bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mechanisms underlying sequence-dependent DNA hybridisation rates in the absence of secondary structure

Sophie Hertel[1,^], Richard E. Spinney[1,2^], Stephanie Y. Xu[1], Thomas E. Ouldridge[3], Richard G. Morris[1,2], and Lawrence K. Lee[1,4*]

[1] EMBL Australia Node for Single Molecule Science, School of Medical Sciences, UNSW Sydney, 2052, Australia
[2] School of Physics, University of New South Wales - Sydney 2052, Australia
[3] Department of Bioengineering and Centre for Synthetic Biology, Imperial College London, London, SW7 2AZ, United Kingdom
[4] ARC Centre of Excellence in Synthetic Biology, University of New South Wales, Sydney, Australia

[^] These authors contributed equally

* To whom correspondence should be addressed. Email: lawrence.lee@unsw.edu.au

ABSTRACT

The kinetics of DNA hybridisation are fundamental to biological processes and DNA-based technologies. However, the precise physical mechanisms that determine why different DNA sequences hybridise at different rates are not well understood. Secondary structure is one predictable factor that influences hybridisation rates but is not sufficient on its own to fully explain the observed sequence-dependent variance. Consequently, to achieve a good correlation with experimental data, current prediction algorithms require many parameters that provide little mechanistic insight into DNA hybridisation. In this context, we measured hybridisation rates of 43 different DNA sequences that are not predicted to form secondary structure and present a parsimonious physically justified model to quantify their hybridisation rates. Accounting only for the combinatorics of complementary nucleating interactions and their sequence-dependent stability, the model achieves good correlation with experiment with only two free parameters, thus providing new insight into the physical factors underpinning DNA hybridisation rates.

INTRODUCTION

DNA is a biopolymer formed from four different nucleotides, adenine, thymine, guanine and cytosine (A,T,G and C respectively), whose order or sequence is used to encode information that is the foundation of biology. Complementary DNA strands hybridise via Watson and Crick base pairing between A-T or G-C bases to form the DNA double helix or duplex (1), whose structural (2) and physical (3-5) properties are well characterised. In addition to its essential role in biology, DNA hybridisation also underpins DNA nanotechnology (6-8), which utilises DNA self-assembly for the construction of rationally designed nanoscale structures and machines (9-16). DNA nanotechnology has led to the development of a broad range of technologies including applications in molecular sensing (17-20), coordinating complex reaction cascades (21-23), drug delivery vessels (24-27) and super resolution imaging methods, such as DNA points accumulation for imaging in nanoscale

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

topography (DNA-PAINT) (28). Thus, understanding the thermodynamics, kinetics and mechanisms for DNA hybridisation is fundamentally important for biology and biotechnology.

The thermodynamics of DNA hybridisation have long been observable via spectrophotometric or viscometric observations of thermal melt curves and are well studied (29-31). The reaction is dominated by states consisting of completely dissociated DNA strands or fully hybridised DNA duplexes. The stability of a DNA duplex can therefore be estimated from the structure of a fully hybridised duplex and is dependent on hydrogen bonds between paired bases in DNA duplexes and hydrophobic base stacking that occurs between neighbouring base pairs; both these interactions are sequence dependent (32, 33). Models predicting the melting temperature T_m of a DNA duplex adopt a two-state nearest neighbour approach, which postulates that the stability of a given base pair depends on the identity of the nucleotide bases involved (A-T or G-C) and its nearest neighbour base pairs. In turn, the T_m associated with the formation of any DNA duplex can be estimated from the sum of the free energy of all 2 contiguous base-pairing interactions, as well as additional parameters to account for the relative stabilities of the ends of the duplex (32, 33). Given that there are only 10 unique combinations of 2-base sequences, nearest neighbour models can be parameterised experimentally including in different buffer conditions (34-41), and algorithmic implementations predict hybridisation T_m reasonably well (42, 43).

The *kinetics* of DNA hybridisation are also sequence-dependent (44-47). However, the pathways to DNA hybridisation are difficult to observe, and therefore the physical basis for sequence-dependent hybridisation rates remains poorly understood. Based on early thermodynamic and kinetic measurements of DNA hybridisation, Pörschke and colleagues proposed a reaction mechanism in which hybridisation proceeds via a slow, rate limiting bimolecular nucleation step, followed by fast monomolecular 'zippering' into a fully formed DNA duplex (30, 48, 49). More recent developments in coarse-grained molecular dynamics (MD) simulations enabled an *in silico* view of hybridisation pathways, in which the rate-limiting nucleation step consisted of a short stretch (~3 bp at 300 K) of contiguous and complementary base-pairing interactions (50-52). Since there are typically many such possible nucleating interactions, it therefore follows that the combination and relative stability of all possible nucleating interactions, which is entirely determined by the DNA sequence, defines the overall activation free energy and therefore the rate of any DNA hybridisation reaction. However, DNA strands can also form intramolecular interactions that result in secondary structures such as hairpins that influence both the rates of hybridisation and melting. Such secondary structure can reduce hybridisation rates either by limiting the availability of a subset of nucleating interactions or by lowering the probability that any given nucleating interaction is stable enough to favour the displacement of the secondary structure, which must be denatured prior to zippering into a fully formed duplex (53-55).

Two algorithms have recently been developed for predicting sequence dependent hybridisation rates. A 'weighted neighbour voting' algorithm was used to examine 50 different sequence-dependent physical 'features' and found 35 different features that correlated with hybridisation rates (56).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Perhaps unsurprisingly, of these features, the ensemble standard free energy of secondary structure emerged as the single best predictor of DNA hybridisation rates, reporting predictions of hybridisation rate constants (k_a) of ~60% accuracy within a factor of two. The inclusion of five additional features resulted in a six-parameter model, which achieved a reported ~80% prediction accuracy within a factor of two. However, apart from secondary structure, the physical mechanisms underpinning how these additional features influence hybridisation rates remain unclear. Hata *et al.* subsequently presented an alternative, physically motivated, model by estimating the relative binding capability for all 3 consecutive base sequences involved in all possible nucleation interactions, including those which were off-register or mis-matched (45). This capability was dependent on an estimate of the propensity of any of the 32 possible 3-base nucleating interactions to seed full hybridisation and on the probability of predicted secondary structures sterically hindering nucleating interactions. Surprisingly however, seeding propensities did not correlate with the stability of nucleating interactions and accurate predictions required these propensities to be determined empirically by fitting 32 free parameters to experimental data. Thus, secondary structure remains the only physically well-defined determinant for algorithms predicting sequence dependent hybridisation rates. However, accurate predictions require multiple additional parameters that are not physically well defined. This suggests that there are other dominating physical factors apart from secondary structure that are yet to be identified.

Coarse grained MD simulations, for example, indicate that nucleation can occur from base pairing interactions that are off-register from a fully formed duplex (50-52). In these instances, off-register nucleation states can progress to metastable intermediaries such as misaligned duplexes that can move into register via inchworming or pseudoknot internal displacement mechanisms, followed by the final zippering step into the fully hybridized DNA double strand (50). Like zippering, these monomolecular rearrangements also occur much more rapidly than nucleation. Consequently, repetitive sequences, which have a greater number of possible off-register nucleating interactions were predicted to hybridise more rapidly than non-repetitive sequences (50).

Here we explored the impact of off-register nucleating states on the hybridisation rate of DNA strands experimentally. To reduce the complexity of hybridisation pathways and to identify physical elements yet to be explicitly accounted for in predictive models, we focused on sequences that were not predicted to form secondary structures. Using surface plasmon resonance (SPR) we measured the hybridisation rates of 43 different DNA strands with varying GC content and degree of sequence repetition and demonstrate that repetitive sequences do indeed hybridise more rapidly than non-repetitive sequences. We also present a simple, physically-justified model, which demonstrates that it is possible to capture much of the variance in sequence dependent hybridisation rates with only two free parameters that account for the combination and stability of all possible nucleating interactions, including those that are off-register.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MATERIALS AND METHODS

DNA oligonucleotides

All DNA was purchased from IDT. The salt purified oligonucleotides were resuspended in milliQwater and stored at -20 °C. To ensure that the measured hybridization kinetics only depended on differences in the sequence, DNA strands were designed to have no or negligible secondary structures (2bp or less) and nearly the same free energy of the lowest energy double stranded complex, using NUPACK and IDT (Table 1).

Surface plasmon resonance experiments

SPR experiments were performed with a Biacore S200 system. A CM5 chip was coated with 4000-5000 RU streptavidin purchased from Sigma-Aldrich. The experimental setup for the surface plasmon resonance measurements was chosen as described before (57), shown in figure 1A. To ensure a Langmuir 1:1 interaction model, an anchor DNA strand was immobilized on the chip surface by biotin-streptavidin coupling at a density of $1.7 \times 10^9\ molecules/mm^2$ so that intermolecular crosslinking of the immobilized DNA strands was minimized. The anchor strand then captured the template strand, which had a free complimentary binding site for the target strand. By referring to the mass and length of the anchor and template strands, the highest signal expected for binding of the target strand to the template was 11 RU (10bp) and 12 RU (14bp), respectively.

The biotinylated anchor strand was immobilized on two flow cells of the sensor chip, leaving two flow cells as blank reference cells. DNA samples were prepared in 10 mM HEPES pH 7.5, 150 mM NaCl, 3 mM EDTA and 0.005 % Tween20 running buffer and SPR experiments were performed in the same buffer at 25 °C and a flow rate of 60 µl/min. The SPR chip could be regenerated for reuse by removing the template strand with a 60 sec injection of 10 mM glycine pH 2.5. Sensorgrams were double-referenced and three repeats of each data set were carried out. The corrected binding curves were fitted with a 1:1 binding model to obtain apparent association constants, k_{app}, and dissociation constants, k_d. k_{app} were plotted as a function of the target concentration and fit to a linear function whose slope corresponded to the association rate constant, k_a. K_D from steady state measurements was calculated using the RU_{max} values obtained from binding curve fits, plotted as a function of the target concentration. All data was fit using Prism and MATLAB. Final k_a and k_d values are averages of at least three replicates and errors reported are standard deviations.

Estimation of binding free energies with NuPACK

Binding free energies were determined using NUPACK version 4.0.0.21 with the following parameters: $25\ ℃$, 0.15 M NaCl, material setting to ‘dna2004’ and ensemble parameter to ‘stacking’.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RESULTS

Repetitive DNA sequences hybridise more rapidly than non-repetitive sequences

To experimentally test predictions that additional off-register nucleating interactions in repetitive sequences result in faster hybridisation rates, we compared association rates of two 14 base sequences previously analysed in coarse grained MD simulations (50). The first was a non-repetitive sequence (14NR) with 50% GC content that was designed to minimise hairpin formation and off-register interactions with its complementary strand. The second consisted of seven successive AC repeats (14AC), which maintains the same GC content as 14NR but allows for more off-register nucleating interactions. The absence of complementary bases precludes the formation of secondary structure *via* intramolecular base pairing. In addition, we measured hybridisation kinetics of a repeated sequence consisting of seven successive AG repeats (14AG). Unlike the 14AC sequence, the AG repeat sequence has the capacity to form G-quadruplexes including GAGA quartets (58), and GAGAGAGA heptads (59). Thus, the 14AG sequence provided a convenient means to assess the relative impact of secondary structures and the additional off-register nucleation sites in repeated sequences on DNA hybridisation rates. We also performed measurements with variants of the 14-base DNA sequences that were truncated to a length of 10 bases. Details of all DNA sequences used in this study are summarised in Table 1.

Table 1. All DNA sequences with associated rate constants measured in this study.

Number	Name	Sequence 5' > 3'	k_a ($M^{-1}s^{-1}$) x 10^6	k_d (s^{-1}) x 10^{-2}
	Anchor	B-TTTGACCTCCTTGGCAGCACTG		
	Template	*X_nTTTCAGTGCTGCCAAGGAGGTC		
1	14NR	GCTGTTCGGTCTAT	1.04 ± 0.12	NA
2	14AC	CACACACACACACA	4.82 ± 0.55	NA
3	14AG	TCTCTCTCTCTCTC	2.02 ± 0.07	NA
4	10NR	GTTCGGTCTA	1.15 ± 0.09	0.89 ± 0.05
5	10AC	ACACACACAC	3.84 ± 0.15	0.67 ± 0.09
6	10AG	TCTCTCTCTC	NA	NA
		50 % GC content		
7		ACCAACCAACCAAC	5.21 ± 0.08	NA
8		CAACAACACCACCA	3.24 ± 0.40	NA
9		AAACCACCCAACAC	2.75 ± 0.13	NA
10		CCACCAACAACAAC	4.06 ± 0.18	NA
11		CAACACCCAAACAC	2.12 ± 0.26	NA
12		ACCAAACCACCAAC	1.19 ± 0.16	NA
13		CAAAACCCCAACAC	1.83 ± 0.08	NA
14		ACCAACACCAACCA	3.19 ± 0.14	NA
15		AACCACCACAAACC	3.66 ± 0.43	NA
16		ACACACACCACACA	4.44 ± 0.33	NA
17		CAACACAACCAACC	3.76 ± 0.40	NA
18		AAACCCACCACACA	1.89 ± 0.40	NA
19		AACCAACACCACCA	3.36 ± 0.33	NA
20		CAACCAACCA	3.96 ± 0.40	0.50 ± 0.04

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

21	ACAACACCAC	2.56 ± 0.18	0.24 ± 0.01
22	ACACCAAACC	2.18 ± 0.31	0.78 ± 0.15
23	CCACCAACAA	2.83 ± 0.33	1.27 ± 0.16
24	CAACACCCAA	2.51 ± 0.34	1.75 ± 0.21
25	ACCAAACCAC	2.17 ± 0.25	0.32 ± 0.01
26	CAAAACCCCA	2.65 ± 0.23	1.79 ± 0.19
27	ACCAACACCA	2.77 ± 0.44	0.87 ± 0.05
28	AACCACCACA	3.87 ± 0.37	0.75 ± 0.05
29	ACACACACCA	4.16 ± 0.37	0.66 ± 0.06
30	CAACACAACC	2.70 ± 0.39	0.40 ± 0.02
31	AAACCCACCA	2.53 ± 0.45	1.64 ± 0.09
32	AACCAACACC	2.66 ± 0.32	0.43 ± 0.04
	57 % GC content		
33	CCCAAACCCAACCA	2.98 ± 0.24	NA
34	CACCACAACCACCA	3.81 ± 0.39	NA
35	CCCCACACAACAAC	3.96 ± 0.67	NA
36	ACACCACCAC	5.61 ± 0.47	0.11 ± 0.01
37	CCCCACACAA	5.45 ± 0.08	0.74 ± 0.10
	42 % GC content		
38	CCAAAACCAACAAC	2.30 ± 0.09	NA
39	AAAAACCCACCCAA	2.43 ± 0.37	NA
40	CAACACCAAACAAC	1.78 ± 0.30	NA
41	CCAAAACCAA	1.67 ± 0.17	4.10 ± 1.04
42	AAAAACCCAC	2.36 ± 0.24	2.08 ± 0.10
43	AAACCACACA	1.81 ± 0.30	3.34 ± 0.31

*X_n represents a DNA stand complementary to the target sequence of length n = 10/14

DNA hybridisation kinetics were measured with SPR as previously described (57). DNA strands that were complementary to target strands were immobilised to the surface of an avidin-coated SPR chip by hybridisation to biotinylated 'anchor' strands (Figure 1A). During association measurements, target strands were flowed over the surface of the chip at fixed concentrations [T] resulting in pseudo-first-order binding kinetics. The response units (RU) from all SPR sensorgrams were therefore well described by the following monoexponential equation:

$$RU_t = -RU_{max}e^{k_{obs}t} + R_0, \quad (1)$$

where RU_{max} is the RU value when all binding sites are occupied and R_0 is the RU value at the zero time point (Figure 1B and C and S1-3). Association rate constants (k_a) could then be calculated from sensorgrams according to:

$$k_a = \frac{k_{obs} - k_{off}}{[T]} \quad (2)$$

The repetitive 14AC sequence ($k_a = 4.8 \pm 0.5 \times 10^6 M^{-1}s^{-1}$) hybridised approximately five times faster than the non-repetitive 14NR sequence ($k_a = 1.0 \pm 0.1 \times 10^6 M^{-1}s^{-1}$) (Figure 1D). This is consistent with MD simulations, suggesting that the additional possible off-register nucleating interactions in repetitive sequences result in faster hybridisation rates (50). As expected, given the

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

propensity for the 14AG sequence to form secondary structures (58, 59), the 14AG sequence hybridised more slowly than the 14AC sequence. Interestingly however, the association rate for the 14AG sequence ($k_a = 2.0 \pm 0.07 \times 10^6 M^{-1}s^{-1}$) was faster than that of the non-repetitive 14NR sequence, suggesting that additional off-register nucleation states in the AG sequences are sufficient to off-set the reduction of the hybridisation rate associated with the presence of secondary structures. Truncating DNA strands to 10 bases did not appear to have a large effect on association rates. There was no detectable difference in association rates between the 14NR and the truncated 10NR sequence ($k_a = 1.2 \pm 0.1 \times 10^6 M^{-1}s^{-1}$) and while slower than the 14AC sequence, the 10AC sequence ($k_a = 3.8 \pm 0.2 \times 10^6 M^{-1}s^{-1}$) still hybridised 4 times faster than the 10NR sequence. Dissociation rates were drastically faster for the AG sequence compared with the AC and NR sequences. The 10AG sequence dissociated too rapidly to be captured within the limits of experimental measurements. Since observed binding curves also depend on dissociation rates (see equation 2), neither association nor dissociation sensorgrams could be fit to monoexponential equations to obtain accurate association or dissociation rates for the 10AG sequence. In contrast, dissociation rates for the 10NR and 10AC sequences were similar and much slower than the 10AG sequence, with a mean dissociation rate of $8.9 \pm 0.5 \times 10^{-3}s^{-1}$ and $6.7 \pm 0.9 \times 10^{-3}s^{-1}$, respectively (Figure 1E). This is consistent with predictions that secondary structures in DNA sequences not only decrease association rates but have a pronounced tendency to increase dissociation rates, possibly originating from the formation of secondary structures during melting (55). It also follows that the relatively slow dissociation rates of the AC and NR sequences reflects the lack of significant secondary structures in these sequences as predicted.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

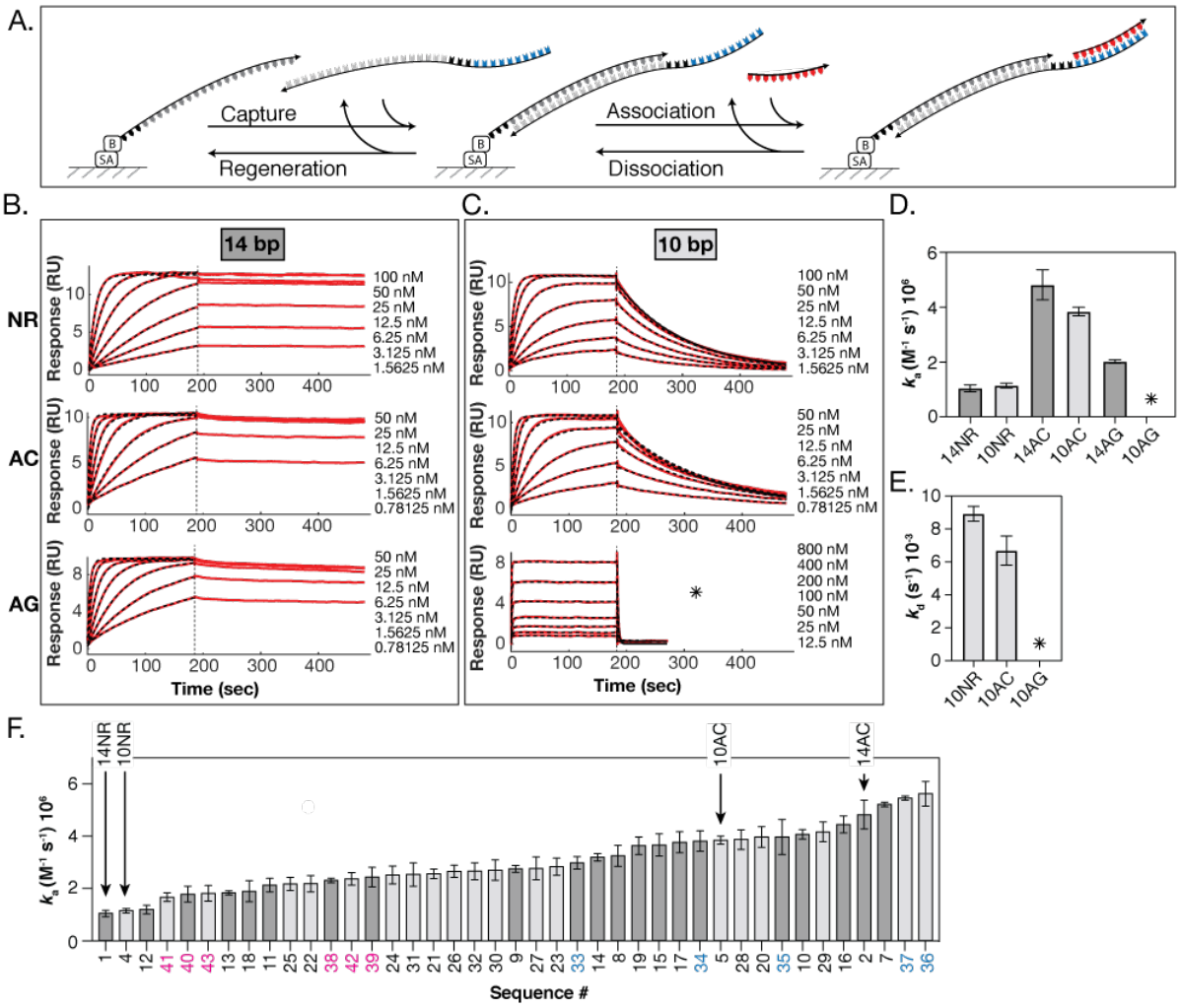

Figure 1: Binding kinetics for non-repetitive and repetitive DNA sequences. (A). Schematic depiction of the surface chemistry used to measure DNA hybridisation kinetics with SPR. First, a 20-base biotinylated DNA strand (anchor, dark grey) binds irreversibly to the streptavidin coated surface of the SPR chip. Second, a longer strand of DNA (template, light grey) that is complementary to the anchor binds (capture). The template strand has an extension consisting of a 3-thymine spacer and a sequence that is complementary to the target strand. Third, association and dissociation kinetics (association and dissociation respectively) of the target strand (red) can then be measured in real time. The chip can be re-used for replicate experiments after a regeneration step that denatures all DNA duplexes leaving only the black anchor strand (regeneration). (B and C) Representative raw SPR sensorgrams (red) with mono-exponential fit (dashed black) to association phase for 14bp sequences (B) and to association and dissociation phase for 10bp sequences, fit locally for each concentration (C). The apparent high association rate of the 10AG sequence was due to the use of high concentration of target strand required to get an appreciable yield, and the fast dissociation, which increases the rate at which the system approaches equilibrium. Replicate data for sequences in (B) and (C) are in figure S1. (D) Association rate constants for 14bp (dark grey) and 10bp (light grey) sequences. (E) Dissociation rate constants for 10bp sequences. * indicates that no kinetic rates could be determined. (F) Association rate constants for all DNA sequences without secondary structure in

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

this study as measured by SPR and indexed according to Table 1. As in (D), light and dark grey correspond to 10 and 14 base sequences respectively. Sequences with 42% and 57% GC content are marked with magenta and cyan labels respectively. All other sequences have 50% GC content. Error bars are standard deviation from at least three independent measurements. Raw SPR sensorgrams fitted with monoexponential equations are in figure S2-3.

Association rates of randomly generated AC sequences

DNA sequences consisting only of AC bases (AC sequences) provide a useful means to explore the mechanisms underlying sequence-dependent hybridisation rates in the absence of secondary structure. We therefore measured the hybridisation rates of an additional 38 randomly generated DNA sequences consisting of only adenine and cytosine bases. These DNA strands were either 10 or 14 bases in length with a GC content between 40% and 60%. As above, kinetic traces of all sequences were consistent with pseudo-first-order binding kinetics (Figure S2-3) allowing for reliable determination of binding rates, which are presented in order of increasing rates in figure 1F. The kinetic rate constants for all sequences in this study are summarised in table 1 and supplementary table 1, which also shows, where applicable, consistent equilibrium dissociation constants (K_D) as calculated from kinetic rates and steady state measurements, further confirming the reliability of SPR data.

The repetitive 14AC and 10AC sequences were among those with the fastest hybridisation rates ranking 4th and 11th respectively, whereas the non-repetitive 14NR and 10NR had the slowest hybridisation rates (Figure 1F). Furthermore, consistent with previous reports (46), sequences with a higher GC content tended to hybridise more rapidly. Thus, increased hybridisation rates appear to broadly correlate with a greater number and stability of possible nucleating interactions. The dependence of DNA hybridisation rates on strand length may also provide important mechanistic insight. Previous studies are in agreement that dissociation rates significantly decrease with increased DNA strand length. However, data on association rates are mixed with reports of weakly increasing rates (60), decreasing rates (61, 62) or an absence of an effect on the rates (63). Our data also shows no obvious correlation between DNA hybridization rates and sequence length. This suggests either that a length dependent effect on hybridisation rates is insignificant between lengths of 10 and 14 bp or that length related hybridisation properties off-set each other to result in the apparent lack of correlation.

Simple physically motivated model for capturing the variance in DNA hybridisation rates

To explore the underlying mechanisms dominating DNA hybridisation kinetics in more detail, we constructed a simple, physically motivated model to quantify the correlation between hybridisation rates and the number and stability of nucleation states, including those that are off-register, that result in a fully formed duplex. This simple model is predicated on the idea that in the absence of other

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

factors such as secondary structure, sequence-dependent hybridisation rates are based fundamentally on two factors, the combinatorics of available nucleation sites, and their stability. As illustrated in figure 2, the model assumes that hybridisation proceeds via a nucleation state consisting of a small sub-sequence of n contiguous intermolecular base pairing interactions (30, 48-50, 52). n thus constitutes a model parameter controlling the effect of combinatorics of the sub-sequences in the strands. From this nucleated state, the strands either dissociate and return to solution or transition into one of a vast number of complicated states associated with various intermediary and meta-stable complexes including partially zippered, off-register structures from where the complex can proceed to a fully hybridised duplex (50).

If transitions from an unbound state into a nucleated state are rate-limiting, and progression to eventual full hybridisation from a meta-stable structure is very likely (50), then a faithful description of the kinetics between the unbound, nucleation and meta-stable states will provide an approximate measure of the total observed rate of hybridisation. We can thus construct a framework for modelling the effective hybridisation rate constant using the simple form:

$$k_a = \sum_{i=1}^{L-n+1} \sum_{j=1}^{L-n+1} k_{i,j}. \quad (3)$$

Here, L is the length of the strand expressed as an integer number of bases, whilst i and j are indices corresponding to the position of the first of n contiguous bases which make up the nucleation state, in the 5' → 3' direction, for the strand and its complement, respectively (Figure 2B – top). The double sum therefore includes $(L-n+1)^2$ contributions from all such nucleation states, regardless of whether they are on-register, or whether they are formed from complementary bases (Figure 2B - bottom). $k_{i,j}$ then quantifies the specific contribution arising from the nucleation state at positions i and j on the strand and its complement, respectively. Crucially, unlike previous models (45), this allows the contribution of any particular nucleation state to vanish in the case of mis-matched bases, thus naturally capturing the combinatorics of nucleating interactions, and for the contribution of nucleating interactions to vary with the stability of the nucleated state when they do match.

To account for the relative stability of each i,j nucleation site we can approximate the associated contributing rate constant $k_{i,j}$ as arising from an idealised sub-system consisting of free or dissociated strands in solution, a single *i,j* nucleation state, and a 'bound' state representing all configurations where the strands are in one of many more complicated subsequent complexes including fully hybridised duplexes or pseudoknots etc. (Figure 2C). Thus, for any given individual i,j nucleation site, this three state sub-system is then fully described with the specification of the rate constants associated with transitions between these states (Figure 2). Various assumptions can then be implemented to define the rates at each step.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

First, we specify a transition rate from a nucleated state to dissociated strands in solution, which we capture as a stability term measured through the free energy of binding of the nucleation state, and thus introducing explicit sequence dependence into the model,

$$r^{i,j}_{nucl\to sol} = r'_{nucl\to sol} e^{\Delta G^0_{i,j}/\mathrm{RT}}. \quad (4)$$

Here $\Delta G^0_{i,j}$ is the free energy of binding of the nucleation state associated with binding locations i and j measured in J/mol, $r'_{nucl\to sol}$ is a rate constant that is independent of the binding sequence, R is the gas constant and T is the temperature in Kelvin. Second, we ignore any entropic effects of unbound DNA bases surrounding the nucleation site and assume that all specific nucleation sites are equally accessible from dissociated strands. The rate for forming any given i,j nucleating interaction is assumed to be constant across all nucleation sites, and can be given by $r^{i,j}_{sol\to nucl} = r_\kappa (L-n+1)^{-2}$ where r_κ is an overall scaling factor representing the rate of a nucleation event occurring in any pair of locations on the strands independently of their length, comprising a rate constant κ and any concentration dependence (e.g. $\mathrm{r}_\kappa = \kappa[T]$ in the pseudo-first-order conditions above), and the $(L-n+1)^{-2}$ term imposes the observed lack of scaling of hybridisation rates with strand length on the model. Third, we assume that the rate of transitions from the bound state to the nucleated state is slow relative to the time scales of hybridisation and hence these transitions are ignored in the model. Finally, as a first approximation, we assume that the microscopic rate of transition from any nucleated site into an intermediary or metastable state is constant across all nucleation sites and sequences ($r^{i,j}_{nucl\to bound} = r_{nucl\to bound}$).

The effective rate for hybridisation via any given i,j nucleating interaction can be arrived at by computing the inverse of the mean first passage time taken to transition from state 1 to state 3. From the rates defined above this rate is given by (Supplementary Note 1):

$$r_{i,j} = \frac{\mathrm{r}_\kappa (L-n+1)^{-2} r_{nucl\to bound}}{r'_{nucl\to sol} e^{\Delta G^0_{i,j}/RT} + \mathrm{r}_\kappa (L-n+1)^{-2} + r_{nucl\to bound}}. \quad (5)$$

While nucleation is the rate limiting step, the rate of transitions away from the nucleated state are much faster than the rate of transitions into the nucleated state ($r_{nucl\to bound}, r'_{nucl\to sol} \gg \mathrm{r}_\kappa (L-n+1)^{-2}$). As such we can simplify this expression, to leading order in $\mathrm{r}_\kappa (L-n+1)^{-2}/k_{nucl\to bound}$, and then convert to the relevant rate constant to find

$$k_{i,j} \approx \frac{\kappa (L-n+1)^{-2}}{\frac{r'_{nucl\to sol}}{r_{nucl\to bound}} e^{\Delta G^0_{i,j}/RT} + 1} = \frac{\kappa (L-n+1)^{-2}}{e^{\gamma + \Delta G^0_{i,j}/RT} + 1} \quad (6)$$

where $\gamma = \ln (r'_{nucl\to sol}/r_{nucl\to bound})$. The rate constant for hybridisation via any single (i,j) nucleation state can thus be directly interpreted as a uniform and limiting rate, $\kappa (L-n+1)^{-2}$, into the nucleation state from solution multiplied by a probability of continuing through to full hybridisation from the nucleation state,

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$p_{i,j}^{hybridise} = \frac{1}{e^{\gamma+\Delta G_{i,j}^{0}/RT}+1}. \quad (7)$$

Consequently, the value of γ coincides with the value of $-\Delta G_{i,j}^{0}/RT$ for which the probability of continuing on to hybridisation is $p_{i,j}^{hybridise} = 1/2$. Substituting equation (6) into equation (3) we arrive at

$$k_a = \kappa(L-n+1)^{-2}\sum_{i=1}^{L-n+1}\sum_{j=1}^{L-n-1}\frac{1}{1+e^{\gamma+\Delta G_{i,j}^{0}/RT}} \quad (8)$$

fully specifying our model up to estimation of the nucleation free energies of the nucleation state. When a nucleation state (i,j) constitutes a mismatch the model considers the nucleation free energy to be infinity such that the probability of hybridisation is zero. For complementary nucleation states, nucleation free energies, $\Delta G_{i,j}^{0}$ were obtained using the NUPACK 4.0.0.21 implementation of the nearest neighbour model (see materials and methods).

Given a fixed n, the terms κ and γ then constitute the free parameters of the model, which can be fit to data. However, of the two, only γ controls the sequence dependence, with κ simply acting as a scaling factor. In physical terms γ controls how sharply the increases in the stability of the nucleation states increases the likelihood of continuing through to full hybridisation. Crucially, the fact that the sequence dependent stability of nucleating interactions is controlled by a single free parameter dramatically restricts model complexity such that over-fitting can be avoided as much as possible.

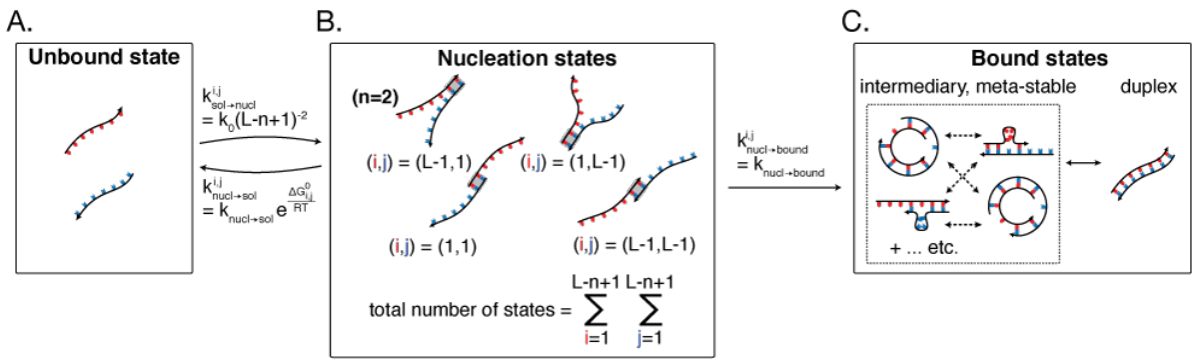

Figure 2. Cartoon depiction of nucleation and hybridisation underpinning the simple model with a nucleation length of two. From the unbound state (A) the system can transition into one of $(L-n+1)^2$ possible nucleation states, where L is the length of the DNA strands and n is the length of the nucleating interaction. (B) illustrates examples of the many possible nucleation states of length $n = 2$. From any of these states the system can either return to an unbound state, or it can continue through to full hybridisation via a complicated network of possible intermediary and meta-stable bound states as illustrated in (C). The model assumes negligible transitions from bound states back to nucleated states. Rates are for transitions to and from a single specified location (i,j), where i and j refer to the index of the first base involved in the nucleating interaction from the 5' end, on each strand respectively. For each possible nucleation location there exists a constant rate of nucleation

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

equal to $r_\kappa(L-n+1)^{-2}$ such that the overall rate of nucleation is independent of length. Then there is a sequence dependent rate from the nucleated state back to solution that depends on the stability of the nucleated binding complex. Finally, there is a constant rate of transitioning from the nucleated state into a bound state. From these rate definitions, the effective rate of hybridisation due to nucleation location (i,j) is taken as the inverse of the mean first passage time from the solution state to the bound state (equation (6), Supplementary Note 1).

Fits to experimental data

We first determine how well nucleation site combinatorics alone correlate with relative hybridisation rates, without accounting for stability. This is achieved simply with the model described above (Equation 8) by replacing the probability of hybridisation with a value of one if the nucleation site is formed from complementary base pairs, or zero with mismatched base pairs. All other aspects of the model are unchanged. The resulting model rates can then be scaled to fit with experimental data by varying the scale factor, κ, using a Nelder-Mead optimisation algorithm taking the sum of the squared residuals as the objective function to be minimised. Fits were performed with nucleation lengths of $n = 1,2,3\ and\ 4$ and predicted rates plotted against measured rates (Figure 3A) from which correlation coefficients were calculated. Given the complexity of DNA hybridisation, and the potential that many rate determining factors were not accounted for in our simple model, it was important to ensure that reported correlations were a true reflection of model accuracy. We therefore performed careful statistical analysis to estimate standard deviations and confidence intervals to quantify the certainty in correlation coefficients. In addition, we performed permutation tests to determine p-values based on the probability that similar correlations could be obtained from null-distributions of randomly reshuffled datasets. Where correlations between model and experiment were high ($\rho > 0.5$) p-values were low ($p < 0.001$), thus providing confidence that the observed trends were not spurious. Further, null distributed correlations were very close to 0 (Supplementary Table 2) providing confidence that the observed correlations were not due to overfitting. A detailed description of error analyses performed in this study is in supplementary note 2. Standard deviations, confidence intervals and p-values for all reported correlation coefficients are in supplementary table 2, and model parameters from all fits in this study are in supplementary table 3. For completeness, we also report correlation coefficients calculated from point values in supplementary table 4, which do not account for experimental uncertainty.

With a nucleation length of $n = 1$, combinatorics alone provides no distinguishing power between different AC sequences with the same GC content, since these sequences necessarily have the same number of possible complementary one base interactions. Consequently, predicted rates using combinatorics alone were essentially flat for AC sequences when $n = 1$ (Figure 3A). Additionally, the NR sequences, which also consist of G and T bases, can make far fewer complementary single base pair interactions and thus have lower predicted hybridisation rates, consistent with experiment. Indeed, any weakly existing correlation when $n = 1$ can be attributed to the slower hybridisation rates for the NR sequences. With increasing nucleation length, there is a larger variation in the number of

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

complementary nucleating interactions. This variation yields a clearly positive correlation between predicted and measured hybridisation rates (Figure 3A) that increases with nucleation length reaching a maximum when $n = 3$, which has a correlation coefficient of $\rho = 0.56 \pm 0.04$.

To incorporate base-pair stability in the model, we next fit the full model in equation (8) to experimental data to determine whether improved fits could be obtained over predictions based on nucleation site combinatorics only. Accounting for stability introduces the single free parameter, γ, which along with the scaling parameter κ, was varied, again taking sum of the squared residuals as the objective function to be minimised for each nucleation length ($n = 1,2,3\ or\ 4$). Relative to combinatorics alone, the full model resulted in an improved correlation between model and experimentally measured hybridisation rates across all nucleation lengths (Figure 3B). We note in particular a high correlation with a nucleation length of $n = 1$ ($\rho = 0.68 \pm 0.03$), where there was a lack of correlation from combinatorics alone and which therefore can be attributed almost exclusively to the inclusion of nucleation site stability.

For higher binding lengths, $n > 1$, stability reliably improves the achieved correlation, but none attain the correlation captured by the n=1 case. We observe however that the model consistently over-estimates the related, and most repetitive sequences, 10AC and 14AC, possibly indicative of a lurking feature limiting the increase of hybridisation rates due to increasing combinatorics not captured by the model. If we omit these two related sequences the maximal correlation between predicted and measured hybridisation rates is improved to $\rho = 0.73 \pm 0.03$, occurring at $n = 3$ (Figure 3B and S4 and Supplementary Table 2).

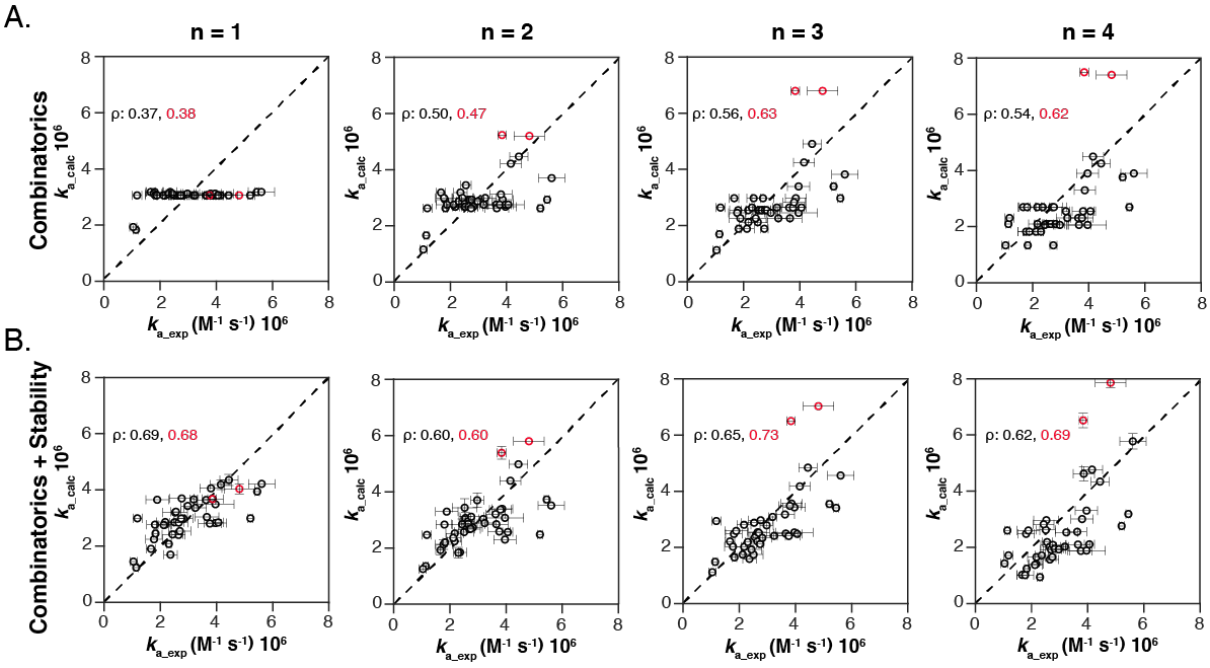

Figure 3. Predicted vs measured hybridisation rates. Predicted rates from combinatorics alone in (A) and from the full model in (B) with $n = 1$ to $n = 4$ from left to right. Errors are standard deviations from at least three independent measurements. Red data points depict rates for repetitive 10AC and

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

14AC sequences. Each plot is labelled with correlation coefficients for the entire dataset (black) and for data omitting the repetitive 10AC and 14AC sequences.

DISCUSSION/CONCLUSION

This study explores the complex processes underpinning DNA hybridisation and sequence-dependent binding kinetics. While previous studies identified secondary structure as a key contributing factor to hybridisation rates (45, 56), we focus on other, equally relevant but poorly defined physical factors to gain a more complete understanding of DNA hybridisation. In particular, by combining careful experimental design and measurements with a physically justified theoretical model, significant progress is made in cementing several principles, proposed to be fundamental to DNA hybridisation mechanisms. These principles are: that the rate of forming nucleating interactions limits the rate of DNA hybridisation (30, 48, 49); that the combination and stability of all possible nucleating interactions is therefore a rate determining factor (45, 50); and that rate-limiting nucleating interactions can be off-register from a fully formed duplex (50). The study experimentally verifies previous predictions that repetitive sequences, which have a greater number of off-register nucleating interactions, hybridise more rapidly than non-repetitive sequences (50). Additional measurements were then performed to capture the variance in hybridisation rates between 41 different strands that have little or no secondary structure (Figure 2). Finally, a simple physically motivated model has been developed that captures a large part of this variance by accounting only for the combination of possible nucleating interactions, including those that are off-register, and their relative stability.

A guiding principle in the construction of our model is to use as few free parameters as possible, to avoid over-fitting and enable the inference of broad physical mechanisms in the context of limited and/or noisy data. However, such a principle is always in tension with a faithful representation of the complicated physical processes that underpin DNA hybridisation. We sought to achieve a useful balance by focussing on two physically plausible rate determiners: the combinatorics and thermodynamic stability of nucleation states. Our model in turn captures these factors with a minimal number of parameters, the nucleation length n and the stability term γ. We have purposely avoided introducing additional confounding factors as much as possible with the exclusive focus on DNA sequences that exhibit very little to no secondary structure. A consequence of this approach is that we do not expect, nor is it our intention, that we will perfectly capture all the observed variance in experimental hybridisation rates. Instead, our goal was to capture as much variance as possible by modelling physically plausible mechanisms with as few parameters as possible. Consequently, the strong correlation achieved between model and experimental data ($\rho = 0.73 \pm 0.03$) provides valuable insight into the physical mechanisms of DNA hybridisation and sequence dependent hybridisation rates.

Our approach is in contrast to several recent studies. For instance, Hata *et al.* achieved a very good correlation with experimental data using a model with a physically motivated component based on

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

secondary structure. However, the high correlation is contingent on an additional sequence dependent component involving 32 free parameters which are difficult to interpret physically (45). Other studies lean more heavily on data driven methodologies, again obtaining reasonable correlations, but requiring the use of a large number of data features. These studies thus provide little insight into physical mechanisms apart from secondary structure, that may be underpinning the process.

The repetitive 10AC and 14AC sequences had the greatest combination of nucleating interactions, particularly at longer nucleation lengths, which resulted in model predictions that were substantially higher than experiment, but only when $n \geq 3$ (Figure 3). Indeed, the results of the model are highly contingent on the choice of the binding nucleation site length, n. Moreover, there was no single choice of nucleation length ($n = 1,2,3\ or\ 4$) that yielded correlations that were drastically better than the others. Here we must emphasise, the restriction to a single nucleation length is highly idealised whereas in reality, nucleation is a progressive and complicated process. Microscopically, a practically innumerable number of nucleation and hybridisation pathways will exist from free strands to full hybridisation, and these pathways will be distinct for different initial interactions between the strands. Along different parts of these pathways further progression will be practically guaranteed, whilst at others it may be highly unlikely. As such, the success of any given binding nucleation length does not imply importance to the exclusion of other characteristic lengths, but simply reflects that it is possible to capture the variance in the data by examining an ostensibly critical part of the nucleation process. In this respect one can view the choice of the nucleation state being n bases as constituting the implicit assumptions that any nucleation state with fewer than n bases will disassociate if they cannot lead to a contiguous nucleation state of n bases, and that binding states of more than n contiguous bases are very likely to lead to full hybridisation. In reality however, it is unlikely that there exists a single threshold nucleating length that fulfils the criteria above. Indeed, plotting the distribution of binding free energies at different n shows considerable overlap in the distribution of binding free energies at different nucleation lengths (Figure S5). These distributions suggest, according to NUPACK predictions, that a larger n does not necessarily translate to a more stable nucleating interaction. Thus, the true threshold length of a nucleating interaction according to the definition above, is highly sequence dependent and will differ not only between DNA strands with different sequences but also within any given DNA strand. One could extend this model to account for variable nucleation lengths, but this would unavoidably lead to a proliferation of associated free parameters, drastically increasing the chance of over-fitting.

The probability that any matched nucleating interaction will proceed towards a fully formed duplex, can be calculated from the stability of nucleating interactions (as determined by NUPACK) combined with the γ value obtained from fits to experimental data according to equation (7). These probabilities for all such interactions in this study are reported in model outputs, which are available in a GitHub repository, with a mean probability of $p_{i,j}^{hybridise} = 0.40 \pm 0.2$, which is similar to previous observations from MD simulations (50). However, model probabilities rely on the stipulation that there is a constant rate of progression from an initial nucleation state to more strongly bound metastable or fully

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

hybridised states, which is also a simplification of the real underlying physical hybridisation process. Attempting to account for such variation would again inevitably introduce large numbers of additional free parameters, risking over-fitting which, even if appropriately implemented, may in turn serve only to obscure more primary physical principles behind the variation of hybridisation rates.

The hybridisation rates of 10 base and 14 base DNA strands were similar with no obvious correlation between sequence length and hybridisation rates (Figure 2I), which informed the decision to normalise the model prediction by the number of possible nucleation states, $(L-n+1)^{-2}$, thus removing length dependence from our model. For completeness however, we also performed optimisations with the negative square replaced by a free exponent, α. Fitted values of α that are less than or greater than -2 would be suggestive of a positive or negative dependence of hybridisation rates with strand length. However, the use of such an additional parameter conferred very little increase in model performance, with optimised values of α very close to -2 (Supplementary Table 2), as emerges from the initial model choice and confirming that the dependence on length in our data is extremely weak. While a weak dependence of hybridisation rates on length cannot be expected to hold true for all lengths of DNA strands, the lack of length dependence in our data could be the result of many possible physical processes. A natural interpretation in terms of the presented model is that the number of nucleation attempts per unit time were constant across sequence lengths, perhaps due to similar effective diffusion coefficients and molecular cross-sections over the range of lengths utilised. In turn this property may be contingent on the designed lack of secondary structure in our data set.

Despite these strong assumptions, our model has favourable properties, which strengthen its claims for a faithful capturing of basal physical processes. First, the model is constructed from minimal free parameters, and second, the variation possible in the model is strongly constrained by physical plausibility arguments. Consequently, the capacity for fitting arbitrary patterns in data is severely constrained. Explicitly, all other factors being equal, the model always assigns greater hybridisation rates to sequences that have a larger number of repetitive sub-sequences and when the stability of those binding states is stronger, with the sole variation controlling the size of such an effect. If some other physical property were more dominant, which conflicted with the property that more stable nucleation sites hybridise faster for example, the model would be unable to capture it. Thus, while our model cannot be taken as a precise account of the hybridisation process, it enables us to conclude the correlations it achieves with data lends strong evidence to the claim that both binding site combinatorics and the stability of those sites are strongly implicated as dominant mechanisms underlying the sequence-dependent hybridisation rates of DNA strands *in vitro*. These findings will be useful for the design of applications in DNA nanotechnology such as DNA PAINT where control over hybridisation kinetics is imperative for achieving adequate signal to noise within practical acquisition times (28, 64, 65) (66, 67). Future work could incorporate our findings with approaches such as by Hata *et al.* (45), whose algorithm explicitly accounts for the consequences of secondary structure on nucleation propensities.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DATA AVAILABILITY

Code for fitting the model to experimental data, all model outputs, and binding energies for all nucleating interactions for all DNA sequences in this study are available in the GitHub repository (https://github.com/llee0905/DNA-bind)

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online. These include Supplementary Figures and Supplementary Tables in Supplementary Data File 1 and Supplemental Notes in Supplementary Data File 2.

FUNDING

This research was supported by the Australian Research Council Centre of Excellence in Synthetic Biology (Grant ID CE200100029) and the National Health and Medical Research Council (Grant ID APP1129234).

REFERENCES

1. WATSON,J.D. and CRICK,F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.

2. Calladine,C.R., Drew,H.R., Luisi,B.F. and Travers,A.A. (2004) Understanding DNA Third Edition. Elsevier Inc.

3. Chen,H., Meisburger,S.P., Pabit,S.A., Sutton,J.L., Webb,W.W. and Pollack,L. (2012) Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc Natl Acad Sci U S A*, **109**, 799–804.

4. Bosco,A., Camunas-Soler,J. and Ritort,F. (2014) Elastic properties and secondary structure formation of single-stranded DNA at monovalent and divalent salt conditions. *Nucleic Acids Res*, **42**, 2064–2074.

5. Chi,Q., Wang,G. and Jiang,J. (2013) The persistence length and length per base of single-stranded DNA obtained from fluorescence correlation spectroscopy measurements using mean field theory. *Physica a-Statistical Mechanics and Its Applications*, **392**, 1072–1079.

6. Seeman,N.C. (1982) Nucleic acid junctions and lattices. *J. Theor. Biol.*, **99**, 237–247.

7. Seeman,N.C. and Sleiman,H.F. (2017) DNA nanotechnology. *Nature Reviews Materials*, **3**, 6451–23.

8. Seeman,N.C. (2003) DNA in a material world. *Nature*, **421**, 427–431.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

9. Shih,W.M., Quispe,J.D. and Joyce,G.F. (2004) A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature*, **427**, 618–621.

10. Goodman,R.P. (2005) Rapid chiral assembly of rigid DNA building blocks for molecular nanofabrication. *Science*, **310**, 1661–1665.

11. Rothemund,P.W.K. (2006) Folding DNA to create nanoscale shapes and patterns. *Nature Publishing Group*, **440**, 297–302.

12. Douglas,S.M., Dietz,H., Liedl,T., Högberg,B., Graf,F. and Shih,W.M. (2009) Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature Publishing Group*, **459**, 414–418.

13. Dietz,H., Douglas,S.M. and Shih,W.M. (2009) Folding DNA into Twisted and Curved Nanoscale Shapes. *Science*, **325**, 725–730.

14. Benson,E., Mohammed,A., Gardell,J., Masich,S., Czeizler,E., Orponen,P. and Högberg,B. (2015) DNA rendering of polyhedral meshes at the nanoscale. *Nature*, **523**, 441–444.

15. Gerling,T., Wagenbauer,K.F., Neuner,A.M. and Dietz,H. (2015) Dynamic DNA devices and assemblies formed by shape-complementary, non-base pairing 3D components. *Science*, **347**, 1446–1452.

16. Berengut,J.F., Wong,C.K., Berengut,J.C., Doye,J.P.K., Ouldridge,T.E. and Lee,L.K. (2020) Self-Limiting Polymerization of DNA Origami Subunits with Strain Accumulation. *Acs Nano*, **14**, 17428–17441.

17. Tang,L. and Li,J. (2017) Plasmon-Based Colorimetric Nanosensors for Ultrasensitive Molecular Diagnostics. *ACS Sens*, **2**, 857–875.

18. Wu,X., Hao,C., Kumar,J., Kuang,H., Kotov,N.A., Liz-Marzán,L.M. and Xu,C. (2018) Environmentally responsive plasmonic nanoassemblies for biosensing. *Chem. Soc. Rev.*, **47**, 4677–4696.

19. Lu,C.-H., Willner,B. and Willner,I. (2013) DNA nanotechnology: from sensing and DNA machines to drug-delivery systems. *Acs Nano*, **7**, 8320–8332.

20. Tyagi,S. and Kramer,F.R. (1996) Molecular beacons: probes that fluoresce upon hybridization. *Nat. Biotechnol.*, **14**, 303–308.

21. Fu,J., Liu,M., Liu,Y., Woodbury,N.W. and Yan,H. (2012) Interenzyme substrate diffusion for an enzyme cascade organized on spatially addressable DNA nanostructures. *J. Am. Chem. Soc.*, **134**, 5516–5519.

22. Fu,J., Yang,Y.R., Johnson-Buck,A., Liu,M., Liu,Y., Walter,N.G., Woodbury,N.W. and Yan,H. (2014) Multi-enzyme complexes on DNA scaffolds capable of substrate channelling with an artificial swinging arm. *Nat Nanotechnol*, **9**, 531–536.

23. Wilner,O.I., Weizmann,Y., Gill,R., Lioubashevski,O., Freeman,R. and Willner,I. (2009) Enzyme cascades activated on topologically programmed DNA scaffolds. *Nat Nanotechnol*, **4**, 249–254.

24. Chang,M., Yang,C.-S. and Huang,D.-M. (2011) Aptamer-Conjugated DNA Icosahedral Nanoparticles As a Carrier of Doxorubicin for Cancer Therapy. *Acs Nano*, **5**, 6156–6163.

25. Douglas,S.M., Bachelet,I. and Church,G.M. (2012) A Logic-Gated Nanorobot for Targeted Transport of Molecular Payloads. *Science*, **335**, 831–834.

26. Zhao,Y.-X., Shaw,A., Zeng,X., Benson,E., Nyström,A.M. and Högberg,B. (2012) DNA origami delivery system for cancer therapy with tunable release properties. *Acs Nano*, **6**, 8684–8691.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

27. Lee,H., Lytton-Jean,A.K.R., Chen,Y., Love,K.T., Park,A.I., Karagiannis,E.D., Sehgal,A., Querbes,W., Zurenko,C.S., Jayaraman,M., *et al.* (2012) Molecularly self-assembled nucleic acid nanoparticles for targeted in vivo siRNA delivery. *Nature Nanotech*, **7**, 389–393.

28. Schnitzbauer,J., Strauss,M.T., Schlichthaerle,T., Schueder,F. and Jungmann,R. (2017) Super-resolution microscopy with DNA-PAINT. *Nat Protoc*, **12**, 1198–1228.

29. Lazurkin,Y.S., Frank-Kamenetskiĭ,M.D. and Trifonov,E.N. (1970) Melting of DNA: its study and application as a research method. *Biopolymers*, **9**, 1253–1306.

30. Pörschke,D. and Eigen,M. (1971) Co-operative non-enzymic base recognition. 3. Kinetics of the helix-coil transition of the oligoribouridylic--oligoriboadenylic acid system and of oligoriboadenylic acid alone at acidic pH. *J Mol Biol*, **62**, 361–381.

31. Morrison,L.E. and Stols,L.M. (1993) Sensitive Fluorescence-Based Thermodynamic and Kinetic Measurements of DNA Hybridization in Solution. *Biochemistry*, **32**, 3095–3104.

32. SantaLucia,J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A*, **95**, 1460–1465.

33. Owczarzy,R., Vallone,P.M., Gallo,F.J., Paner,T.M., Lane,M.J. and Benight,A.S. (1997) Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers*, **44**, 217–239.

34. Gotoh,O. and Tagashira,Y. (1981) Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles. *Biopolymers*, **20**, 1033–1042.

35. Vologodskii,A.V., Amirikyan,B.R., Lyubchenko,Y.L. and Frank-Kamenetskiĭ,M.D. (1984) Allowance for heterogeneous stacking in the DNA helix-coil transition theory. *J. Biomol. Struct. Dyn.*, **2**, 131–148.

36. Breslauer,K.J., Frank,R., Blocker,H. and Marky,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A*, **83**, 3746–3750.

37. Delcourt,S.G. and Blake,R.D. (1991) Stacking energies in DNA. *J Biol Chem*, **266**, 15160–15169.

38. SantaLucia,J., Allawi,H.T. and Seneviratne,P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.

39. Sugimoto,N., Nakano,S., Yoneyama,M. and Honda,K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res*, **24**, 4501–4505.

40. Ghosh,S., Takahashi,S., Ohyama,T., Endoh,T., Tateishi-Karimata,H. and Sugimoto,N. (2020) Nearest-neighbor parameters for predicting DNA duplex stability in diverse molecular crowding conditions. *Proc Natl Acad Sci USA*, **117**, 14194–14201.

41. Banerjee,D., Tateishi-Karimata,H., Ohyama,T., Ghosh,S., Endoh,T., Takahashi,S. and Sugimoto,N. (2020) Improved nearest-neighbor parameters for the stability of RNA/DNA hybrids under a physiological condition. *Nucleic Acids Res*, **48**, 12042–12054.

42. Zadeh,J.N., Steenberg,C.D., Bois,J.S., Wolfe,B.R., Pierce,M.B., Khan,A.R., Dirks,R.M. and Pierce,N.A. (2011) NUPACK: Analysis and Design of Nucleic Acid Systems. *J Comput Chem*, **32**, 170–173.

43. Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

44. Schickinger,M., Zacharias,M. and Dietz,H. (2018) Tethered multifluorophore motion reveals equilibrium transition kinetics of single DNA double helices. *Proc Natl Acad Sci U S A*, **115**, E7512–E7521.

45. Hata,H., Kitajima,T. and Suyama,A. (2018) Influence of thermodynamically unfavorable secondary structures on DNA hybridization kinetics. *Nucleic Acids Res*, **46**, 782–791.

46. Pörschke,D., Uhlenbeck,O.C. and Martin,F.H. (1973) Thermodynamics and kinetics of the helix-coil transition of oligomers containing GC base pairs. *Biopolymers*, **12**, 1313–1335.

47. Zhang,D.Y. and Winfree,E. (2009) Control of DNA strand displacement kinetics using toehold exchange. *J. Am. Chem. Soc.*, **131**, 17303–17314.

48. Pörschke,D. (1971) Cooperative nonenzymic base recognition II. thermodynamics of the helix-coil transition of oligoadenylic + oligouridylic acids. *Biopolymers*, **10**, 1989–2013.

49. Eigen,M. and Pörschke,D. (1970) Co-operative non-enzymic base recognition. I. Thermodynamics of the helix-coil transition of oligoriboadenylic acids at acidic pH. *J Mol Biol*, **53**, 123–141.

50. Ouldridge,T.E., Sulc,P., Romano,F., Doye,J.P.K. and Louis,A.A. (2013) DNA hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic Acids Res*, **41**, 8886–8895.

51. Sambriski,E.J., Schwartz,D.C. and de Pablo,J.J. (2009) Uncovering pathways in DNA oligonucleotide hybridization via transition state analysis. *Proc Natl Acad Sci USA*, **106**, 18125–18130.

52. Markegard,C.B., Gallivan,C.P., Cheng,D.D. and Nguyen,H.D. (2016) Effects of Concentration and Temperature on DNA Hybridization by Two Closely Related Sequences via Large-Scale Coarse-Grained Simulations. *J Phys Chem B*, **120**, 7795–7806.

53. Gao,Y., Wolf,L.K. and Georgiadis,R.M. (2006) Secondary structure effects on DNA hybridization kinetics: A solution versus surface comparison. *Nucleic Acids Res*, **34**, 3370–3377.

54. Chen,C., Wang,W., Wang,Z., Wei,F. and Zhao,X.S. (2007) Influence of secondary structure on kinetics and reaction mechanism of DNA hybridization. *Nucleic Acids Res*, **35**, 2875–2884.

55. Schreck,J.S., Ouldridge,T.E., Romano,F., Sulc,P., Shaw,L.P., Louis,A.A. and Doye,J.P.K. (2015) DNA hairpins destabilize duplexes primarily by promoting melting rather than by inhibiting hybridization. *Nucleic Acids Res*, **43**, 6181–6190.

56. Zhang,J.X., Fang,J.Z., Duan,W., Wu,L.R., Zhang,A.W., Dalchau,N., Yordanov,B., Petersen,R., Phillips,A. and Zhang,D.Y. (2017) Predicting DNA hybridization kinetics from sequence. *Nat Chem*, **10**, 91–98

57. Palau,W. and Di Primo,C. (2013) Simulated single-cycle kinetics improves the design of surface plasmon resonance assays. *Talanta*, **114**, 211–216.

58. Kocman,V. and Plavec,J. (2017) Tetrahelical structural family adopted by AGCGA-rich regulatory DNA regions. *Nat Commun*, **8**, –15.

59. Matsugami,A., Ouhashi,K., Kanagawa,M., Liu,H., Kanagawa,S., Uesugi,S. and Katahira,M. (2001) An intramolecular quadruplex of (GGA)(4) triplet repeat DNA with a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad, and its dimeric interaction. *J Mol Biol*, **313**, 255–269.

60. Jungmann,R., Steinhauer,C., Scheible,M., Kuzyk,A., Tinnefeld,P. and Simmel,F.C. (2010) Single-molecule kinetics and super-resolution microscopy by fluorescence imaging of transient binding on DNA origami. *Nano Lett.*, **10**, 4756–4761.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.17.473246; this version posted December 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

61. Yazawa,K. and Furusawa,H. (2018) Probing Multiple Binding Modes of DNA Hybridization: A Comparison between Single-Molecule Observations and Ensemble Measurements. *ACS Omega*, **3**, 2084–2092.

62. Dupuis,N.F., Holmstrom,E.D. and Nesbitt,D.J. (2013) Single-molecule kinetics reveal cation-promoted DNA duplex formation through ordering of single-stranded helices. *Biophys J*, **105**, 756–766.

63. Whitley,K.D., Comstock,M.J. and Chemla,Y.R. (2017) Elasticity of the transition state for oligonucleotide hybridization. *Nucleic Acids Res*, **45**, 547–555.

64. Jungmann,R., Avendaño,M.S., Dai,M., Woehrstein,J.B., Agasti,S.S., Feiger,Z., Rodal,A. and Yin,P. (2016) Quantitative super-resolution imaging with qPAINT. *Nat. Methods*, **13**, 439–442.

65. Baker,M.A.B., Nieves,D.J., Hilzenrat,G., Berengut,J.F., Gaus,K. and Lee,L.K. (2019) Stoichiometric quantification of spatially dense assemblies with qPAINT. *Nanoscale*, **11**, 12460–12464.

66. Strauss,S. and Jungmann,R. (2020) Up to 100-fold speed-up and multiplexing in optimized DNA-PAINT. *Nat. Methods*, **17**, 789–791.

67. Schueder,F., Stein,J., Stehr,F., Auer,A., Sperl,B., Strauss,M.T., Schwille,P. and Jungmann,R. (2019) An order of magnitude faster DNA-PAINT imaging by optimized sequence design and buffer conditions. *Nat. Methods*, **16**, 1101–1104.