_j^{(i)}), \quad j = 1, 2, \ldots, N\right\}, \quad i = 1, 2, \ldots, M$$

In each numerical data set, t_j and n_j are the same as in the experimental data set, and $p_j^{(\mathrm{i})}$ is calculated as

$$p_j^{(\mathrm{i})} = \frac{m_j^{(\mathrm{i})}}{n_j}$$

where $m_j^{\text{(i)}}$ is a random sample drawn from the binomial distribution with number $= n_j$ and probability $= p_j$. In MATLAB, we can generate $m_j^{\text{(i)}}$ using

$$m_j^{\text{(i)}} = \text{sum}(\text{rand}(1, n_j)) < p_j)$$

We map each numerical data set to the determined values of q and r using the least square fitting described above

$$q^{\text{(i)}} = q(\text{D}^{\text{(i)}}), \quad r^{\text{(i)}} = r(\text{D}^{\text{(i)}})$$

We then calculate the standard deviations of $\{q^{\text{(i)}},\ i = 1, 2, \ldots, M\}$ and $\{r^{\text{(i)}},\ i = 1, 2, \ldots, M\}$.

$$\langle q \rangle \approx \frac{1}{M} \sum_{i=1}^{M} q^{\text{(i)}}, \quad \text{std}(q) \approx \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (q^{\text{(i)}} - \langle q \rangle)^2}$$
$$\langle r \rangle \approx \frac{1}{M} \sum_{i=1}^{M} r^{\text{(i)}}, \quad \text{std}(r) \approx \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (r^{\text{(i)}} - \langle r \rangle)^2}$$

The 95% confidence intervals for q_{exp} and r_{exp} can be approximated as

$$[q_{\text{exp}} - 2\text{std}(q), \quad q_{\text{exp}} + 2\text{std}(q)]$$
$$[r_{\text{exp}} - 2\text{std}(r), \quad r_{\text{exp}} + 2\text{std}(r)]$$

In the paper, we report std(q) and std(r) as the standard errors for q_{exp} and r_{exp}.

3 Modeling duplex lifetime samples between 0.7 ms and 300 ms

Let τ denote the (random) lifetime of a DNA duplex when it is pulled against the pore by a voltage reversal to -20 mV. We need to point out a few aspects of experimentally measured samples of τ before we write out the mathematical model.

- There is a lower cut-off threshold for τ at $x_1 = 0.7$ ms. Hybridization events with dwell time $\tau < x_1$ are not detected. In experiments, the capacitive transient after a sudden voltage change obscures the electric current measurement, and thus, prevents us from detecting a hybridization event with very short dwell time.

- There is an upper cut-off threshold for τ at $x_2 = 300$ ms. Events with dwell time $\tau > x_2$ are recognized as hybridization events. In that case, DNAs are ejected at 300 ms by changing the voltage to -120 mV. As a result, the values of the dwell time larger than 300 ms are not measured.

- At low probing voltage, the measured samples of τ between $x_1 = 0.7$ ms and $x_2 = 300$ ms do not follow a single exponential distribution.

We model the dwell time between $x_1 = 0.7$ ms and $x_2 = 300$ as the sum of two exponential modes. Mathematically, we use the probability density

$$p(\tau, v) = \alpha\lambda_1 \exp(-\lambda_1(\tau - x_1)) + (1-\alpha)\lambda_2 \exp(-\lambda_2(\tau - x_1))$$

There are three parameters in the model, and we put them into one parameter vector: $v = (\lambda_1, \lambda_2, \alpha)$. Here λ_1 is the rate of slow dissociation, λ_2 the rate of fast dissociation, relatively within the range of 0.7 to 300 ms, and α is the fraction of slow dissociation. Note that there is at least one more dissociation mode with even slower rate in the range of $\tau > 300$ ms. For less pure DNA targets (59B), there is also at least one dissociation mode with very fast rate in the range of $\tau < 0.7$ ms, which is not observable in our experiments.

To deal with the constraint $\tau > x_1$ in a more mathematically convenient way, we consider random variable $Y = (\tau - x_1)$. The constraint $\tau < x_2$ becomes $Y < t_c$ where $t_c = x_2 - x_1$ is the cut-off threshold for Y. The conditional probability density of Y truncated at $t_c = x_2 - x_1$ has the expression

$$p(y, v|Y < t_c) = \frac{\alpha\lambda_1 \exp(-\lambda_1 y) + (1-\alpha)\lambda_2 \exp(-\lambda_2 y)}{1 - \alpha\exp(-\lambda_1 t_c) - (1-\alpha)\exp(-\lambda_2 t_c)}, \qquad y < t_c$$

This is our mathematical model for the duplex lifetime constrained to the range of (0.7 ms, 300 ms). The experimental data consists of all measured dwell times between 0.7 ms and 300 ms. In terms of random variable Y, the experimental data has the form

$$\mathrm{D}_{\mathrm{exp}} = \{Y_j, \quad j = 1, 2, \ldots, n\}$$

where n is the number of dwell time samples between $x_1 = 0.7$ ms and $x_2 = 300$ ms. Below we are going to do two things

1. use the maximum likelihood estimation to determine the values of λ_1, λ_2 and α; and
2. estimate the uncertainties in the estimated values of these three parameters

3.1 Estimating the Values of λ_1, λ_2, and α

The log likelihood function is

$$\begin{aligned}\log L(v) &= \sum_{j=1}^{n} \log p(Y_j, v|Y < t_c) \\ &= \sum_{j=1}^{n} \log[\alpha\lambda_1 \exp(-\lambda_1 Y_j) + (1-\alpha)\lambda_2 \exp(-\lambda_2 Y_j)] \\ &\quad - n\log[1 - \alpha\exp(-\lambda_1 t_c) - (1-\alpha)\exp(-\lambda_2 t_c)]\end{aligned}$$

To find the maximum, we differentiate the log likelihood function with respect to $v = (\lambda_1, \lambda_2, \alpha)$.

$$\begin{aligned}\frac{\partial}{\partial \lambda_1}(\log L(v)) =&\alpha \cdot \sum_{j=1}^{n} \frac{\exp(-\lambda_1 Y_j)(1-\lambda_1 Y_j)}{\alpha\lambda_1 \exp(-\lambda_1 Y_j) + (1-\alpha)\lambda_2 \exp(-\lambda_2 Y_j)} \\ &- \alpha \cdot n \cdot \frac{\exp(-\lambda_1 t_c) t_c}{1-\alpha\exp(-\lambda_1 t_c) - (1-\alpha)\exp(-\lambda_2 t_c)}\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \lambda_2}(\log L(v)) =&(1-\alpha) \cdot \sum_{j=1}^{n} \frac{\exp(-\lambda_2 Y_j)(1-\lambda_2 Y_j)}{\alpha\lambda_1 \exp(-\lambda_1 Y_j) + (1-\alpha)\lambda_2 \exp(-\lambda_2 Y_j)} \\ &- (1-\alpha) \cdot n \cdot \frac{\exp(-\lambda_2 t_c) t_c}{1-\alpha\exp(-\lambda_1 t_c) - (1-\alpha)\exp(-\lambda_2 t_c)}\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \alpha}(\log L(v)) =&\sum_{j=1}^{n} \frac{\lambda_1 \exp(-\lambda_1 Y_j) - \lambda_2 \exp(-\lambda_2 Y_j)}{\alpha\lambda_1 \exp(-\lambda_1 Y_j) + (1-\alpha)\lambda_2 \exp(-\lambda_2 Y_j)} \\ &+ n \cdot \frac{\exp(-\lambda_1 t_c) - \exp(-\lambda_2 t_c)}{1-\alpha\exp(-\lambda_1 t_c) - (1-\alpha)\exp(-\lambda_2 t_c)}\end{aligned}$$

So the maximum likelihood estimate v satisfies the equation

$$\overrightarrow{F}(v) = 0$$

where function $\overrightarrow{F}(v) = (F_1(v), F_2(v), F_3(v))^T$ is defined as

$$\begin{aligned}F_1(v) =&\sum_{j=1}^{n} \frac{\exp(-\lambda_1 Y_j)(1-\lambda_1 Y_j)}{\alpha\lambda_1 \exp(-\lambda_1 Y_j) + (1-\alpha)\lambda_2 \exp(-\lambda_2 Y_j)} \\ &- n \cdot \frac{\exp(-\lambda_1 t_c) t_c}{1-\alpha\exp(-\lambda_1 t_c) - (1-\alpha)\exp(-\lambda_2 t_c)}\end{aligned}$$

$$\begin{aligned}F_2(v) =&\sum_{j=1}^{n} \frac{\exp(-\lambda_2 Y_j)(1-\lambda_2 Y_j)}{\alpha\lambda_1 \exp(-\lambda_1 Y_j) + (1-\alpha)\lambda_2 \exp(-\lambda_2 Y_j)} \\ &- n \cdot \frac{\exp(-\lambda_2 t_c) t_c}{1-\alpha\exp(-\lambda_1 t_c) - (1-\alpha)\exp(-\lambda_2 t_c)}\end{aligned}$$

$$\begin{aligned}F_3(v) =&\sum_{j=1}^{n} \frac{\lambda_1 \exp(-\lambda_1 Y_j) - \lambda_2 \exp(-\lambda_2 Y_j)}{\alpha\lambda_1 \exp(-\lambda_1 Y_j) + (1-\alpha)\lambda_2 \exp(-\lambda_2 Y_j)} \\ &+ n \cdot \frac{\exp(-\lambda_1 t_c) - \exp(-\lambda_2 t_c)}{1-\alpha\exp(-\lambda_1 t_c) - (1-\alpha)\exp(-\lambda_2 t_c)}\end{aligned}$$

We use Newton's method to solve this non-linear system where the Jacobi is approximated using numerical differentiation. Once the values of $v = (\lambda_1, \lambda_2, \alpha)$ are determined, we can calculate two probabilities

$$\begin{aligned}
p_1 =& \text{ probability that the two dissociation modes described by } v = (\lambda_1, \lambda_2, \alpha) \\
& \qquad \text{produces a dwell time sample below } x_1 = 0.7 \text{ ms} \\
p_2 =& \text{ probability that the two dissociation modes described by } v = (\lambda_1, \lambda_2, \alpha) \\
& \qquad \text{produces a dwell time sample above } x_2 = 300 \text{ ms}
\end{aligned}$$

In terms of parameter values $v = (\lambda_1, \lambda_2, \alpha)$, probabilities p_1 and p_2 have the expressions

$$\begin{aligned}
p_1 =& \frac{\alpha \exp(\lambda_1 x_1) + (1-\alpha)\exp(\lambda_2 x_1) - 1}{\alpha \exp(\lambda_1 x_1) + (1-\alpha)\exp(\lambda_2 x_1)} \\
p_2 =& \frac{\alpha \exp(-\lambda_1 (x_2 - x_1)) + (1-\alpha)\exp(-\lambda_2 (x_2 - x_1))}{\alpha \exp(\lambda_1 x_1) + (1-\alpha)\exp(\lambda_2 x_1)}
\end{aligned}$$

The number of dwell time samples below $x_1 = 0.7$ ms that are from the two dissociation modes described by $v = (\lambda_1, \lambda_2, \alpha)$ is

$$n_1 = n \cdot \frac{p_1}{1 - p_1 - p_2}$$

The number of dwell time samples above $x_2 = 300$ ms that are from the two dissociation modes described by $v = (\lambda_1, \lambda_2, \alpha)$ is

$$n_2 = n \cdot \frac{p_2}{1 - p_1 - p_2}$$

Note that in general n_1 is not the number of all dwell time samples below $x_1 = 0.7$ ms. In experiments, the dwell time samples below $x_1 = 0.7$ ms (which are not detected) may include samples from very fast dissociation modes that are not well manifested in the range of 0.7 ms to 300 ms. Similarly in experiments, the dwell time samples above $x_2 = 300$ ms may include samples from very slow dissociation modes that are not well manifested in the range of 0.7 ms to 300 ms.

3.2 Estimating the uncertainty in the determined parameters $(\lambda_1, \lambda_2, \alpha)$

Let $\lambda_1(\mathrm{D})$, $\lambda_2(\mathrm{D})$ and $\alpha(\mathrm{D})$ denote the mapping from the data set D to the determined values of parameters λ_1, λ_2 and α using the maximum likelihood method described above.

To facilitate the discussion below, we use $\mathrm{D}_{\mathrm{exp}}$ to denote the experimental data set and use $\mathrm{D}^{(\mathrm{i})}$ to denote the i-th numerical data set. Let

$$\begin{aligned}
\lambda_{1,\mathrm{exp}} &= \lambda_1(\mathrm{D}_{\mathrm{exp}}) \\
\lambda_{2,\mathrm{exp}} &= \lambda_2(\mathrm{D}_{\mathrm{exp}}) \\
\alpha_{\mathrm{exp}} &= \alpha(\mathrm{D}_{\mathrm{exp}})
\end{aligned}$$

To estimate the uncertainty in $\lambda_{1,\text{exp}}$, $\lambda_{2,\text{exp}}$ and α_{exp}, we generate M independent numerical data sets:

$$\mathrm{D}^{(\mathrm{i})} = \left\{ Y_j^{(i)}, \quad j = 1, 2, \ldots, n \right\}, \quad i = 1, 2, \ldots, M$$

Basically, each numerical data set is a collection of n independent samples drawn from the conditional distribution

$$p(y, v | Y < t_c) = \frac{\alpha\lambda_1 \exp(-\lambda_1 y) + (1-\alpha)\lambda_2 \exp(-\lambda_2 y)}{1 - \alpha \exp(-\lambda_1 t_c) - (1-\alpha)\exp(-\lambda_2 t_c)}, \qquad y < t_c$$

We draw n independent samples from the conditional distribution in two steps.

1. Draw m independent samples from the unconditional distribution

   $$p(y, v) = \alpha\lambda_1 \exp(-\lambda_1 y) + (1-\alpha)\lambda_2 \exp(-\lambda_2 y)$$

   Here m is sufficiently larger than n, for example, $m = 2n$. In MATLAB, m independent samples from the unconditional distribution are generated using

   $$\begin{aligned} &rates = lambda2 + (lambda1 - lamda2). * (rand(1, m) < alpha) \\ &samples = -log(rand(1, m))./rates \end{aligned}$$

2. Collect all samples out of the m samples that satisfy the constraint $Y < tc$ and then take n samples. In MATLAB, this is done using

   $$\begin{aligned} samples &= samples(find(samples < tc)) \\ samples &= samples(1 : n) \end{aligned}$$

We map each numerical data set to the determined values of λ_1, λ_2 and α using the maximum likelihood method described above

$$\begin{aligned} \lambda_1^{(\mathrm{i})} &= \lambda_1(\mathrm{D}^{(\mathrm{i})}) \\ \lambda_2^{(\mathrm{i})} &= \lambda_2(\mathrm{D}^{(\mathrm{i})}) \\ \alpha^{(\mathrm{i})} &= \alpha(\mathrm{D}^{(\mathrm{i})}) \end{aligned}$$

We then calculate the standard deviations of $\{\lambda_1^{(\mathrm{i})}\}$, $\{\lambda_2^{(\mathrm{i})}\}$ and $\{\alpha^{(\mathrm{i})}\}$.

$$\begin{aligned} \mathrm{std}(\lambda_1) &\approx \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(\lambda_1^{(\mathrm{i})} - \langle\lambda_1\rangle)^2} \\ \mathrm{std}(\lambda_2) &\approx \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(\lambda_2^{(\mathrm{i})} - \langle\lambda_2\rangle)^2} \\ \mathrm{std}(\alpha) &\approx \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(\alpha^{(\mathrm{i})} - \langle\alpha\rangle)^2} \end{aligned}$$

The 95% confidence intervals for $\lambda_{1,\text{exp}}$, $\lambda_{2,\text{exp}}$ and α_{exp} can be approximated as

$$\begin{gathered}[\lambda_{1,\text{exp}} - 2\text{std}(\lambda_1),\ \ \lambda_{1,\text{exp}} + 2\text{std}(\lambda_1)] \\ [\lambda_{2,\text{exp}} - 2\text{std}(\lambda_2),\ \ \lambda_{2,\text{exp}} + 2\text{std}(\lambda_2)] \\ [\alpha_{\text{exp}} - 2\text{std}(\alpha),\ \ \alpha_{\text{exp}} + 2\text{std}(\alpha)]\end{gathered}$$

In the paper, we report $\text{std}(\lambda_1)$, $\text{std}(\lambda_2)$ and $\text{std}(\alpha)$ as the standard errors for $\lambda_{1,\text{exp}}$, $\lambda_{2,\text{exp}}$ and α_{exp}.

4 Repetition of Hybridization Experiments to Test Consistency

To test the consistency of our hybridization results, experiments were repeated twice for the 59A-10A and 59B-10B syntheses, and for the 59A-20 synthesis. In this section, we report the measured hybridization probabilities and their standard deviations at multiple fishing (exposure) times, for each experiment, and we compare the model parameters determined from different experiments.

Calculation of the standard deviation for the measured probability at each fishing time is done as follows. Recall that for each fishing time t_j, n_j denotes the number of fishing events and $p_j^{(e)}$ denotes the exact probability of hybridization. The measured number of bound events X_j at fishing time t_j is a binomially distributed random variable, with distribution probability $p_j^{(e)}$ and number of trials n_j. The expected value of X_j is $\text{E}[X_j] = np_j^{(e)}$ and the variance is $\text{Var}[X_j] = np_j^{(e)}(1 - p_j^{(e)})$. The exact probability of hybridization is approximated by the measured fraction of hybridization: $p_j^{(e)} \approx p_j = X_j/n_j$. Note that while $p_j^{(e)}$ is a deterministic number, the measured fraction of hybridization $p_j = X_j/n_j$ is still a random number. The expected value of p_j is $\text{E}[p_j] = p_j^{(e)}$ and the variance is $\text{Var}[p_j] = p_j^{(e)}(1 - p_j^{(e)})/n_j$. Since the exact value of $p_j^{(e)}$ is unknown, the variance of p_j is approximated as: $\text{Var}[p_j] \approx p_j(1 - p_j)/n_j$. From the variance, we calculate the standard deviation as $std[p_j] \approx \sqrt{p_j(1 - p_j)/n_j}$. For all three syntheses (59A-10A, 59B-10B and 59A-20), the measured hybridization probability and 95% confidence intervals (2 × the standard deviation) from separate experiments showed consistent trends (Fig. S2).

We also examined the differences in the fitted parameters that model the data for each experiment. Table S1 reports the model parameters determined from each experiment, showing that the observed experiment-to-experiment variability was relatively small. For each experimental condition, combined data sets from both experiments were reported and modeled in the main text (Fig. 2; Table 1, *i*-*iii*).

References

[1] Imagej - a public domain java image processing program. Available at http://rsb.info.nih.gov/ij, developed by Wayne Rasband, National Institutes of Health, Bethesda, MD.

Table S1: **Comparison of fitted hybridization model parameters for repeated experiments.**

Target-ODN[*] **Synthesis**	**Exprm.** **No.**[†]	**Total No.** **of events**	r **± s.d.**[‡] ($\mathbf{s^{-1}}$)	$k_{\text{on}}^{\P}$ **± s.d.** $\times 10^6$ ($\mathbf{M^{-1}s^{-1}}$)	q **± s.d.**[‡]
59A-10A	1	4557	117.7 ± 3.8	21.4 ± 0.7	0.97 ± 0.004
–	2	2541	105.7 ± 5.1	19.2 ± 0.9	0.97 ± 0.004
59A-20	1	2040	45.9 ± 2.5	8.3 ± 0.5	0.99 ± 0.01
–	2	2962	54.0 ± 2.2	9.8 ± 0.4	0.99 ± 0.01
59B-10B	1	2929	127.9 ± 8.6	23.3 ± 1.6	0.79 ± 0.01
–	2	2647	112.1 ± 10.2	20.4 ± 1.9	0.79 ± 0.01

[*] Oligodeoxynucleotide = ODN. Each experiment had 59 mer target DNA in the *cis* chamber, and in separate experiments 10 mer or 20 mer complimentary ODNs in the *trans* chamber at $[O] = 5.5$ μM. Two syntheses of 59 mer target DNA (59A,B) and 10 mer binding ODNs (10A,B) were used separately in each experiment listed. A single synthesis of 20 mer ODN was used.

[†] Combined data sets were modeled and results reported in the main text (Fig. 2; Table 1, *i*-*iii*).

[‡] Method of estimating $\{q, r\}$ and their uncertainty is detailed in Section 2. All parameters are reported in the form of (estimated value) ± (standard deviation)

[¶] Rate constant and standard deviation computed using $k_{\text{on}} = r/[O]$, given the value and standard deviation for r.

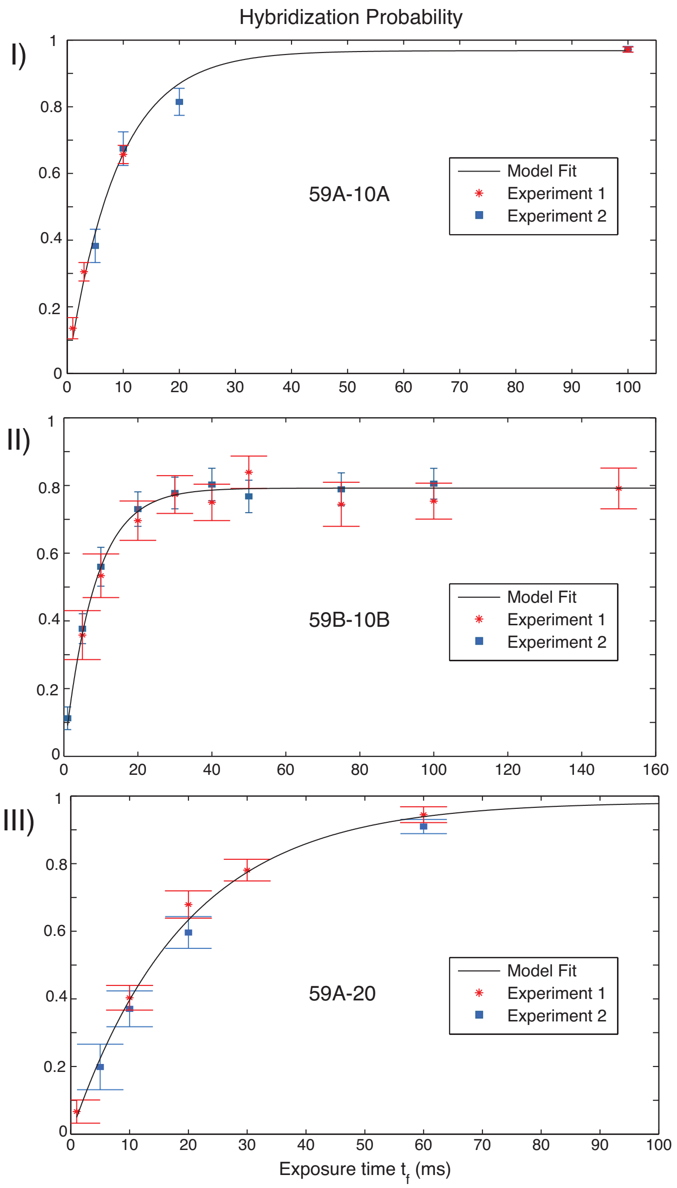

Figure S2: **Measured hybridization probability are consistent in repeated experiments for (I) 59A-10A, (II) 59B-10B, and (III) 59A-20 syntheses.** For each experiment, the measured probability data points p_j are shown with error bars representing 95% confidence intervals $(p_j - 2\sqrt{p_j(1-p_j)/n_j}, \quad p_j + 2\sqrt{p_j(1-p_j)/n_j})$, where $p_j = X_j/n_j$ and X_j is the number of measured hybridization events out of n_j trials at exposure times shown. The model curves shown are fitted to the combined data sets (*black*) and have parameter values reported in the main text (Table 1, *i*-*iii*). Fitted model parameters for each data set are reported in Table S1. Longer exposure times (up to 500 ms) measured in (II-III) were consistent with fitted q parameters, but are not displayed to show a comparative close-up of the transient and steady-state trends. Experimental conditions are reported in the main text.