学位論文

Influence of Single-Stranded DNA Structures on Hybridization Kinetics

（一本鎖 DNA の構造がハイブリダイゼーションの速度に及ぼす影響）

平成 27 年 12 月 博士 (理学) 申請

東京大学大学院理学系研究科
物理学専攻

畑 宏明

Abstract

DNA hybridization is a process in which two single-stranded DNA (ssDNA) molecules bind each other through base-pairing to form a double-stranded DNA (dsDNA). Hybridization, especially of oligonucleotides, plays a central role in many important techniques, such as PCR, microarrays, and DNA origami. Therefore, kinetic rates of hybridization crucially affect the efficiencies of those techniques. In this thesis, I describe novel influences of two types of ssDNA structures on hybridization kinetics: thermodynamically unfavorable (ΔG positive) secondary structures and single-strand base stacking (SSBS).

Thermodynamically unfavorable secondary structures show positive Gibbs free energy changes (ΔG) with self-folding. The influence of ΔG positive secondary structures on solution hybridization kinetics was studied using stopped-flow experiments. Observed hybridization kinetics significantly depended on the base sequence, and determined hybridization rate constants differed by two orders of magnitude among the sequences. The difference was correlated with the stability of the secondary structures. To understand the mechanisms underlying the secondary structure dependence of hybridization rate, I proposed a reaction model for the hybridization with positive ΔG secondary structures. This model enabled me to calculate hybridization rate constants from base sequences, and the calculated rates quantitatively agreed with the experimental rate constants. In addition, the analysis of kinetic data derived using the model suggested that SSBS affects hybridization kinetics.

Influences of positive ΔG secondary structures on surface hybridization kinetics were also studied using DNA microarrays. The DNA microarrays provided kinetic data for the hybridization of one hundred strands with different base sequences. I found a similar secondary structure dependence of hybridization rates to that observed in the study of solution hybridization, when the concentration of free ssDNA strands is much higher than that of immobilized strands. However, the dependence was not observed when the free strand concentration is much lower than the immobilized strand concentration. To understand the mechanisms underlying the concentration dependence of secondary structure influence, I developed a reaction model by expanding the

traditional model of hybridization kinetics on solid surfaces.

The SSBS was suggested to have a substantial influence on hybridization kinetics from my study of solution hybridization. To evaluate the influences of SSBS, molecular dynamics (MD) simulations of nucleotide tetramers were carried out. Obtained MD trajectories showed a significant sequence dependence of SSBS stability. The dependence was compared with that suggested from the solution hybridization experiments, which showed that the dependence is compatible with the reaction model developed in the study of solution hybridization.

In the future, the ssDNA structures described in this thesis will be utilized as tools for accurately controlling hybridization rate, which is necessary for further development of various DNA-related techniques. In addition, my insights into DNA hybridization will be applicable to RNA hybridization. Therefore, this study provides further insight into the mechanisms of RNA-RNA interactions, such as RNA interference in gene expression.

CONTENTS

Chapter 1

Nucleic Acid Hybridization

Nucleic acid hybridization is a process in which two single-stranded nucleic acid molecules bind each other through base-pairing to form the double strand. Especially, hybridization between short single-stranded nucleic acid molecules, often called oligonucleotides, (Figure 1.0.1) has been utilized in various biotechnologies such as gene detection [1], amplification [2], and mutation [3]. Oligonucleotide hybridization is highly sequence specific due to the Watson-Crick rules for base-pairing [4]. These rules state that each of the four types of nucleic acid bases binds to a specific partner through hydrogen bonding; for DNA, adenine (A) binds to thymine (T), and guanine (G) binds to cytosine (C). This specificity of hybridization is the origin of the high-fidelity replication of genetic information.

Studies of oligonucleotide hybridization are often divided into two groups by the property focused on: studies of thermodynamics and those of kinetics (Table 1.0.1). The former have investigated how temperature affects the concentration ratio of single strands to double strands at equilibrium. The latter have investigated how the concentration ratio of single strands to double strands changes as a function of time, and have provided insight into the detailed mechanisms underlying hybridization specificity.

The thermodynamics of oligonucleotide hybridization has been studied since the 1960s [5]. The thermodynamic properties have been revealed to depend on several factors; including the length of nucleic acid strand [12], the GC content

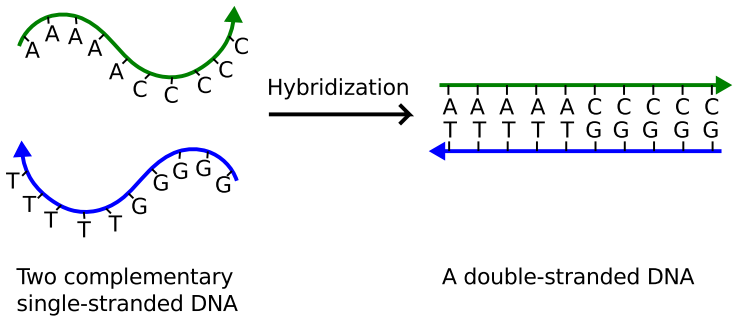

Figure 1.0.1: A schematic of deoxyribooligonucleotide hybridization. Sugar-phosphate backbones of DNA are represented by green and blue arrows (5′-to-3′ direction). Four bases of adenine, thymine, guanine, and cytosine are represented by A, T, G and C, respectively.

Table 1.0.1: Studies about oligonucleotide hybridization/dehybridization.

’53	DNA double helical structure [4]	
’60s	Length, Temperature, pH, GC content, Ionic strength (e.g., [5])	
	KINETICS	THERMODYNAMICS
’70s	N-Z model [6]	Two-state model [7]
’80s		Sequence dependence [8]
’90s		NN parameter [9] (K values was calculated from sequences)[a]
’00s-	Secondary structures [10]	
’06	Stable secondary structures [11]	
’16	Unstable secondary structures [this thesis]	
’16-	(k_{on} needs to be calculated from sequences)[b]	

[a] K represents the equilibrium constant of hybridization/dehybridization.
[b] k_{on} represents the rate constant of hybridization.

(the amount of G and C bases) in the strand [5], and the ionic concentration of solvent [13]. To describe the thermodynamic behavior of hybridization/dehybridization of oligonucleotides, the two-state (all-or-none) model has been widely accepted. In that model, the transition between the two single strands and a double strand is represented by a simple reversible model: A + B $\leftrightarrow$ AB, where A and B denote two complementary single strands, and AB denotes the double strand composed of A and B. Partially hybridized intermediates are ignored. This model introduced the concept of the melting temperature (T_{m}) of double-stranded nucleic acids. At the melting temperature, single strands and double strands exist at the same concentration, and T_{m} values are often used to represent the thermodynamic stabilities of double strands. In order to calculate T_{m} values from base sequences, unified sets of parameters, called Nearest-Neighbor (NN) parameters, were derived from previous experimental observations of the thermodynamics [9, 14]. The NN parameters provide values of enthalpy and entropy changes for every adjacent two base-pair formation, which has enabled the research community to calculate thermodynamic properties for any sequences of double-stranded nucleic acids.

On the other hand, the kinetics of oligonucleotide hybridization has also been studied since the 1960s [15]. Earlier researchers found that hybridization kinetics depends on various conditions, including temperature, length (molecular weight), and the pH of solvent [15]. In addition, the nucleation-zipper model for oligonucleotide hybridization was proposed [1, 16]. In this model, the hybridization process is described as follows:

$$\mathrm{A} + \mathrm{B} \xrightarrow{\textit{nucleation}} \mathrm{AB}^{*} \xrightarrow{\textit{zippering}} \mathrm{AB}, \tag{1.1}$$

where AB^{*} denotes a partially hybridized intermediate, and A, B, and AB are the same as described above. The intermediate AB^{*} contains three or four successively hybridized base-pairs [6]. This model provided a widely accepted explanation for the non-Arrhenius behavior evident in the temperature dependence of hybridization rates as described later. Building on this classical model, researchers later investigated more quantitatively the relationship between base sequence and hybridization kinetics [17–19]. However, compared

with the current understanding of the thermodynamics of oligonucleotide hybridization [9, 20, 21], our understanding of the underlying kinetics is not as sophisticated [22], and the calculation of hybridization rates from sequences remains challenging.

Attempts to further the understanding of hybridization kinetics have recently attracted much attention [23], especially due to the dramatic developments in nucleotide-based nanotechnology, such as dynamic DNA devices [24], *in vivo* DNA computing [25], and the cotranscriptional folding of RNA nanostructures [26]. For example, in the case of DNA computing, the computation is driven by hybridization, and thus, the speed of computation crucially depends on the hybridization rate. Moreover, these nucleotide-based systems are often coupled with other biomolecules like proteins, and therefore, utilization under isothermal conditions is desirable to avoid denaturation of the coupled biomolecules. Under isothermal conditions, the temperature control, which is currently used to avoid kinetic trapping in nucleic acid interactions, is prohibited. Consequently, for the further development of those technologies, a more accurate control of hybridization rates under isothermal conditions is necessary. Thus, detailed insights into the hybridization kinetics and a methodology to calculate hybridization rates are highly desired.

To understand the physical mechanism of oligonucleotide hybridization, the influence of structural properties of single-stranded nucleic acids on hybridization kinetics must be considered. Single-stranded nucleic acid structures are often divided into three levels: primary, secondary, and tertiary structures (Figure 1.0.2) [16]. The primary structure corresponds to a linear chain composed of successive nucleotides (Figure 1.0.2a). This structure is determined by the one-dimensional base sequence information. The secondary structure corresponds to self-folded structure formed by internal base-pairing, which is usually described two-dimensional illustrations (Figure 1.0.2b). For example, hairpin structures, often used in the design of molecular beacons, are classified as a secondary structure. The tertiary structure is formed by three-dimensional folding, or winding, of the secondary structure. For example, the L shape of tRNA, which is formed by folding of the cloverleaf secondary structure, is classified as a tertiary structure. Interestingly, these structures of

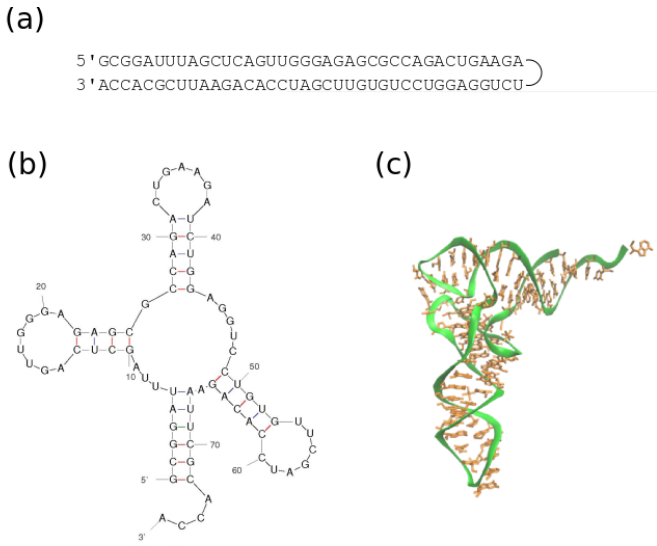

Figure 1.0.2: Three levels of single-stranded nucleic acid structures: examples of yeast phenylalanine tRNA [27]. (a) The primary structure. (b) The secondary structure. (c) The tertiary structure.

single-stranded nucleic acids are dominated by the base sequence information. Therefore, my study of the influence of these structures on hybridization kinetics reveals the relationship between base sequences and hybridization kinetics, which will provide a way to control the time development of interactions between nucleic acid molecules by editing the sequence information.

In this thesis, I report novel influences arising from two ssDNA structures on hybridization kinetics: thermodynamically unfavorable secondary structures and single-strand base-stacking. Both structures have not been considered in hybridization kinetics, because such structures are much less stable than that were investigated in previous researches. In other words, influences of such unstable ssDNA structures on hybridization kinetics have been considered to be ignorable. However, recent developments in DNA-related technologies have made such unstable structure influences unignorable. In the following chapters, I describe effects of unstable ssDNA secondary structures in solution hybridization (Chapter 2), on surface hybridization (Chapter 3), and effects of single-strand base-stacking on hybridization rates (Chapter 4).

Chapter 2

Effects of Unstable Secondary Structure on Solution Hybridization

2.1 Introduction

A ssDNA strand often forms self-folded structures through intramolecular base-paring. These ssDNA structures are called secondary structure. Influences of the secondary structure on the kinetic properties of hybridization in solution phase have been investigated by many researchers [10, 11, 28, 29]. According to those previous researches, ssDNA secondary structures prevent the zippering process in hybridization, because there is a large energy barrier to destruct the secondary structure and to make a fully base-paired dsDNA [11, 28]. This reaction model has explained qualitatively the hybridization kinetics with secondary structures. However, to make a quantitative discussion of the secondary structure effect, much more number of secondary structures need to be studied. Therefore, the quantitative prediction of hybridization rates from base sequences is still far from being available [30].

Furthermore, those previous studies focused only on stable secondary structures, which show a negative free energy change (ΔG) with secondary structure formation (Figure 2.1.1, Left). Unstable secondary structures, which show a positive ΔG ($|\Delta H| < |T\Delta S|$, $\Delta H < 0$, and $\Delta S < 0$, in Gibbs free energy calculation) can also form (Figure 2.1.1, Right). Although such structures are thermodynamically unfavorable, they exist at almost the same concentration with

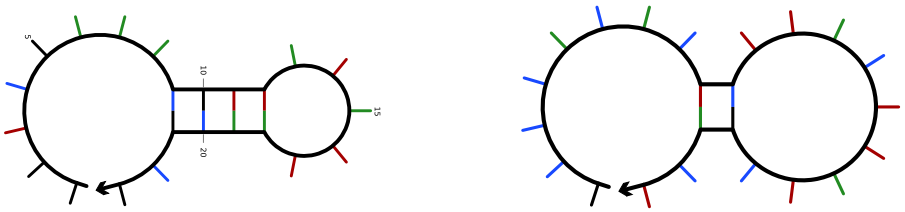

Figure 2.1.1: The classification of ssDNA secondary structures by their thermodynamical stabilities, which is used in this thesis. (Left) An example of thermodynamically stable secondary structures. The predicted value of ΔG is -2.5 kcal/mol ([Na^+]=0.195 mM, $T = 25°C$). The ΔG value was calculated using UNAFold. The image was generated using NUPACK (available at http://www.nupack.org/). (Right) An example of thermodynamically unfavorable secondary structures. The predicted value of ΔG is 2.2 kcal/mol.

the unstructured coil when the absolute value of ΔG is very small. Thus, positive ΔG secondary structures should no longer be ignored in hybridization kinetics, especially under isothermal conditions involved in numerous recent biosensor techniques and DNA network systems [31–35]. Nevertheless, to our knowledge, influences of positive ΔG secondary structures on hybridization kinetics have never been investigated.

In this chapter, I report influences of positive ΔG secondary structures on DNA hybridization kinetics. The hybridization kinetics for 47 pairs of 23-mer oliginucleotides was measured using stopped-flow fluorescence spectroscopy at different temperatures, and their secondary structures and duplex stabilities were studied by thermal melting. The observed hybridization kinetics was significantly dependent on the base sequence, and the rate constant varied by more than two orders of magnitude among sequences that have no negative ΔG secondary structures. The rate constant tended to decrease with the probability of intramolecular base-pairing of positive ΔG secondary structures. Their hybridization obeyed second-order reaction kinetics, and the temperature dependence of the rate constant showed non-Arrhenius behavior. Therefore, their hybridization must be nucleation limited. On the other hand, in the case of the reaction model presented by previous researchers [11, 28] to explain the effect of

stable secondary structures on hybridization kinetics, the rate-limiting process is destruction of secondary structure during zippering process. Thus, the model for stable secondary structures was found not to be applicable in cases of positive ΔG secondary structure. Here, I derived a model describing how these positive ΔG secondary structures affect hybridization kinetics, and the model enabled us to calculate hybridization rate constants from base sequences that agreed well with the experimental hybridization rate constants.

2.2 Materials and Methods

2.2.1 DNA oligonucleotides

A total of 47 pairs of complementary DNA sequences were designed. I denoted one strand in the pair as A and the other as B. Sequences of the A strands are shown in Table 2.2.1, where they are written in the 5′-to-3′ direction. The DNA oligonucleotide sequences used in this study were based on those of the DCN (DNA corded number) components described by Gotoh et al. [36] and were designed *in silico* (i) to contain 23 nucleotides, (ii) to have a similar melting temperature, (iii) to have neither stable misnucleation nor mishybridization, and (iv) to prevent the formation of very stable secondary structures.

Melting temperatures were calculated according to the work of SantaLucia [9]. Sequences were designed to have melting temperatures of 63.0 ± 2.0°C under conditions of $[Na^+]$ = 0.205 M and $[Mg^{2+}]$ = 0 M.

The stability of misnucleation was calculated, and the result indicated that five or more unwanted successive base-pairs rarely formed (Figure A17). In addition, the calculations for strand A (or B) using UNAFold [37] indicated that over 99% of A (or B) existed as monomers rather than dimers (AA or BB) under our experimental conditions (25°C, 50 nM ssDNA, $[Na^+]$ = 0.195 M, and $[Mg^{2+}]$ = 0 M). Therefore, effects of misnucleation and mishybridization were negligible, and we confined this study to intramolecular secondary structures.

Secondary structure prediction of DNA oligonucleotides was performed using UNAFold under our experimental conditions (same as above). Additionally, ‘no isolated base pairs’ and ‘no dangle’ options were employed, according to

Mathews et al. [38] and Zhang et al. [17], respectively. The probabilities of ssDNA intramolecular base pair formation were calculated using the hybrid2.pl program in UNAFold. The predicted ΔG values of the secondary structures ranged from −3.0 to 2.2 kcal/mol (Table 2.2.1). This range included both previously investigated and unexplored range values. Gao et al. [11] reported the effects of secondary structures having negative ΔG values on DNA hybridization kinetics, and the ΔG values of these secondary structures were −4.4 and −1.7 kcal/mol under our experimental conditions.

Of the total 47 DNA sequences used in this study, 32 sequences were predicted to have only positive ΔG secondary structures (Table 2.2.2). Such secondary structures and their influence on hybridization kinetics were studied here for the first time. The other 15 sequences had negative ΔG secondary structures for either or both strands, and they were studied for comparison with previous reports on the effects of a stable (negative ΔG) secondary structure on hybridization kinetics.

All DNA oligonucleotides were commercially synthesized, HPLC purified (SIGMA-Genosys, Hokkaido, Japan), and stored in 1× TE buffer (10 mM Tris·HCl, 1 mM EDTA, pH 8.0) at −20°C until use.

2.2.2 Thermodynamic measurements

Melting curves of ssDNA and dsDNA were recorded in the presence of 0.5× SYBR® Green I fluorescent dye (Lonza, Basel, Switzerland) using a CFX96 real-time PCR detection system (Bio-Rad Laboratories, Inc., Hercules, CA, USA). The excitation wavelength was 494 nm, and emission wavelength was 521 nm (optimal fluorescence signal for SYBR® Green I). A total reaction volume of 50 μL and DNA concentrations of 50 nM were used in the experiments. The buffer solution was saline-sodium citrate (SSC) (150 mM NaCl, 15 mM sodium citrate). The reaction mixtures were cooled from 95°C to 15°C at a rate of 0.5°C/min, and subsequently heated to 95°C at the same rate. Fluorescence intensity changes were recorded every 1 min. The melting temperatures for DNA sequences (Table 2.2.1) were determined from peaks in

Table 2.2.1: List of DNA sequences used in this study, ΔG values, hybridization rate constants, and experimental duplex melting temperatures.

No.[a]	Sequence	ΔG (kcal/mol)[b]		k_{app}	T_m[c]
		A strand	B strand	10^5 $M^{-1}s^{-1}$	°C
0	GCCCACACTCTTACTTATCGACT	2.2	1.7	35 ± 2	61.5
1	AGAGGCTTATAACTGTGTCGGGT	1.6	1.8	21.0 ± 0.7	63.0
2	TGTTCTAAGATTATCCTCCCGCC	1.8	1.7	22.1 ± 0.7	62.0
3	GGCGGCTATAACAATTTCATCCA	1.8	1.4	28 ± 2	63.5
4	TAGCCCAGTGATTTATGACATGC	1.5	1.4	4.3 ± 0.1	64.5
5	GCATCTACACTCAATACCCAGCC	1.4	1.4	43 ± 2	62.5
6	GCCCGTACTGTTGAGATTATGGT	0.96	1.8	9.2 ± 0.4	64.5
7	GCACCTCCAAATAAAAACTCCGC	1.6	0.94	48 ± 4	64.5
8	AGATCAGAGATAGTTACGCCGCA	1.2	1.2	12.5 ± 0.5	66.0
9	TATGTTCCTTACCCCGTTTACCA	1.3	0.96	7.3 ± 0.1	62.0
10	TAGCCAACTCTAAATAACGGACG	1.1	1.1	10.2 ± 0.2	63.5
11	GAAGGAATGTTAAAATCGTCGCG	1.0	1.0	7.7 ± 0.2	64.0
12	TTTGTTTTCCTTATGAGCCAGCC	1.2	0.94	14.3 ± 0.5	62.5
13	GCCCCGATATCTATTTTAGGACG	1.0	1.0	13.5 ± 0.8	62.5
14	CGCAGGAGAGTTAAACGAAAGCA	1.0	0.99	1.038 ± 0.009	64.5
15	GGCTCTATACGATTAAACTCCCC	1.2	0.76	15.5 ± 0.6	61.5
16	CATCTGAACGAGTAAGGACCCCA	1.2	0.76	12.1 ± 0.2	64.5
17	CGTCTATTGCTTGTCACTTCCCC	0.96	0.86	5.23 ± 0.04	65.5
18	AGTCCTTGGTTATCATTCCCTCT	0.85	0.93	10.5 ± 0.5	61.5
19	GGTCTCAGCTAATTTCACACAGA	0.85	0.85	6.3 ± 0.2	62.5
20	TGGGGGGCATAAAACGATACTAG	0.60	1.1	4.70 ± 0.08 *	61.5
21	TGCTCACTTACATTACGTCCATG	0.93	0.70	7.0 ± 0.2	63.0
22	ACCCTTTATCCTGTAACTTCCGC	0.94	0.68	26 ± 2	63.0
23	GCCTAGTGAAACCGTAAGTGCAT	0.91	0.68	1.73 ± 0.03	65.0
24	TCAGCACTCTACTTGACGGACTT	0.74	0.82	0.874 ± 0.007	65.5
25	AGGTTAGGATTTGTCGGGAGATG	-0.049	1.3	31 ± 3	63.0
26	GTCCCGGAAAATACTATGAGACC	0.83	0.36	0.368 ± 0.003	61.5
27	GGCGCTTAAATCATCTTTCATCG	0.73	0.45	5.9 ± 0.2	64.5
28	CCGTCGTGTTATTAAAGACCCCT	0.15	0.62	0.759 ± 0.005	62.5
29	GAGTCAATCGAGTTTACGTGGCG	0.34	0.43	0.488 ± 0.006	65.0
30	TTCGGTTCTCTCCAAAAAAAGCA	0.85	-0.18	2.17 ± 0.03	62.0
31	TGGCACTTATAGCTGTCGGAAGA	-0.15	0.71	1.58 ± 0.01	63.5
32	GAGTCCGCAAAAATATAGGAGGC	-0.083	0.38	22.0 ± 0.5	61.5
33	CGAGAGTCTGTAATAGCCGATGC	0.25	0.031	10.6 ± 0.2	63.0
34	GCCTCACATAACTGGAGAAACCT	0.08	0.17	10.8 ± 0.4	61.5
35	GGCTGTCAATTTATCAGGGAGGC	0.14	0.044	28.3 ± 0.9	63.0
36	TTCGCTGATTGTAGTGTTGCACA	0.29	-0.44	0.773 ± 0.007	65.0
37	ATGGGAACCTAAAAGTGTGGCTA	-0.53	0.14	1.96 ± 0.03	63.5
38	GCATTGAGGTATTGTTGCTCCCA	-0.048	-0.37	30 ± 1	62.5
39	CCATCAGGAATGACACACACAAA	-1.2	-0.037	14.7 ± 0.4	61.0
40	ATGCACCGGTAATATTCCTCTGC	-0.44	-1.0	0.96 ± 0.02 *	64.0
41	CGCAGGAATTAACATGATGAGCG	-1.0	-1.0	1.07 ± 0.02 *	62.0
42	GAAACACTGGATACCTGTGGGAC	-1.9	-0.49	0.342 ± 0.002 *	63.0
43	GGGATAGAACTCACGTACTCCCC	-1.2	-1.6	0.146 ± 0.003 *	62.5
44	GGGATCAGTTGTACACTCCCTAG	-1.3	-1.5	0.585 ± 0.008 *	63.5
45	ATGCGTAACACTCCGTATTGCAT	-1.7	-2.0	0.103 ± 0.001 *	65.0
46	GGTCGAAACGTTATATTAACGCG	-2.5	-3.0	1.93 ± 0.02	64.5

[a] Sequences were aligned from one having the largest value of sum of ΔG for A strand and B strand to one having the smallest value.

[b] Predicted Gibbs free energy change of the most stable ssDNA secondary structure.

[c] The difference in T_m values between duplicated experiments was less than ±0.5°C.

[*] Determined by linear regression of the initial linear region of the data.

Table 2.2.2: The number of sequences having positive/negative ΔG secondary structures.

ΔG		Number of sequences
A strand	B strand	
+	+	32
+	−	2
−	+	4
−	−	9
Total		47

the differential melting curve.

2.2.3 Kinetic measurements

An RX-2000 stopped-flow instrument (Applied Photophysics, Surrey, UK) was used to determine the rate of DNA hybridization for the various sequences. A circulating water bath regulated the temperature of the working solutions, which were held in two separate syringes. Each syringe contained a complementary DNA strand solution of equal DNA strand concentrations (50 nM) that was allowed to equilibrate the indicated temperature before the reaction. The buffer solution was 1× SSC. PicoGreen® fluorescent dye (Invitrogen, Carlsbad, CA, USA) [39, 40], a dye for quantitating dsDNA in the presence of ssDNA, was used to observe duplex formation. The syringe drive delivered equal volumes (0.5 mL) of each solution into the reaction chamber through a rapid mixing chamber, with a dead-time of 8 ms. After the solution was inside the reaction chamber, the fluorescence wavelength at 523 nm (with an excitation wavelength of 502 nm) was measured every 0.5 or 1 s, depending on the total time of the experiment, until equilibrium was reached using a fluorescence spectrophotometer LS 55 (Perkin Elmer, Waltham, MA, USA). The total reaction time was 1800–86400 s. For each duplex forming reaction, at least three injections were performed.

The PicoGreen® intercalating dye bound quickly enough to monitor hybridization (less than 0.5 s, see Figure A1). The dye concentration recommended by the manufacturer was used, and the fluorescence intensity is proportional to the dsDNA concentration under our experimental conditions (Figure A2). Additionally, we confirmed that the hybridization rates obtained by

this method agreed well with those obtained by an absorbance method (Figure A3). In this fluorescence method, hybridization could be monitored with DNA concentrations lower than those used in the absorbance method, thus allowing for (i) slower hybridization, such that we could measure hybridization kinetics with a higher time resolution, and (ii) lower costs for DNA synthesis, which further enabled us to obtain hybridization data for more types of DNA sequences than previous studies.

The normalized fluorescence intensity was obtained in such a way that one normalized unit (n.u.) of intensity corresponded to the average of the final 200 data points of fluorescence intensity. A fluorescence intensity of 0 n.u. corresponded to the minimum fluorescence data point of each trace, and $t = 0$ also corresponds to this point.

2.2.4 Analysis of kinetic data

DNA hybridization rate constants were determined by linear and nonlinear second-order fitting of the fluorescence data, according to Gao et al. [11]. Hybridization of two complementary DNA strands was modeled using the simple irreversible second order process $\mathrm{A} + \mathrm{B} \xrightarrow{k} \mathrm{AB}$. A is an ssDNA strand, B is the complement of A, and AB is the perfect duplex composed of A and B. The concentrations of A, B, and AB were designated by C_A, C_B, and C_AB, respectively. The second-order rate constant for AB duplex formation was notated as k. Initially at time $t = 0$, $C_\mathrm{A} = C_\mathrm{A0}$, $C_\mathrm{B} = C_\mathrm{B0}$ and $C_\mathrm{AB} = 0$. When $t > 0$, $C_\mathrm{A} = (C_\mathrm{A0} - C_\mathrm{AB})$ and $C_\mathrm{B} = (C_\mathrm{B0} - C_\mathrm{AB})$. Hybridization of equal molar A and B strands ($C_\mathrm{A} = C_\mathrm{B} = C$ and $C_\mathrm{A0} = C_\mathrm{B0} = C_0$) can be described by second-order reaction kinetics:

$$\frac{1}{C_0 - C_\mathrm{AB}} - \frac{1}{C_0} = kt. \tag{2.1}$$

The observed fluorescence intensity at time t, $F(t)$, and the maximal fluorescence intensity, F_∞, should be proportional to C_AB and C_0, respectively. At time $t = 0$, $F(0) = 0$. Substituting $F(t)$ and F_∞ into Equation 2.1 yields the equation:

$$\frac{F_\infty}{F_\infty - F(t)} - 1 = C_0 kt. \tag{2.2}$$

Here, I describe a plot for the left hand side of Equation 2.2 versus time as a second-order rate plot. Then, solving for $F(t)$ from Equation 2.2 yields the following:

$$F(t) = \frac{F_\infty C_0 kt}{1 + C_0 kt}. \tag{2.3}$$

The rate constant k was determined by linear fitting (Equation 2.2) or nonlinear fitting (Equation 2.3) to the fluorescence data for three independent experiments using SciPy (available at http://www.scipy.org/). In the former fitting, k was the adjustable fitting parameter. The maximal fluorescence intensity F_∞ was constant and determined by the average of the final 200 fluorescence intensity data points. In the latter fitting, fitting of the data up to 80% hybridization were performed with k and F_∞ as the adjustable fitting parameters. Standard errors of fitting parameters were obtained from the fitting to data containing all of three experiments. The hybridization rate constants obtained were denoted as k_{app}.

2.3 Results

2.3.1 DNA hybridization kinetics

Oligonucleotide hybridization kinetics was determined using a stopped-flow apparatus with fluorescence spectroscopy. Figure 2.3.1 shows DNA hybridization kinetics data obtained at 25°C for 47 DNA sequences. The total number of sequences was much higher than that of previous hybridization kinetic studies (e.g., three sequences in Gao et al. [11]). For every sequence, hybridization was measured three times (Figure 2.3.1, experiments 1–3), and the obtained kinetic traces were highly reproducible. Only results from the initial 70 seconds are displayed, but the reaction was measured until equilibrium was reached, when the fluorescence is unity. I found that the hybridization rate was strongly dependent on the sequence. For example, sequence no. 35 (blue traces in the first panel) formed 90% of duplexes within 60 sec, while sequence no. 29 (purple traces in the second panel) formed only 20%.

In order to test whether the observed oligonucleotide hybridization kinetics were second-order, traditional second-order rate plots of hybridization kinetics

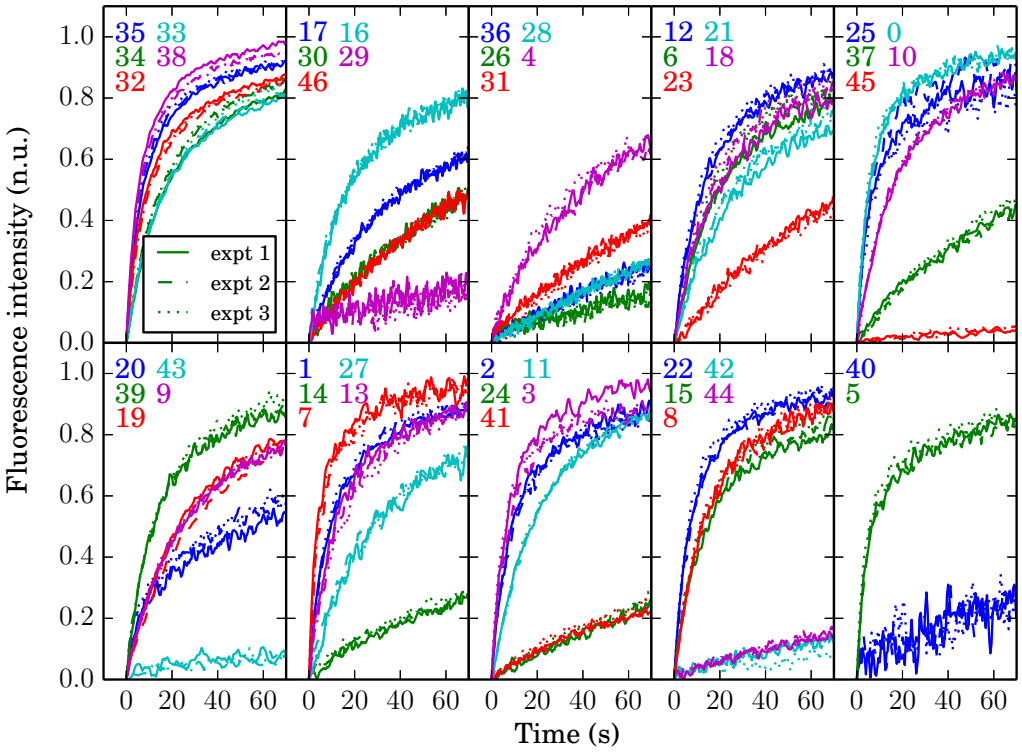

Figure 2.3.1: Hybridization kinetics of 47 DNA sequences. Two complementary DNA strands at 50 nM each were rapidly mixed at 25°C in 1× SSC buffer with PicoGreen®. Colors indicate specific sequences, and the sequence numbers are shown in the upper left of each panel. Time courses for the initial 70 seconds of three independent experiments are shown.

Table 2.3.1: The number of sequences which could be fit with second-order reaction kinetics.

ΔG		Number of sequences	
A strand	B strand	Second-order	Total
+	+	31	32
+	−	2	2
−	+	4	4
−	−	3	9
Total		40	47

for all 47 DNA sequences were constructed (Figure A4), and those of typical sequences are shown in Figure 2.3.2. Figure 2.3.2a–c are the plots for sequences whose hybridization kinetics were able to be fit with a straight line (Equation 2.2), and panels a–c correspond to data from sequences showing relatively fast, medium, and slow hybridization, respectively. Hybridization kinetics for 40 of the total 47 sequences were able to be fit with a straight line. Thus, their hybridization obeyed second-order kinetics. The resulting hybridization rate constants for the 40 sequences are shown in Table 2.2.1, and the number of sequences which could be fit with second-order reaction kinetics was listed in Table 2.3.1. The distributions of rate constants were shown in Figure 2.3.3. The rate constants varied by two orders of magnitude. For 31 of the 40 sequences, the ΔG value of the most stable ssDNA secondary structure was positive for both strands. Six sequences had slightly negative ΔG (e.g., -0.53 kcal/mol) secondary structures for either of the two complementary strands. The other three had negative ΔG values for both strands.

On the other hand, for the other seven of 47 sequences, nos. 20, 40, 41, 42, 43, 44, and 45, their kinetic traces could not be fit to a straight line and curved slightly upward (Figure 2.3.2d and Figure A4). This kind of kinetic trace is similar to that obtained when first-order reactions are plotted with a second-order rate plot. In addition, their hybridization was relatively slow. They completed 80% duplex formation in $\sim 10^3$ seconds, whereas other sequences took $\sim 10^2$ seconds. The predicted ΔG values for these seven sequences were negative for

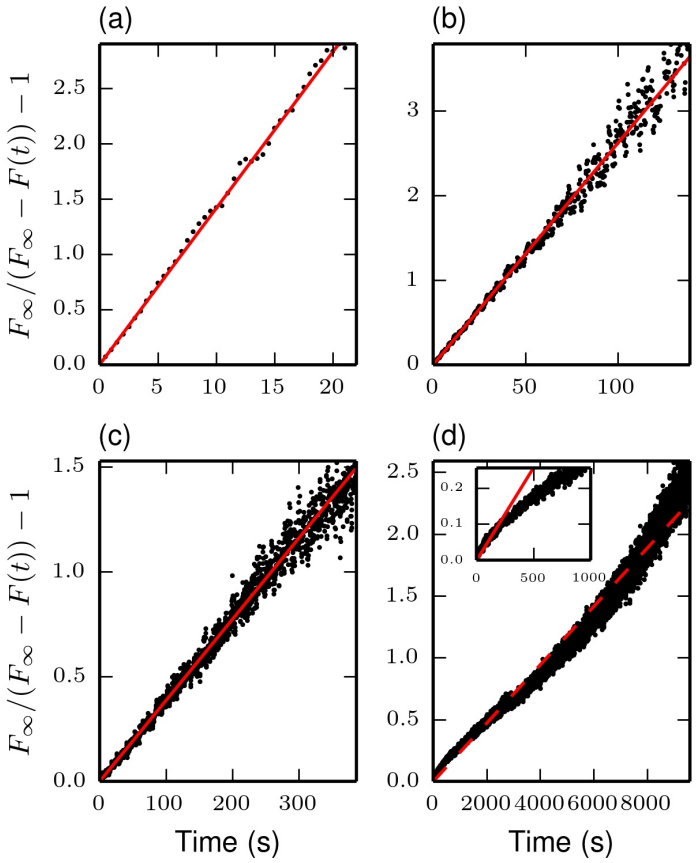

Figure 2.3.2: Second-order rate plots of DNA hybridization kinetics. (a–c) The data from typical sequences with hybridization obeying second-order kinetics. The sequences are (a) no. 35, (b) no. 17, and (c) no. 36. A red line is the best fit to the data up to 80% hybridization of three independent experiments; data from one experiment is shown here. (d) The data from a typical sequence (no. 45) for which hybridization could not be described using second-order reaction kinetics. The broken line is the best fit. The inset contains a plot of the data for the initial 1,000 s. The red line is fit to the initial linear region of data by linear regression.

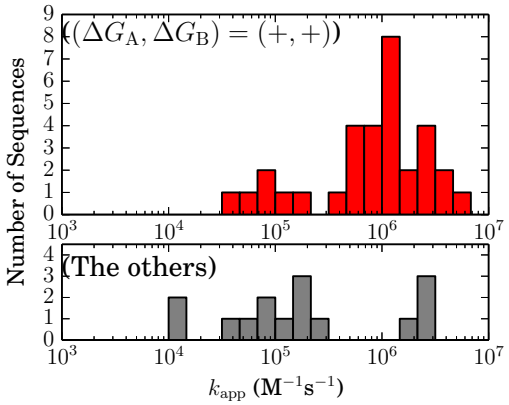

Figure 2.3.3: Distributions of rate constants, k_{app}. (Top) The histogram for sequences having only positive ΔG secondary structures. (Bottom) The histogram for the other sequences.

both of the complementary strands, except for sequence no. 20.

In order to obtain second-order rate constants for the seven sequences, I applied the procedure described by Gao et al. [11], and performed a linear fit of the initial data points to Equation 2.2. The inset in Figure 2.3.2d shows the initial fluorescence data points of sequence no. 45. The kinetic trace initially increased linearly, but then it curved downward after about 200 s. Only data from no. 45 is displayed here, but this trend was also observed for the other six sequences (Figure A5). Gao et al. observed this trend using UV absorbance spectroscopy, and they proposed that the first linear kinetic regime corresponds to the nucleation process. Thus, second-order rate constants for nucleation of the seven sequences were obtained by linear fitting, and the results are shown in Table 2.2.1 with asterisks. These values were used as the k_{app} for each of the seven sequences.

2.3.2 Melting curves

Melting curves of ssDNA and dsDNA were recorded by the fluorescence method. For dsDNA, melting temperatures obtained from these curves are listed in Table 2.2.1. Melting curves for typical sequences are shown in Figure A6. Their melting temperatures fell within the narrow temperature range of 63.5 ± 2.5°C. This result ruled out effects of duplex stability on the hybridization

kinetics measured in this study as shown later. The average T_m was higher than the predicted value at 0.5°C. This would be because an intercalating dye, SYBR® Green I, increases the thermo stability of sequences. Actually, the calculation, obtained following the methods of Mcghee et al. [41] indicated that an approximate 1°C increment in the T_m value was due to dye intercalation. Therefore, the effect of dye intercalation on stability of dsDNA is not substantial in this case.

For ssDNA melting, typical melting curves are shown in Figure 2.3.4. Sequences predicted to have only positive ΔG secondary structures showed a small fluorescence intensity change (i.e., the difference between the fluorescence intensity at a temperature and that at 95°C), ΔF, compared to the change in dsDNA (Figure 2.3.4a). In contrast, ssDNA having negative ΔG secondary structures showed relatively large ΔF values (Figure 2.3.4b). The fluorescence intensity from ssDNA increases as the number of intramolecular base-pairs increases. Thus, the stability of secondary structures (which increases with the number of existing base-pairs) for each sequence qualitatively agreed with the prediction using UNAFold. Quantitatively, in a plot with data from all sequences (Figure A7), I observed a positive correlation between the change in fluorescence intensity within ssDNA melting and the predicted intramolecular base-pair number.

2.3.3 Hybridization rates and secondary structures

I observed that the hybridization rate decreased with the number of base-pairing in ssDNA. Figure 2.3.5 displays this decrease with a plot of the measured hybridization rate constants k_{app} versus the ΔF of ssDNA at 25°C (the correlation coefficient for the data of sequences which could be fit with second-order reaction kinetics, was -0.34 ($p = 0.03$). As the ΔF value increased, the upper limit of the observed hybridization rate constant decreased. The similar trend was found in the same plot for the data of sequences having only positive ΔG secondary structures (Figure A8).

Then, the hybridization rate constants were compared with the ΔG values of

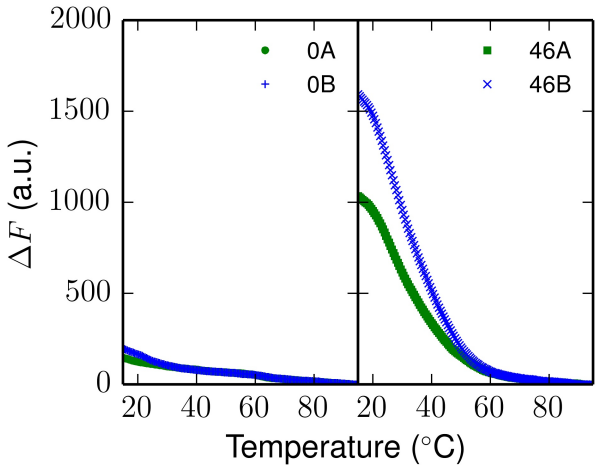

Figure 2.3.4: Melting curves of ssDNA for typical DNA sequences monitored by fluorescence spectroscopy. (a) Thermal melting of two complementary ssDNA strands (0A and 0B) for a sequence with no known negative ΔG secondary structure. The sequence is shown in Table 2.2.1 (sequence no. 0). Samples of 50 nM ssDNA in 1× SSC and 0.5× SYBR® Green I were heated at a ramp rate of 0.5°C/min. The ΔF value represents the difference between the fluorescence intensity at a temperature and that at 95°C. (b) Thermal melting of two complementary ssDNA, as in (a), but for a sequence having negative ΔG ssDNA secondary structures (sequence no. 46).

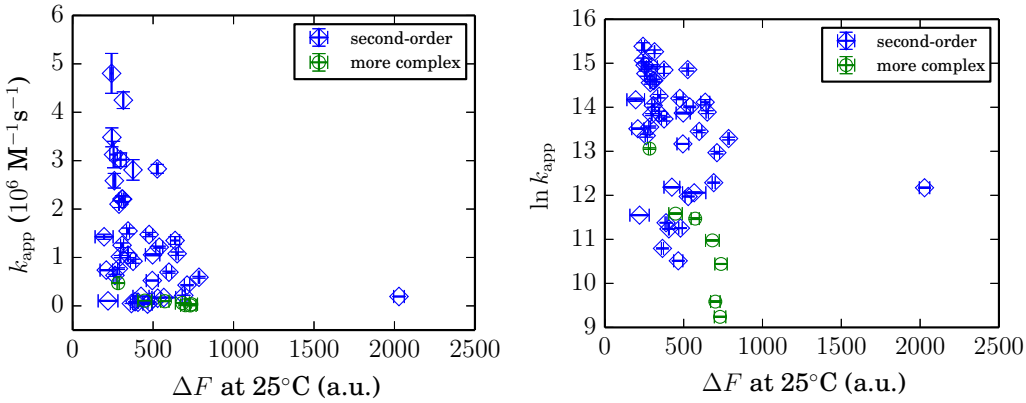

Figure 2.3.5: Comparison of hybridization rates obtained from the stopped-flow experiments and ΔF at 25°C obtained from the ssDNA melting experiments. (Left) The hybridization rate constants were plotted against ΔF values. Blue diamonds correspond to sequences showing second-order reaction kinetics (nos. 0–19 and nos. 21–39). Green circles correspond to sequences showing more complex reaction kinetics (no. 20 and nos. 40–45). Vertical error bars correspond to standard deviations calculated from three independent experiments. Horizontal error bars were calculated from duplicate experiments ($N = 2$). The correlation coefficients between the two values for the second-order sequences, for the more complex sequences, and for all of the sequences were -0.34 ($p = 0.03$), -0.90 ($p = 0.005$) and -0.37 ($p = 0.01$), respectively. (Right) The logarithms of hybridization rate constants were plotted against ΔF values. The correlation coefficients between the two values for the second-order sequences, for the more complex sequences, and for all of the sequences were -0.28 ($p = 0.08$), -0.90 ($p = 0.005$), and -0.37 ($p = 0.01$), respectively.

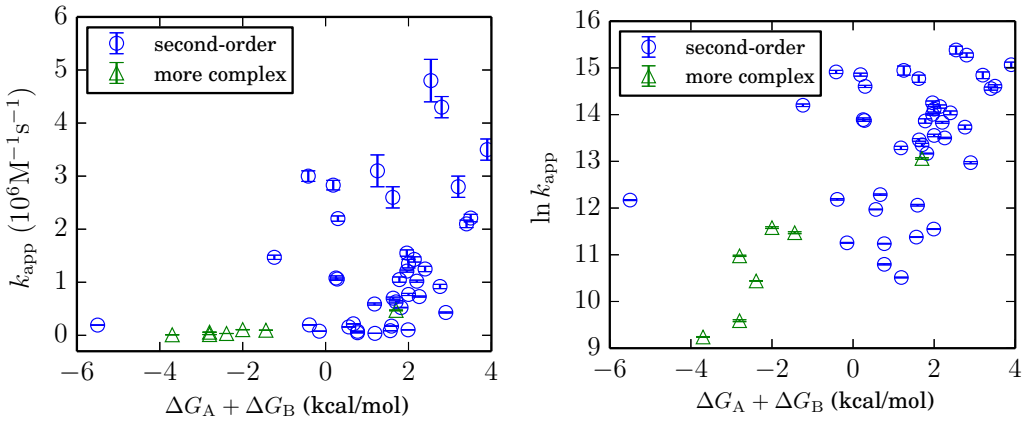

Figure 2.3.6: Comparison of the hybridization rate constants k_{app} and ΔG values predicted by UNAFold. (Left) The k_{app} values were plotted against ΔG values. Blue circles correspond to sequences showing second-order reaction kinetics (nos. 0–19 and nos. 21–39). Green triangles correspond to sequences showing more complex reaction kinetics (no. 20 and nos. 40–45). Vertical error bars correspond to standard deviations calculated from three independent experiments. The correlation coefficients between the two values for the second-order sequences, for the more complex sequences, and for all of the sequences were 0.32 ($p = 0.05$), 0.97 ($p = 0.0004$), and 0.44 ($p = 0.002$), respectively. (Right) The logarithm of k_{app} was plotted against the ΔG value. The correlation coefficients between the two values for the second-order sequences, for the more complex sequences, and for all of the sequences were 0.34 ($p = 0.03$), 0.90 ($p = 0.005$), and 0.61 ($p = 4 \times 10^{-6}$), respectively.

secondary structure (Figure 2.3.6). The decrease of hybridization rates with the degree of secondary structure was also found in this plot (the correlation coefficient for the data of sequences which could be fit with second-order reaction kinetics, was 0.32 ($p = 0.05$). In addition, plots of hybridization rate constants versus the number of predicted base-pairs in ssDNA were made (Figure A9, A10). In the calculation of the number of base-pairs in ssDNA, not only the optimal secondary structures but also suboptimal secondary structures were considered. This plot also showed the tendency of the hybridization rates decrease with the degree of secondary structure. Furthermore, I confirmed that a minor difference in duplex stability does not affect the trend (i.e., a decrease in the upper limit of the hybridization rate constant with the value ΔF) (Figure A11).

2.3.4 Temperature dependence and activation energy of DNA hybridization

To study the temperature dependence of hybridization kinetics, time courses of hybridization at several different temperatures were also determined for some sequences. Figure 2.3.7 shows the kinetic data from two typical sequences; one shows second-order hybridization (Figure 2.3.7a), and the other shows more complex hybridization (Figure 2.3.7b) (for other sequences, see Figure A12). The hybridization rate of the former sequence changed little in the temperature range compared with that of the latter sequence. This is due to a difference in the stability of secondary structures, and a similar trend was observed in a comparison between the hybridization of unstructured coils and hybridization with stable secondary structures [28, 42, 43]. Thus, hybridization with a positive ΔG secondary structure showed a similar temperature dependence as the unstructured-coil hybridization.

For sequences showing second-order hybridization, hybridization rate constants were obtained at 25, 30, 35, 40, 45, and 50°C, and plotted in Figure 2.3.8. The temperature dependence of the hybridization rate constants exhibited non-Arrhenius behavior, and apparent activation energies of hybridization changed from negative at high temperatures to positive at low temperatures. The values of activation energies extracted from the plot are shown in Table 2.3.2. These negative activation energies indicated the existence of metastable intermediates, because the activation energies should be positive independently of temperature when the hybridization is elementary reaction (there is only single energy barrier from reactants to products).

2.4 Discussion

2.4.1 Influence of thermodynamically unfavorable secondary structures on DNA hybridization kinetics in solution

In this chapter, the influences of thermodynamically unfavorable secondary structures on hybridization kinetics were studied. Thermodynamically

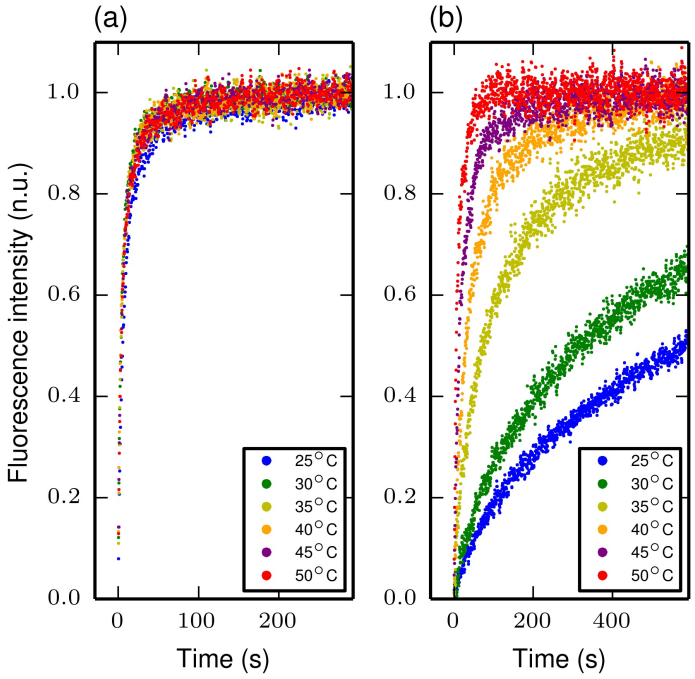

Figure 2.3.7: Hybridization kinetics at 25, 30, 35, 40, 45, and 50°C. (a) Kinetic data from sequence no. 0. which only has positive ΔG secondary structures. Hybridization kinetics were measured for three times for each temperature, and data from one measurement is plotted here. (b) The kinetic data of sequence no. 44. which has some negative ΔG secondary structures.

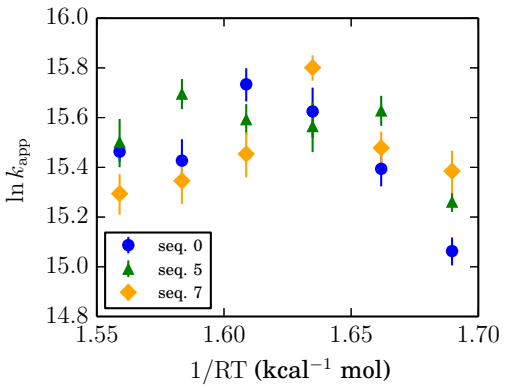

Figure 2.3.8: Arrhenius plots for the hybridization rate constants of sequences nos. 0, 5, and 7 determined by non-linear fitting of the data from three independent experiments. Error bars indicate standard deviations of the fitted variable.

Table 2.3.2: Apparent activation energies in the respective temperature ranges.

No.	E_a [†]	E_a [‡]
	kcal/mol	
0	11 ± 1	-7 ± 4
5	8 ± 4	0 ± 3
7	8 ± 2	-3 ± 1

[†] Fitting temperature range: 25–35°C.
[‡] Fitting temperature range: 40–50°C.

unfavorable, in this thesis, means that the free energy change ΔG with the formation of secondary structure is positive. Of the total 47 DNA sequences used in this study, 32 sequences were predicted to have only positive ΔG secondary structures. Such secondary structures and their influence on hybridization kinetics were studied here for the first time. The other 15 sequences had negative ΔG secondary structures for either or both strands, and they were studied for comparison with previous reports on the effects of a stable (negative ΔG) secondary structure on hybridization kinetics.

For sequences with positive ΔG secondary structures, the observed hybridization kinetics differed significantly from sequence to sequence (Figure 2.3.1), and the determined hybridization rate constants varied two orders of magnitude among the sequences (Table 2.2.1 and Figure 2.3.3). Despite the large differences in rate constants, all of the hybridization obeyed second-order reaction kinetics, except for sequence no. 20 (Figures 2.3.2 and A4). This result means that the hybridization of these sequences is a bimolecular reaction.

In addition, the temperature dependence of the hybridization rates for sequences with positive ΔG secondary structures showed non-Arrhenius behavior (Figure 2.3.8). The same behavior has been found in the hybridization of nucleic acid sequence with no secondary structure [6]. To understand this behavior, the following explanation has been applied; at low temperatures, a diffusion-controlled nucleation process is the origin of positive activation energy E_a, and the negative E_a at high temperature is due to the existence of metastable intermediates in the nucleation process. The metastable intermediates were found to be composed of two complementary ssDNA strands with three or four

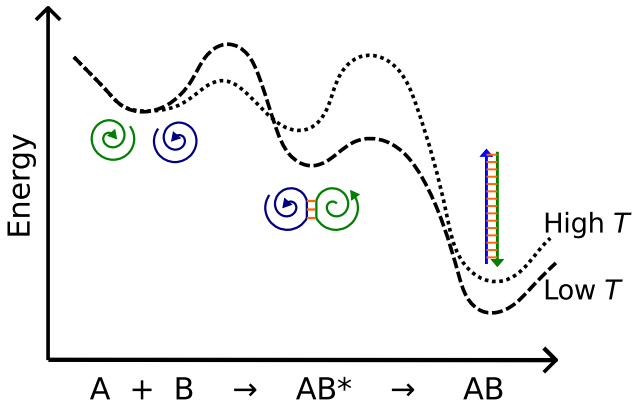

Figure 2.4.1: Schematic view on the profiles of energy versus an effective reaction coordinate for hybridization for high temperatures (dotted line), and for low temperatures (broken line).

successive intermolecular base-pairs [6]. Figure 2.4.1 shows a schematic view on the energy profiles. At low temperatures (Figure 2.4.1, broken line), such intermediates (AB*) are stable enough than dehybridized states (A and B). Therefore, diffusion-controlled nucleation rates increase with temperature, and apparent hybridization rates also increase with temperature. On the other hand, at high temperatures (Figure 2.4.1, dashed line), the intermediates become unstable, and dissociation rates of nuclei increase. Therefore, apparent rate constants of hybridization decrease with temperature. This explanation should apply to the hybridization of sequences with positive ΔG secondary structures and, therefore, their hybridization must be nucleation limited.

The hybridization rates of sequences with positive ΔG secondary structures tended to decrease with the probability of base-pairing in ssDNA (Figure 2.3.5). The probability of base-pairing increases with the stability of positive ΔG secondary structures. Therefore, positive ΔG secondary structures should prevent the nucleation process in hybridization, rather than such thermodynamically unfavorable secondary structures being too unstable to have any substantial effects on hybridization kinetics, as previously believed.

For some sequences with negative ΔG secondary structures, I observed non-second-order hybridization. The second-order rate plots for their hybridization showed the same trend as results reported in [11] (Figure 2.3.2). In

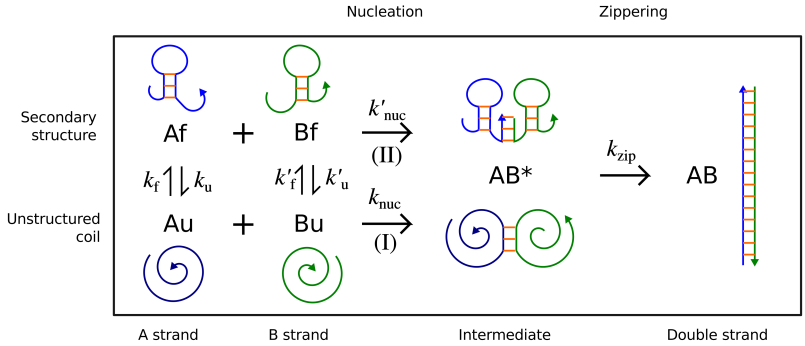

Figure 2.4.2: Schematic of my reaction model for hybridization with positive ΔG secondary structures.

the case of negative ΔG secondary structures, after the nucleation process, there was a large free energy barrier with the destruction of secondary structures [11, 28]. This energy barrier makes the hybridization kinetics more complex. For the negative ΔG secondary structures, my results were in agreement with previous studies.

2.4.2 Hybridization model with thermodynamically unfavorable secondary structures

Here, I derive a reaction model for hybridization with positive ΔG secondary structures from our experimental observations. Figure 2.4.2 shows a schematic of the hybridization. To simplify the problem, both A (or B) strands were assumed to have two conformational states: an unstructured coil Au (or Bu) and a secondary structure Af (or Bf).

Based on the traditional nucleation-zipper model for oligonucleotide hybridization [16], a nucleus forms first from a small number of successive base pairs (nucleation process), and, subsequently, the double helix can zip up (zippering process). In addition, my experimental observations showed that the hybridization was a second-order reaction. Therefore, the following conditions were satisfied because the hybridization was nucleation limited: (i) the zippering rate (k_{zip}) was much larger than the nucleation rate (k_{nuc}), and (ii) transitions between the unfolded state and folded states were much faster than the

nucleation rate (i.e. k_f, $k_u \gg k_{nuc}$). From these two conditions, effects of destruction of positive ΔG secondary structures in zippering process should not be substantial for hybridization kinetics. Also, the destruction before nucleation should not be rate-limiting. Therefore, the observed decrease in the hybridization rate with the stability of positive ΔG secondary structures can be attributed to the decrease in k_{nuc}. At this point, a reaction model in which stable positive ΔG secondary structures increase hybridization rates can not be considered. Thus, the hybridization occurs much more between unstructured coils (Nucleation I in Figure 2.4.2) than between secondary structures (Nucleation II in Figure 2.4.2).

First, I assumed that the hybridization occurs only between unstructured coils (Nucleation I in Figure 2.4.2). Based on this assumption, the effect of a decrease in a concentration of unstructured coils due to the secondary structure formation, can be evaluated. In this case, the rate equation of hybridization is

$$\frac{dC_{AB}}{dt} = k_0 C_{Au} C_{Bu}, \tag{2.4}$$

where C represents the concentration of each conformation. Here, $C_A = C_{Au} + C_{Af}$ and $C_B = C_{Bu} + C_{Bf}$. The nucleation rate between unstructured coils is denoted as k_0 (and k_0 is sequence-independent). By substituting C_A (or C_B) into C_{Au} (or C_{Bu}), the following expression is obtained:

$$\frac{dC_{AB}}{dt} = \frac{k_0}{(1 + K_A)(1 + K_B)} C_A C_B, \tag{2.5}$$

where K_A (or K_B) is the equilibrium constant of transitions between Au (or Bu) and Af (or Bf), and so $K_A = C_{Af}/C_{Au}$ (or $K_B = C_{Bf}/C_{Bu}$). From Equation 2.5, the apparent rate constant of hybridization depends on the value of $1/(1 + K_A)(1 + K_B)$. Figure 2.4.3 shows the plot of the experimental hybridization rate constants k_{app} against the values of $1/(1 + K_A)(1 + K_B)$. The values of K_A and K_B were calculated as $K_A = \exp(-\Delta G_{Af}/RT)$ and $K_B = \exp(-\Delta G_{Bf}/RT)$, respectively. ΔG_{Af} (or ΔG_{Bf}) is the Gibbs free energy change with secondary structure formation at 25°C, and the ΔG value was calculated for the optimal secondary structure for each strand. R is the gas constant, and T is the absolute temperature. In Figure 2.4.3, a weak positive

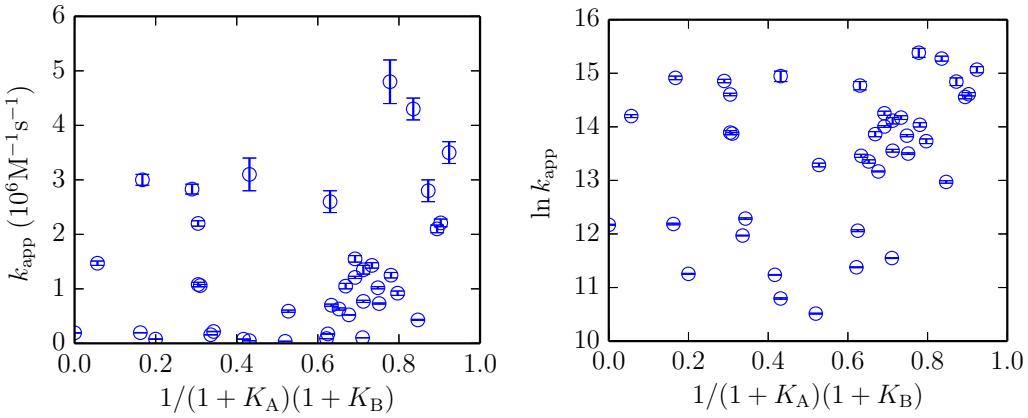

Figure 2.4.3: Comparisons of the hybridization rate constants k_{app} and the values of $1/(1+K_{\mathrm{A}})(1+K_{\mathrm{B}})$. (Left) The hybridization rate constants were plotted in a linear scale. The data of sequences whose hybridization could be fit with second-order reaction kinetics were plotted. The correlation coefficient was 0.27 ($p = 0.09$). (Right) The logarithms of hybridization rate constants were plotted against the values of $1/(1+K_{\mathrm{A}})(1+K_{\mathrm{B}})$. The correlation coefficient was 0.32 ($p = 0.04$).

correlation ($r = 0.27$, $p = 0.09$) was observed. This result suggested that hybridization rates increase with concentrations of unstructured coils, which is compatible with the model in which positive ΔG secondary structures decrease hybridization rates by reducing concentrations of unstructured coils.

Next, effects of suboptimal secondary structures on hybridization were considered. SsDNA strands can form not only the most stable secondary structure (i.e., the optimal secondary structure) but also quasi stable secondary structures (i.e., suboptimal secondary structures). The effect of a decrease in a concentration of unstructured coils due to these suboptimal secondary structures should be considered. When an A (or B) strand has N (or N') secondary structures, the rate equation is

$$\frac{\mathrm{d}C_{\mathrm{AB}}}{\mathrm{d}t} = \frac{1}{(1+\sum_{i=1}^{N} K_{\mathrm{A}i})(1+\sum_{j=1}^{N'} K_{\mathrm{B}j})} k_0 C_{\mathrm{A}} C_{\mathrm{B}}, \tag{2.6}$$

where, $K_{\mathrm{A}i}$ (or $K_{\mathrm{B}j}$) denotes the equilibrium constant of transition between the unstructured coil and the i-th (or j-th) secondary structure. From this equation, the apparent rate constant of hybridization depends on the value of

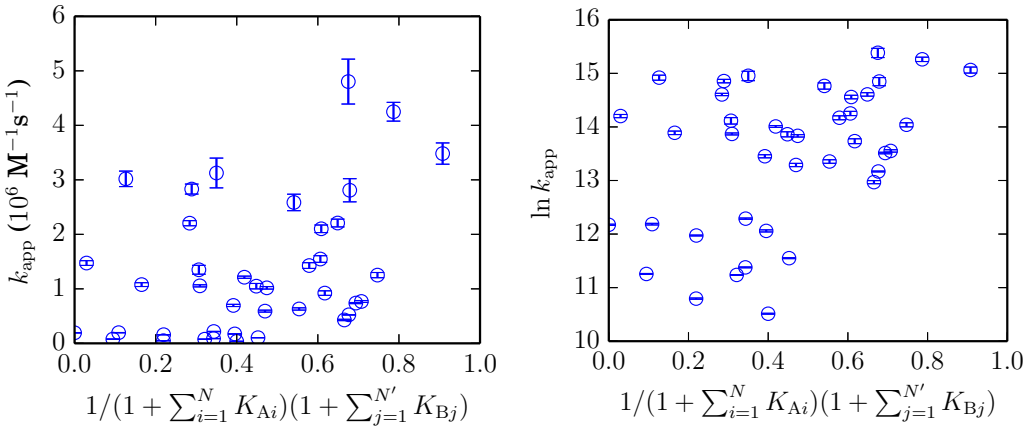

Figure 2.4.4: Comparisons of the hybridization rate constants k_{app} and the values of $1/(1+\sum_{i=1}^{N} K_{\mathrm{A}i})(1+\sum_{j=1}^{N'} K_{\mathrm{B}j})$. (Left) The hybridization rate constants were plotted in a linear scale. The correlation coefficient was 0.38 ($p = 0.02$). (Right) The logarithms of hybridization rate constants were plotted against the values of $1/(1+\sum_{i=1}^{N} K_{\mathrm{A}i})(1+\sum_{j=1}^{N'} K_{\mathrm{B}j})$. The correlation coefficient was 0.42 ($p = 0.007$).

$1/(1+\sum_{i=1}^{N} K_{\mathrm{A}i})(1+\sum_{j=1}^{N'} K_{\mathrm{B}j})$. Figure 2.4.4 shows the plot of the hybridization rate constants, k_{app}, against the values of $1/(1+\sum_{i=1}^{N} K_{\mathrm{A}i})(1+\sum_{j=1}^{N'} K_{\mathrm{B}j})$. In the calculation of values of $1/(1+\sum_{i=1}^{N} K_{\mathrm{A}i})(1+\sum_{j=1}^{N'} K_{\mathrm{B}j})$, suboptimal secondary structures with ΔG values in the range from ΔG_{opt} to ΔG_{opt} + 1 kcal/mol, where ΔG_{opt} represents the ΔG of the optimal secondary structure, were considered. Here, $C_{\mathrm{A}} = C_{\mathrm{Au}} + \sum_{i=1}^{N} C_{\mathrm{Af}i}$, $C_{\mathrm{B}} = C_{\mathrm{Bu}} + \sum_{j=1}^{N'} C_{\mathrm{Bf}j}$, $C_{\mathrm{Af}i}/C_{\mathrm{Au}} = \exp(-\Delta G_{\mathrm{Af}i}/RT)$, and $C_{\mathrm{Bf}j}/C_{\mathrm{Bu}} = \exp(-\Delta G_{\mathrm{Bf}j}/RT)$, where $\Delta G_{\mathrm{Af}i}$ (or $\Delta G_{\mathrm{Bf}j}$) is the Gibbs free energy change with formation of the i-th (or j-th) secondary structure at 25°C. In Figure 2.4.4, the correlation between these two values was 0.38, which increased by 0.11 compared with that obtained when only the optimal secondary structures are considered. This shows that including suboptimal secondary structures into the consideration is reasonable to assess secondary structure effects on hybridization kinetics.

Furthermore, there exists nucleation not only (i) between unstructured coils but also (ii) between a unstructured coil and a secondary structure, and (iii) between a secondary structure and a secondary structure. Although these nucleation ((ii) and (iii)) occurs less frequently than that between unstructured coils as described

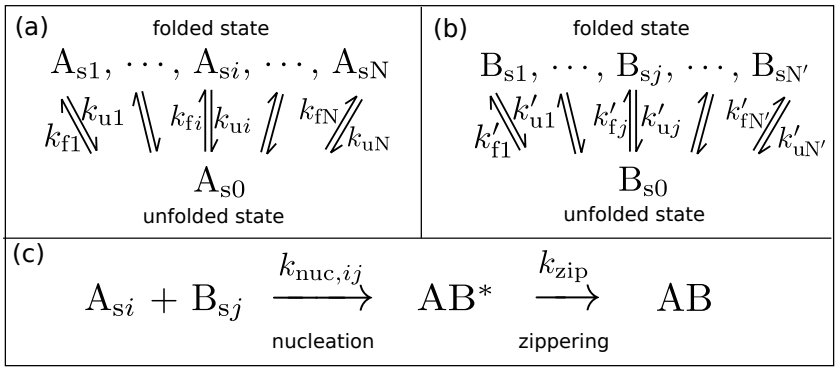

Figure 2.4.5: A schematic of hybridization with positive ΔG secondary structures. (a) An A strand has an unfolded state and N folded states (positive ΔG secondary structures). Transitions between these states occur with state-dependent rate constants. (b) Transitions for B strands. (c) Hybridization is divided into two processes of nucleation and zippering. Nucleation occurs between pair of any strand states of A and B. The nucleation rate constant depends on the pair, because the number of nucleation sites differs from state to state. The subsequent zippering is much faster than the nucleation, and the dependence of the rate constant on zippering of the pair is not considered substantial.

above, consideration of these reaction can improve the evaluation of secondary structure effects on hybridization rates. Figure 2.4.5 shows a schematic of the reaction. In this case, nucleation can occur between any pair of strand states of A and B, which is followed by zippering (Figure 2.4.5c). The rate equation is

$$\frac{\mathrm{d}C_{\mathrm{AB}}}{\mathrm{d}t} = \sum_{i=0}^{N}\sum_{j=0}^{N'} \frac{k_{ij}C_{\mathrm{A}si}C_{\mathrm{B}si}}{C_{\mathrm{A}}C_{\mathrm{B}}} C_{\mathrm{A}}C_{\mathrm{B}}, \tag{2.7}$$

where A_{s0} (or B_{s0}) represents the unstructured coil. $\mathrm{A}_{si}(1 \le i \le N)$ (or $\mathrm{B}_{sj}(1 \le j \le N')$) represent N (or N') secondary structures. Nucleation rates between $\mathrm{A}_{si}(0 \le i \le N)$ and $\mathrm{B}_{sj}(0 \le j \le N')$ is denoted as k_{ij}. From this equation, the apparent rate constant of hybridization depends on the value of $\sum_{i=0}^{N}\sum_{j=0}^{N'}(k_{ij}C_{\mathrm{A}si}C_{\mathrm{B}si})/(C_{\mathrm{A}}C_{\mathrm{B}})$.

Here, to estimate the values of k_{ij}, two factors of secondary structure which can affect nucleation rates were considered: (i) the number of nucleation sites in a strand, and (ii) the nucleation capability of each nucleation site. The former can

be determined by the predicted secondary structure. When the nucleation rate k_{ij} depends on the number of nucleation sites in secondary structures, the following expression can be obtained:

$$k_{ij} = k_0 \frac{n_{\mathrm{A}_{si}}}{n_{\mathrm{A}_{\mathrm{S}0}}} \frac{n_{\mathrm{B}_{sj}}}{n_{\mathrm{B}_{\mathrm{S}0}}}, \tag{2.8}$$

where n represents the number of nucleation sites for each conformation. The nucleation rate between unstructured coils is denoted as k_0. Substituting Equation 2.8 into Equation 2.7 yields

$$\frac{\mathrm{d}C_{\mathrm{AB}}}{\mathrm{d}t} = k_0 \sum_{i=0}^{N} \frac{C_{\mathrm{A}_{si}}}{C_{\mathrm{A}}} \frac{n_{\mathrm{A}_{si}}}{n_{\mathrm{A}_{\mathrm{s}0}}} \sum_{j=0}^{N'} \frac{C_{\mathrm{B}_{sj}}}{C_{\mathrm{B}}} \frac{n_{\mathrm{B}_{sj}}}{n_{\mathrm{B}_{\mathrm{s}0}}} C_{\mathrm{A}} C_{\mathrm{B}}, \tag{2.9}$$

where the subscript i (or j) denotes the i-th (or j-th) conformational state of the strand. The 0-th state is the unstructured coil, and the 1st strand state is the optimal (most stable) secondary structure. From this equation, the apparent rate constant of hybridization depends on the value of $k_0 \sum_{i=0}^{\mathrm{N}} (C_{\mathrm{A}_{si}} n_{\mathrm{A}_{si}})/(C_{\mathrm{A}} n_{\mathrm{A}_{\mathrm{s}0}}) \sum_{j=0}^{\mathrm{N}'} (C_{\mathrm{B}_{sj}} n_{\mathrm{B}_{sj}})/(C_{\mathrm{B}} n_{\mathrm{B}_{\mathrm{s}0}})$. This value is denoted as k_{calc} in the following.

Then, the nucleation capability of each nucleation site can be determined mainly by three factors: the steric hindrance, flexibility, and base sequence of each nucleation site. Figure 2.4.6 is a schematic view of these factors. For steric hindrance, parameter δ_{sth} was adopted. Nucleation sites located at stems, and also around stem regions, can be inaccessible due to steric hindrance. In such nucleation sites, $\delta_{\mathrm{sth}} = 0$; otherwise, $\delta_{\mathrm{sth}} = 1.0$. For flexibility, bases involved in loops, and also stems, are less flexible than bases in coil regions. This effect was represented by parameter δ_{flx}, which is unity when the nucleation site is in coil regions. For the base sequence, parameter δ_{seq} was adopted. The value of δ_{seq} represents the difference in the nucleation capability among base sequences at a nucleation site. To sum up, the nucleation capability, δ, of a nucleation site is represented by the following:

$$\delta = \delta_{\mathrm{sth}} \delta_{\mathrm{flx}} \delta_{\mathrm{seq}}. \tag{2.10}$$

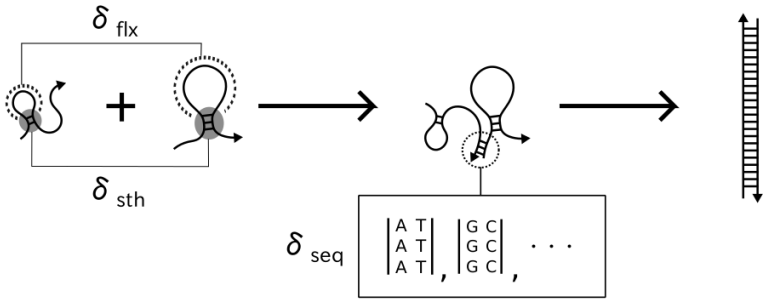

Figure 2.4.6: A schematic view of three parameters (δ_{sth}, δ_{flx}, and δ_{seq}) determining the nucleation capability. Triangle-head arrows represent single DNA strands. Lines connecting the arrows represent base pairs. In gray regions, nucleation is prevented by steric hindrance, whose effect is represented by δ_{sth}. Dotted lines indicate loops in which strand flexibility decreases, whose effect is represented by δ_{flx}. The dotted circle indicates a nucleus, which can be composed of different base sequences and whose effect is represented by δ_{seq}.

Considering the nucleation capability, an effective number of nucleation sites, n^*, of a secondary structure formed by a 23-nt strand is calculated by the equation: $n^* = \sum_{m=1}^{23-L_{\mathrm{nuc}}+1} \delta_m$, where L_{nuc} is the number of bases composing a nucleation site, and δ_m represents δ of the m-th nucleation site in a strand. Substituting n^* for n in k_{calc} calculation yields an equation for evaluating the nucleation rate:

$$k_{\mathrm{calc}} = k_0 \sum_{i=0}^{\mathrm{N}} \frac{C_{\mathrm{A}_{si}}}{C_{\mathrm{A}}} \frac{n^*_{\mathrm{A}_{si}}}{n^*_{\mathrm{A}_{s0}}} \sum_{j=0}^{\mathrm{N}'} \frac{C_{\mathrm{B}_{sj}}}{C_{\mathrm{B}}} \frac{n^*_{\mathrm{B}_{sj}}}{n^*_{\mathrm{B}_{s0}}}. \tag{2.11}$$

As a first approximation, we assumed $\delta_{\mathrm{seq}} = 1$ for all base sequences. The values of δ_{sth} and δ_{flx} were determined to maximize the correlation coefficient between k_{app} and k_{calc}. As a result, for nucleation sites located in stem regions and within two bases from a stem regions, $\delta_{\mathrm{sth}} = 0$, otherwise, $\delta_{\mathrm{sth}} = 1.0$. Also, $\delta_{\mathrm{flx}} = 1.0$ for all nucleation sites even though they are in loop regions. These results were obtained with the condition $L_{\mathrm{nuc}} = 3$ (Figure A13a). In the case of $L_{\mathrm{nuc}} = 4$, a similar trend was observed (Figure A13b). Thus, we used the above values of δ_{sth} and δ_{flx} with $L_{\mathrm{nuc}} = 3$. The k_{calc} value calculated with an approximation of $\delta_{\mathrm{seq}} = 1$ is denoted as k'_{calc}. Figure 2.4.7a shows a plot of k_{app}

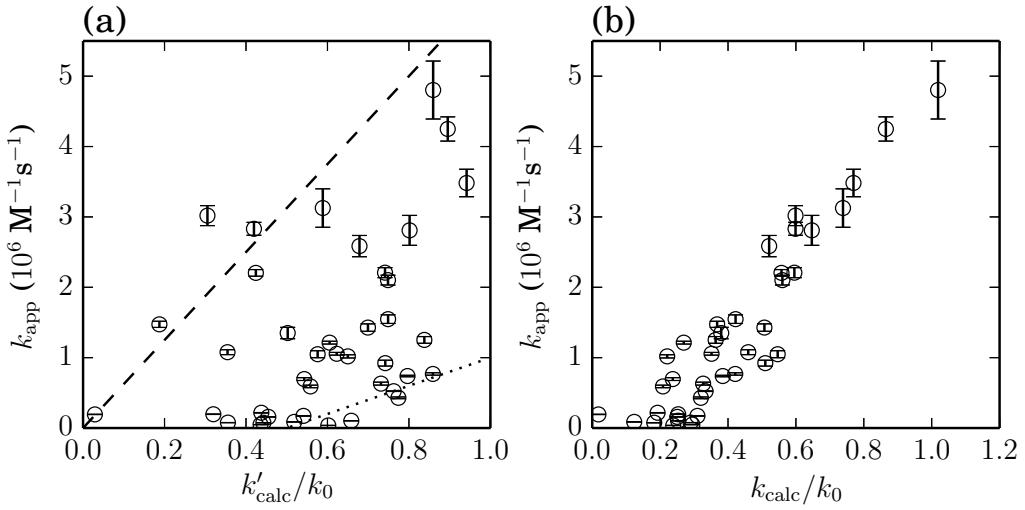

Figure 2.4.7: A comparison of the observed hybridization rate, k_{app}, and the calculated hybridization rate. (a) The value of k'_{calc}/k_0 calculated under the assumption of $\delta_{\mathrm{seq}} = 1$ for all base sequences. Dashed and dotted line are drawn to guide the eye. The dashed and the dotted line have slopes of 6.3 and 2.0, respectively. (b) The k_{calc}/k_0 value was determined with δ_{seq} values listed in Table 2.4.1. The correlation coefficient is 0.92.

versus k'_{calc}. The upper bound (dashed line) and the lower bound (dotted line) of k_{app} increased with k'_{calc}. However, the relationship between them is not completely proportional. Even though the sequences have almost the same k'_{calc} value, there is a large difference in the hybridization rates for the different sequences. For example, the sequences with k'_{calc}/k_0 values of about 0.8 have rate constants that vary by at least an order of magnitude.

Then, we parameterized δ_{seq} for every base triplet. The values of δ_{seq} were determined to maximize the correlation coefficient between k_{app} and k_{calc}. As a result, we obtained a clear proportional relationship between them (Figure 2.4.7b). Determinations were made using Scipy (see Appendix C) with a constraint that triplets complimentary each other have the same δ_{seq} value. The obtained δ_{seq} values are shown in Table 2.4.1. All triplets were found at least 16 times in the sequences studied (Table B1), thus over-fitting did not occur in this case. In addition, the uncertainties of determined δ_{seq} parameters were evaluated using Monte Carlo simulations (see Appendix C). Furthermore, to evaluate the

adequacy of δ_{seq} values, samples were randomly divided into two groups: one for the determination of δ_{seq} values (learning samples), and the other for testing the utility of the k_{calc} calculation (test samples). The results are shown in Table B2. The correlation of k_{app} and k_{calc} for test samples (r_{test}) was improved by the δ_{seq} determination for any number of learning samples (Δr_{test} was positive for all cases). This shows that the obtained δ_{seq} values were not artifacts, but were related to kinetic properties of sequence-dependent nucleus formation. The value of r_{test} increased with the number of learning samples, so we used δ_{seq} values determined from all 40 samples (Table 2.4.1) in the following discussion. The parameterization of δ_{seq} for every base singlet and doublet were also performed, but the optimized correlations (0.53 and 0.64, respectively) between k_{app} and k_{calc} were lower than that for triplet.

Table 2.4.1: List of δ_{seq} parameters for base triplets.

Sequence	δ_{seq}	Sequence	δ_{seq}
AGG/CCT	3.1	GCG/CGC	0.41
GTG/CAC	2.4	AAA/TTT	0.33
GAG/CTC	2.0	TAG/CTA	0.18
TTG/CAA	2.0	CGG/CCG	0.16
ATC/GAT	1.8	CAG/CTG	0.079
GGC/GCC	1.7	AGA/TCT	0.049
AAT/ATT	1.6	GGG/CCC	0.0095
ATA/TAT	1.6	AAC/GTT	0.0
TGG/CCA	1.6	ATG/CAT	0.0
ACA/TGT	1.4	AGT/ACT	0.0
ACC/GGT	1.3	AGC/GCT	0.0
TCC/GGA	1.3	ACG/CGT	0.0
TCG/CGA	1.0	TAA/TTA	0.0
TAC/GTA	0.97	TTC/GAA	0.0
TGC/GCA	0.84	TGA/TCA	0.0
AAG/CTT	0.60	GAC/GTC	0.0

2.4.3 Sequence dependence of nucleation capability

In order to gain insights into the mechanisms underlying the sequence dependence of nucleation capability (δ_{seq}), we first compared δ_{seq} values with

nucleus stability (ΔG_{nuc}) calculated from the NN parameter [9], but no clear correlation was observed (Figure A14). Then, the δ_{seq} values were compared with entropy change (ΔS_{nuc}) in nucleation. The plot is shown in Figure 2.4.8. The ΔS_{nuc} values were calculated from the NN parameters for all base triplets. In Figure 2.4.8, the δ_{seq} values showed a tendency to increase with ΔS_{nuc}. This result suggests that a sequence having a smaller entropy change with nucleation is preferable as a nucleation site in hybridization kinetics.

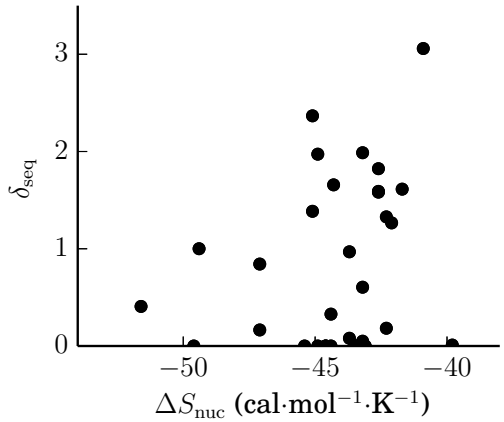

Figure 2.4.8: A plot of the δ_{seq} values vs. the ΔS_{nuc} values for all base triplets. The value of ΔS_{nuc}, the entropy change in nucleation (i.e., three successive base pairs formed), was calculated using Nearest-Neighbor parameters [9].

A promising candidate causing the tendency observed between ΔS_{nuc} and δ_{seq}, is the sequence dependence on the stability of single-strand base-stacking. A negative ΔS_{nuc} value represents a loss of degrees of freedom in the conformational change from ssDNA to dsDNA. Assuming that dsDNA has no sequence dependence for the degrees of freedom of conformation, the difference in the value of ΔS_{nuc} among sequences is attributed to the difference in the degrees of freedom of the ssDNA conformation. A sequence with a smaller reduction in entropy has a ssDNA conformation with a smaller degree of freedom. The degree of conformational freedom in ssDNA conformation is reduced by single-strand base-stacking [16, 44], whose stability is known to have a sequence dependence [45]. The structure of a strand with single-strand base-stacking is similar to that of a strand in dsDNA. Therefore, a base-stacked

nucleation site in a single strand should show a smaller conformational change during the nucleation process than a base-unstacked one. This would make the base-stacked nucleation site form a nucleus more rapidly. The nucleation rate increases with the value of δ_{seq}. Thus, the δ_{seq} value showed a tendency to increase with ΔS_{nuc} as shown in Figure 2.4.8. This idea that the sequence dependence of single-strand base stacking stability is related to that of the nucleation rate could shed light on detailed mechanisms of recognition and binding between nucleic acid strands.

2.4.4 Temperature dependence of the hybridization rate

To evaluate the effect of positive ΔG secondary structures on the temperature dependence of the hybridization rate, we calculated the temperature dependence of the value of k_{calc} using UNAFold with the determined δ_{seq} parameters (Figure A15). The value of k_{calc} gradually increased with temperature, because intramolecular base pairs are less stable at higher temperatures. The increments of k_{calc} were within 3–6% within the temperature range from 25–35°C, while k_{app} increased 41–93% in the same temperature range. Thus, the decrement in the stability of positive ΔG secondary structures corresponded to only 4–15% in the increment of k_{app}. This indicates that hybridization in such a low temperature range is a diffusion-controlled reaction, which is consistent with our hybridization model described above.

Chapter 3

Effects of Unstable Secondary Structure on Surface Hybridization

3.1 Introduction

Nucleic acid hybridization on surfaces, where either strands of complementary pairs are immobilized on phase-interfaces, is a process exploited by various recent biotechnologies; including DNA microarrays [46], Northern/Southern blotting [47], and certain biosensors [48]. The efficiency of those technologies (e.g., accuracy, sensitivity, and assay time) is strongly affected by the hybridization rate.

The concentration of nucleic acid strands is one of the most fundamental factors influencing the hybridization rate. However, for surface hybridization, the hybridization rate does not simply increase with the strand concentration. Due to the immobilization of either strand, both the amount and the surface density of the immobilized strand has substantial influence on hybridization rate. To comprehensively understand the influence of strand concentration, previous researchers have described a reaction mechanism [49], in which two pathways for the hybridization of complementary strands were considered: (i) direct hybridization from bulk solution to the surface, and (ii) non-specific adsorption of free strands onto the surface followed by two-dimensional diffusion to hybridize with immobilized strands. This model successfully explained the influence of strand immobilization conditions on the hybridization rate (or

efficiency), and also enabled the determination of optimal strand densities maximizing the surface hybridization rate.

Stable secondary structures have also been considered in the design of nucleic acid sequences used for those technologies utilizing surface hybridization [50]. Similar to what is seen during solution hybridization, the secondary structure of a single-stranded nucleic acid is known to delay hybridization [11]. However, thermodynamically unfavorable secondary structures, which are described in the previous chapters, have never been investigated for their influence on surface hybridization kinetics. If the same influence could be found on surface hybridization as on the hybridization in solution, the positive ΔG secondary structures should be considered in the sequence design for various techniques described above. Doing this would likely improve the efficiency of those techniques.

In this chapter, I report my investigations into the influence of thermodynamically unfavorable secondary structures on hybridization kinetics using a microarray method. The kinetics of surface hybridization was measured at different concentrations of free DNA strands, which elucidated that this influence is concentration dependent.

3.2 Materials and Methods

3.2.1 DNA sequences

A list of DNA sequences used in this study is shown in Table 3.2.1. Similar to those described in the study of solution hybridization, these sequences were designed in the previous study [36]. In contrast to the measurements of hybridization in solution, these sequences were mixed into a single solution for the following measurement. To make sequences hybridize only between complementary pairs, base-pair formation between non-complementary pairs was considered in the sequence design. These DNA sequences do not form successive base-pairs more than six between non-complementary pairs. In addition, the cross hybridization between non-complementary pairs was experimentally evaluated to be negligible as described later. Secondary structure

Table 3.2.1: DNA sequences used in this study.

No		Sequence	ΔG		No		Sequence	ΔG	
C3[a]	C2[b]		A strand	B strand	C3[a]	C2[b]		A strand	B strand
1		TGTTCTCTGACCAATGAATCTGC	0.72	0.56	51		GCCTATATGAACCAAGCCACTGC	0.25	0.031
2		TGGAACTGGGAACGCTTTAGATG	0.85	1.8	52		CGCCGTCAGTACTTGTATAGATG	0.88	1.1
3		TTCGCTTCGTTGTAATTTCGGAC	0.72	0.62	53		GTCGGTATCGAAAAGGTACTGCA	-1.1	-1.0
4	4	TAGCCCAGTGATTTATGACATGC	1.5	1.4	54		AGGCAGTTCAACCTATATCTGCG	-2.00	-2.1
5		CGCTCTGGTTACTATTGGACGTT	1.3	1.4	55		GGTCGTAACATTGAGAGGAGACG	0.001	-0.01
6	9	TAGCCAACTCTAAATAACGGACG	1.1	1.1	56		GGCGATTTATTGCTAACTGGCTA	-0.01	-0.53
7		TTCGGTTGTCGATATGAGGATCT	-1.3	-1.2	57		GCACTACCGCTAACTATACGCTA	0.65	-0.96
8		GGGGGGTACTTCATACAAGATGC	0.11	0.098	58		GGCTCGTAGTACTCCTTACATGC	0.65	0.42
9		GAGTAGCAGGCAAATACCCTAGA	0.34	-0.15	59		GGCTCTACAAACTTGTGTCCATG	-0.64	-0.64
10		GCCTATTAAGGTCTACGTCATCG	-1.0	-0.41	60		GGTGGAGTGAATCTCACTAGACT	-2.3	-2.3
11		AGTCATACAGTGAGGACCAAATG	-0.24	-0.29	61		CTAGCACAATTAATCAATCCGCC	0.98	0.65
12		CATTCGACATAAGCTGTTGATGC	-1.5	-0.45	62		GCAGCTGAATTGCTATGATCACC	-0.80	-0.80
13	21	TGCTCACTTACATTACGTCCATG	0.93	0.70	63		GCCTATAGTGTCGATTGTCCTCG	-0.26	-0.54
14		TACACCTATCAACTCGTAGAGCA	-1.4	0.20	64		CGATCACGGATTAATGTCACCCC	-0.15	-0.053
15		AGGTCCGGTAGTAATTTAGGTGC	1.3	1.7	65		AAGAGATTTAACTTGAGCTCGCC	0.63	0.73
16		TGCACTCTGATATATACAGGCCA	0.044	0.14	66		TTTGTTGTTCGATATCAGGCGTG	0.92	1.0
17		GCAGCCCTTATAGATAACGGGAC	-0.45	-1.4	67		GCCCGGGAATAGATTATAACGCA	0.99	1.4
18		GAAGCCATGATACTGTTCAGGGT	-1.5	-0.20	68		GCATTTTTAGTAATCCGAGCGCC	1.2	1.2
19		TATTCTACCAACGACATCACTGC	1.4	-0.39	69		CATGGATAAGTTTTCAAGCTGCG	1.5	1.5
20		CCATCAGTTATTCGGAGGGACTC	-1.4	0.59	70		GAGACAGGTAAACCCTCAGAGCA	-0.62	-1.1
21		CCATATCCGATTATTAGCGACGG	-1.2	-1.1	71		TAGCACCCGTTAAAACGGAAATG	-3.3	-3.3
22		CATCTCCAAGAATTGACCCACCA	-0.13	0.57	72		TATGTTTAGTTGTTGAACCGGCG	0.30	0.43
23	28	CCGTCGTGTTATTAAAGACCCCT	0.15	0.62	73		CGATCAGCTCTATTTCCCTCCCA	1.4	1.1
24		GAAGGATCGCTTTTATCTGGCAT	-0.30	0.19	74		AGTCAGTTAATCAGACGTGAGCA	0.13	0.15
25		CATTTGTCAGGTACAGTCCACTT	-0.35	-0.34	75		TGGCAATACAATAACGTATCGCG	-0.18	-2.7
26	0	GCCCACACTCTTACTTATCGACT	2.2	1.7	76		CGCAGTTTGCAAGAACGAACAAA	-0.78	-0.70
27		CGCTGTTACTGTAAGCGTACTAG	-2.5	-2.2	77		CGCGATAATTGATACCTACGGGC	0.37	-0.53
28		CGCGATTCCTATTGATTGATCCC	1.0	0.94	78		GGGGTGTGAGAGCTTTTTAGACG	1.1	2.4
29		CCGTCTGGGTTAAAGATTGCTAG	0.30	0.62	79		GGGATCCGTAACAAGTGTGTTAG	-1.9	-1.7
30		AGTCAGTCCAAATCTCAGGATGG	-0.11	-0.09	80		ACCACTATGATTGAGGAAACGCG	1.3	1.3
31		CGCCTAAATGAAACTCACTCTGC	0.57	0.57	81		CGTCTTTAGTATCAACCCTCCGC	1.2	1.4
32		GGGGTCAAACCAACAATTGATCT	-0.55	-0.23	82		GCATACGAACTTCTATATCGGCG	-0.37	-0.048
33		GCCCATTGATAGAATTACGAGGC	-0.40	-0.64	83		CCGTGTGTATGAGTATGACAGCA	0.62	0.61
34		ATGCCGTTGTCAAGAGTTATGGT	1.1	1.8	84		TGCTGTCTTCGTGTTTTACCTAG	1.1	1.2
35		TGCCGGCTATCGTAAGTATATGC	-0.057	0.027	85		CGATCATGTAAAGCTAACTCGCG	0.42	0.13
36		GCACCTCATACCTTCATAGAGCA	0.42	0.42	86		TGCCGTCATTTAAACGTAAGGGT	-0.29	-0.43
37		CGCGACATTTAGTCCAGGAGATG	-0.82	-0.82	87		TGGCAATTACAGTTGTTAACGCA	-0.14	0.55
38		CTAGTCCATTGTAACGAAGGCCA	1.3	1.3	88		GAGTCGAAGACCTCCTCCTACTC	-1.6	-1.8
39		AGACAATTAGAATCAGTGCCCCT	1.3	0.84	89		ATGCCAATATGTACTCGTGACTC	1.9	2.00
40	38	GCATTGAGGTATTGTTGCTCCCA	-0.048	-0.37	90		GCATATAGTGACGGTAAGGCGAA	0.36	0.85
41	33	CGAGAGTCTGTAATAGCCGATGC	0.25	0.031	91		GCCTCACTTGTAATAAGCGGGAC	0.16	-0.74
42		TGCCGTGATACTTAACTACGCTA	-0.19	-0.10	92		GTCCCAAAAGCTTCTTACGGACG	-1.00	-1.4
43	32	GAGTCCGCAAAAATATAGGAGGC	-0.083	0.38	93		CTAGGTACAACACCAACTGTCTC	-0.74	-0.83
44	34	GCCTCACATAACTGGAGAAACCT	0.08	0.17	94		TGCCGGTTATACCTTTAAGGACG	0.069	0.069
45		CGCCAATGACAATAAGTTGAGGC	-0.77	-0.71	95		GGCTGGTTAAATGTAAATCCGCG	0.3	0.59
46		CGCGATATAACATTAACCGAGGC	1.3	1.0	96		CGCGGTACTATTAGAAAGGGCTA	1.1	1.2
47		CACGCTTAGTTCCTACCTTAGGC	-0.85	-0.85	97		AGTCGCTTAATTACTCCGGATGG	0.39	0.62
48		CGCGTCGAATTACTTAATCACCA	0.70	0.13	98		TGCAGTGTAAGCAACTATTGTCT	-0.70	-0.70
49		GGGATAGGTATTATGCTCCAGCC	-0.18	-0.18	99		CGCGTTGTTTTGGACATATGACT	-0.49	-0.35
50		CGCCATTATACAACGGTTCATGC	0.84	-0.47	100		CCGTGCGTGAAAGTATTCGTTAG	-0.21	0.41

[a] Numbering in this chapter.
[b] Numbering in Chapter 2.

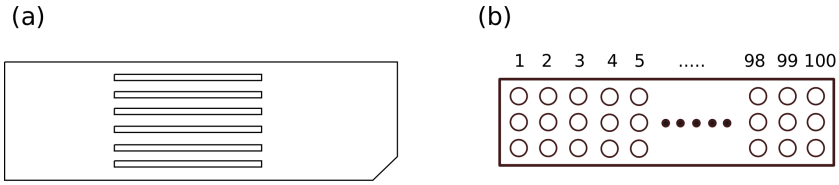

Figure 3.2.1: Design of the DNA microarray used in this study. (a) The plate design. A plate has six lanes, and in every lane, the same set of DNA strands with 100 distinct sequences were immobilized. (b) A close-up view on one of the lanes. Circles represent spots where DNA strands were immobilized. For each of the 100 sequences, the spot was triplicated, and so the total number of spots was $100 \times 3 = 300$ for each lane.

predictions for these sequences were performed using the same method described in Chapter 2 under the same conditions. Of total 100 sequences, 47 sequences were predicted to have only positive ΔG secondary structures for both of A and B strands.

3.2.2 Immobilization of oligonucleotides

DNA immobilization was performed commercially (Nihon Gaishi, Nagoya, Japan). Each oligonucleotide shown in Table 3.2.1 was labeled on the 5′-terminus with an amine. This amine can covalently bond to an active ester attached on the glass surface of plate. Solutions containing a single oligonucleotide species were placed on the surface using the ink-jet spotting method. Figure 3.2.1 shows the design of the DNA microarray plate.

3.2.3 Measurement of hybridization kinetics

Oligonucleotides whose 5'-terminus had been labeled with Cy5 fluorescent dye were purchased (BEX Co. Ltd., Tokyo, Japan). A total of 100 distinct oligonucleotides were mixed into a single solution with TE buffer (10 mM Tris·HCl, 1 mM EDTA, pH 8.0), and stored with protection from light at −30°C until use. In the mixed solution, the concentration of each oligonucleotide was adjusted to 1 μM using UV absorbance.

For measurements of hybridization, the stock solution of oligonucleotides was diluted with SSC (150 mM NaCl, 15 mM sodium citrate) buffer and equilibrated

at room temperature for at least 30 min prior to measurement. Subsequently, the solution was loaded by pipetting to the surface where oligonucleotides with the complementary sequences were immobilized. Six lanes of DNA arrays were spatially separated with a silicon plate cover. The strand concentrations in the solutions were 2, 10, 100, and 1000 nM. The total volume of loaded solution was 30 μL. The complementary DNA strands hybridized at room temperature (24–26°C). Total hybridization time, and the time points at which the fluorescence intensity was measured, are shown in Figures 3.3.2 and A19. After the hybridization, unhybridized free ssDNA strands were washed out at room temperature by washing with 1× SSC buffer for 2 min, followed by pure water for 2 min and by pure water for 2 min. The six microarray lanes on the plate provided fluorescence intensities at six different time points for every sequence. The fluorescence intensity was measured by a DNA microarray scanner (GenePix 4000B unit, Axon Instruments, Molecular Devices, USA) at automatically-adjusted photomultiplier voltages. Subsequently, the scanned images were analyzed by the software package GenePix Pro 5.1 (Axon Instruments). The local background-subtracted median intensities of Cy5 were used in further calculations. A signal-to-noise ratio (SNR) of 3 was employed as the cut-off value for the signal intensities of Cy5. The SNR was defined as follows: (signal-background)/(standard deviation of background).

Cross-hybridization between non-complementary sequences was evaluated as following by the previous work [51]. The oligonucleotide solution containing each sequence was loaded onto the microarray surface where the 100 distinct oligonucleotide sequences were immobilized. Sum of fluorescence intensities from spots of non-complementary sequences were at least 50 times lower than that of the complementary sequence for any sequence loaded (data not shown).

3.2.4 Hybridization rate determination

Hybridization of a free ssDNA strand (A strand) to the immobilized complement (B strand) was modeled as an irreversible second-order process: $A + B \xrightarrow{k} AB$, where k is a rate constant. From this model, the following rate equation was

derived:

$$\frac{\mathrm{d}C_{\mathrm{AB}}}{\mathrm{d}t} = kC_{\mathrm{A}}C_{\mathrm{B}}, \tag{3.1}$$

where C denotes the concentration of strands, and t is time. Solving the equation for C_{AB} yields:

$$C_{\mathrm{AB}} = \frac{1 - e^{-(C_{\mathrm{B0}} - C_{\mathrm{A0}})kt}}{\frac{1}{C_{\mathrm{A0}}} - \frac{1}{C_{\mathrm{B0}}}e^{-(C_{\mathrm{B0}} - C_{\mathrm{A0}})kt}}, \tag{3.2}$$

where C_{A0} (or C_{B0}) represents the initial concentration of A (or B) strands. The fluorescence intensity F is proportional to C_{AB}, and the maximum fluorescence intensity F_∞ is obtained when the hybridization process reaches equilibrium. When $C_{\mathrm{A0}} > C_{\mathrm{B0}}$, $C_{\mathrm{AB}}/C_{\mathrm{B0}} = F/F_\infty$. Substituting F for C_{AB} into Equation 3.2 yields:

$$F = F_\infty \frac{1 - e^{-(C_{\mathrm{B0}} - C_{\mathrm{A0}})kt}}{\frac{C_{\mathrm{B0}}}{C_{\mathrm{A0}}} - e^{-(C_{\mathrm{B0}} - C_{\mathrm{A0}})kt}}. \tag{3.3}$$

Furthermore, an approximation of $C_{\mathrm{A0}} \gg C_{\mathrm{B0}}$ yields:

$$F = F_\infty(1 - e^{-C_{\mathrm{A0}}kt}). \tag{3.4}$$

This equation was used for the determination of rate constant k, when the initial concentration of free strands, C_{A0}, is 1 μM.

On the other hand, when $C_{\mathrm{A0}} \ll C_{\mathrm{B0}}$, I arrived at the following:

$$F = F_\infty(1 - e^{-C_{\mathrm{B0}}kt}). \tag{3.5}$$

This equation was used for determination of the k values when $C_{\mathrm{A0}} = 2$ nM.

The rate constant, k, was determined by fitting the data obtained from two independent experiments using the non-linear least squares curve-fitting routines in SciPy (available at http://www.scipy.org/). In those fittings, F_∞ and $C_{\mathrm{A0}}k$ or ($C_{\mathrm{B0}}k$) were adjustable fitting parameters. Each time course of fluorescence intensity from the two experiments was normalized at the final time point, and coupled for each sequence. In the following, the determined value of k, arrived using the non-linear fit, is denoted by k_{app}.

When $C_{\mathrm{A0}} = 2$ nM, values of C_{B0} were estimated using F_∞ values supplied from the kinetic data at $C_{\mathrm{A0}} = 1\mu$M ($F_{\infty,1\,\mu\mathrm{M}}$). Thus, the rate constant of

hybridization at C_{A0} = 2 nM was denoted by k_{rel}, which was calculated to be the following:

$$k_{rel} = \frac{C_{B0}k}{F_{\infty,1\,\mu M}}. \tag{3.6}$$

By calculations using UNAFold, 99.9% of DNA strands exist as dsDNA rather than ssDNA under the experimental conditions ([Na^+]=0.195 M, 25°C, C_{A0} = 2 nM, and C_{B0} = 2 nM). Therefore, the dissociation (i.e, AB→A+B) was considered to be insubstantial.

3.3 Results

3.3.1 Amount of oligonucleotides immobilized on the surface

The amount of oligonucleotides immobilized on the surface has a crucial influence on the hybridization kinetics. Therefore, determination of the exact amount is essential for a proper analysis of the kinetic data. To assess the amount of immobilized strands (B strands), increasing concentrations of free strands (A strands) were hybridized to B strands up to saturation. This hybridization-based assessment provides the amount of B strands that are hybridization-capable. The 5′-termini of the A strands were labeled with a fluorescent dye, and the intensity of fluorescence emitted from the surface was proportional to the amount of hybrid duplexes, since free (unhybridized) A strands were washed out after the hybridization. To minimize experimental difficulties arising from difference in hybridization rate among the sequences, which could affect the amount of hybrid duplexes formed, ten fast-hybridizing sequences were selected and used for further experimentation. The hybridization of these sequences reached equilibrium at a lower A strand concentration (10 nM) than that of the previous hybridization (the A strand concentrations were 50, 100, and 500 nM) (Figure A18). Therefore, the hybridization for these ten sequences should reach equilibrium in the previous hybridization.

Figure 3.3.1a shows the fluorescence intensity changes upon addition of increasing concentrations of A strands (C_{A0}) for the ten sequences. The hybridization was monitored at room temperature for 60 min. The concentrations of A strands used were 50, 100, and 500 nM. In the plot, values of F_{100}/F_{500}, that

is the ratio of fluorescence intensities after hybridization with A strands at concentration of 100 nM and that of 500 nM, were almost unity for all ten sequences. This indicated that the amount of immobilized B strands is less than the amount of strands that exist in the solution at a strand concentration of 100 nM. On the other hand, values of F_{50}/F_{500} were lower than that of F_{100}/F_{500} by an average of 15%. This indicated that the amount of B strands is higher than that in solution at a strand concentration of 50 nM. Therefore, the values of C_{B0}, which is the amount of B strands represented as a concentration, were assessed for the ten sequences as the following: 50 nM $< C_{B0} <$ 100 nM.

For the remaining 90 sequences, values of C_{B0} were assessed using values of F_∞, those are the fluorescence intensities at the equilibrium of hybridization. Figure 3.3.1b shows the values of F_∞, which were obtained using data fitting with Equation 3.4 to the time course of hybridization with A strands at a concentration of 1 μM (data for some sequences are shown in Figure 3.3.2). Under conditions conducive to hybridization, all of the B strands hybridized to A strands, and therefore, the value of F_∞ is proportional to the value of C_{B0} for every sequence. Values of F_∞ were in the range of ±50% of the average value of the ten fast-hybridizing sequences selected above, with the exception of only seven sequences. Thus, values of C_{B0} for these 83 sequences were found to be at least within the range of 25 nM $< C_{B0} <$ 150 nM. For the other seven sequences, the C_{B0} value could be smaller than 25 nM. These seven sequences were omitted from the following data analysis.

3.3.2 Hybridization kinetics

DNA hybridization kinetics on the surface was measured using the microarray method. Figure 3.3.2 shows the hybridization kinetics of ten sequences. Hybridization kinetics was measured at different concentrations of A strands (C_{A0}). For the case of $C_{A0} = 1$ μM, the shape of kinetic trace differed significantly from sequence to sequence (Figure 3.3.2a). For example, the hybridization of sequence no. 73 showed a relatively high value of fluorescence intensity at the initial time point, while the hybridization of sequence no. 50 showed a large change in fluorescence intensity in a time range of 0–240 s. Also,

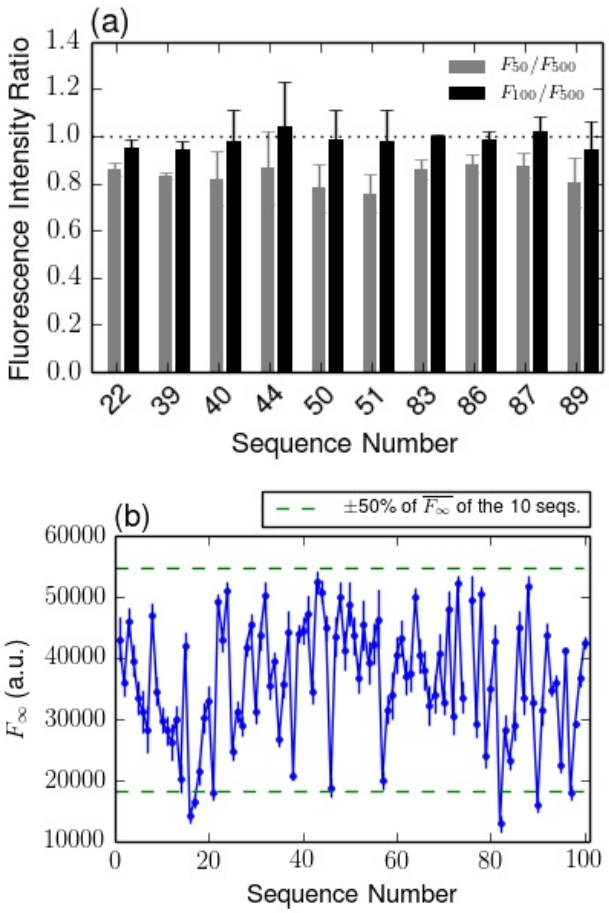

Figure 3.3.1: The amount of immobilized oligonucleotides. (a) The change of fluorescence intensity with increasing concentration of A strands. Values of F_{50}, F_{100}, and F_{500}, correspond to fluorescence intensities after A strands at concentrations of 50, 100, and 500 nM hybridized, respectively. Ten fast-hybridizing sequences were selected. Error bars were calculated from data of duplicated experiments (N=2). (b) The fluorescence intensity of hybridization at equilibrium (F_∞), which was obtained by fitting Equation 3.4 to the time course of hybridization with A strands at a concentration of 1 μM. Broken lines correspond to values of ±50% of the average fluorescence intensity of the ten sequences used in (a). Error bars represent standard-errors. For the sequence no.75, the F_∞ value was not determined due to a significant lack of fluorescence intensity.

the hybridization of sequence no. 17 showed only a relatively small change of fluorescence intensity over the measurement duration. This dependence of hybridization kinetics on the sequence was similar to that observed in the study of hybridization in solution (Chapter 2).

On the other hand, no sequence-dependent difference in hybridization kinetics could be clearly identified in the case of C_{A0} = 2 nM, and a similar kinetic trace was observed among the sequences (Figure 3.3.2b). The fluorescence intensity constantly increased over a time range of 0–180 min. The change of fluorescence intensity subsequently became small during the time range of 240–300, except for two sequences (nos. 66 and 17). These two sequences showed a constant increase in fluorescence intensity during the measurement. This indicated that the hybridization of those two sequences was slower than that of the other eight sequences.

The kinetic traces of these sequences for C_{A0} = 100 nM and 10 nM are shown in Figure A19. Sequence dependence was observed for both hybridization conditions, but the differences in the traces among relatively fast-hybridizing sequences tended to decrease from hybridization at C_{A0} = 100 nM to C_{A0} = 10 nM. Consequently, the hybridization kinetics at four different concentrations of A strands demonstrated that the sequence dependence weakens as the A strand concentration decreases.

3.3.3 Hybridization rate and secondary structure

Hybridization rate constants were determined from the obtained hybridization kinetic data. Values of the rate constant were plotted in Figure 3.3.3 against the number of predicted base-pairs in ssDNA strands (self-pairing). The number of base-pairs was calculated with consideration of all possible secondary structures formed by a ssDNA. For the case when C_{A0} = 1 μM (Figure 3.3.3a), the upper-bound of the hybridization rate constant k_{app}, arrived at by non-linear fitting, clearly decreased with the number of base-pairs, with the sole exception of sequence no. 88. The difference in k_{app} among sequences in the upper-bound was about two orders of magnitude. This decrease of k_{app} with the number of base-pairs in ssDNA, was the same trend as that observed in the study of

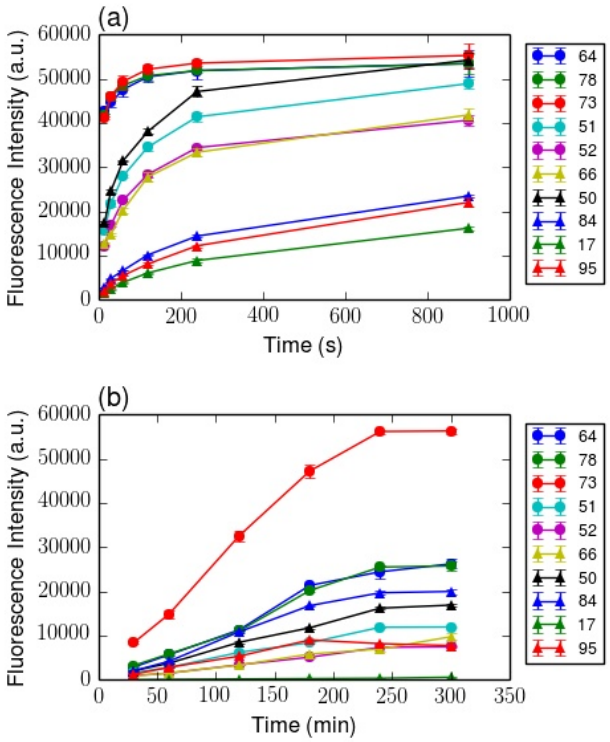

Figure 3.3.2: Observed surface hybridization kinetics. (a) Time courses of hybridization for ten sequences at an A strand concentration of 1 μM. Data of three fast-hybridizing sequences (nos. 64, 78, and 73), four moderate-hybridizing sequences (nos. 50, 51, 52, and 66), and three slow-hybridizing sequences (nos. 84, 17, and 95) are shown. The error bar represents the standard deviation, as calculated from fluorescence intensities of three array spots for each sequence. The fluorescence intensity was measured at 15, 30, 60, 120, 240, and 900 s. (b) Hybridization kinetics of the same sequence, but the when concentration of the A strands was 2 nM. The fluorescence intensity was measured at 30, 60, 120, 180, 240, and 300 min.

hybridization in solution (Chapter 2). This result indicates that ssDNA secondary structures also appear to delay hybridization on surfaces.

On the other hand, for the case of C_{A0} = 2 nM, no dependence of hybridization rate on the number of predicted base-pairs in the ssDNA was observed (Figure 3.3.3c, d). In the plots, the relative value of hybridization rate constant, k_{rel}, was shown, because C_{B0} values were estimated using F_{∞} values (for detail, see Materials and Methods). The data of 31 sequences were removed in the plots due to large standard errors for the determined rate constants (the coefficient of variation was larger than 10). The result shown in Figure 3.3.3c (and d) indicated that the influence of ssDNA secondary structures on hybridization rates on surfaces becomes insubstantial as the concentration of A strands becomes low. Thus, these results obtained from measurements of hybridization kinetics on a surface at different concentrations of A strands, indicated that the influence of ssDNA secondary structures on surface hybridization kinetics depends on the final concentration of the A strands.

The value of the hybridization rate constant, k_{app}, was determined using the non-linear fitting with Equation 3.5, which was derived under the assumption that the hybridization obey second-order kinetics. However, in the study of hybridization in solution (Chapter 2), some sequences with negative ΔG secondary structures did not obey second-order kinetics. Sequences with negative ΔG secondary structures were used in this study of hybridization on the surface (53 in total 100 sequences). For such sequences, the value of the hybridization rate constant may be improper. Therefore, I generated the same plot as Figure 3.3.3, but without those sequences with negative ΔG secondary structures (Figure A20). I found the same trend in the relationship between the hybridization rate constant and the number of predicted base-pairs in the ssDNA (self-pairing). Therefore, the relationship observed in Figure 3.3.3 is not substantially affected even though the surface hybridization of the sequences with negative ΔG secondary structures is unlikely to obey second-order kinetics.

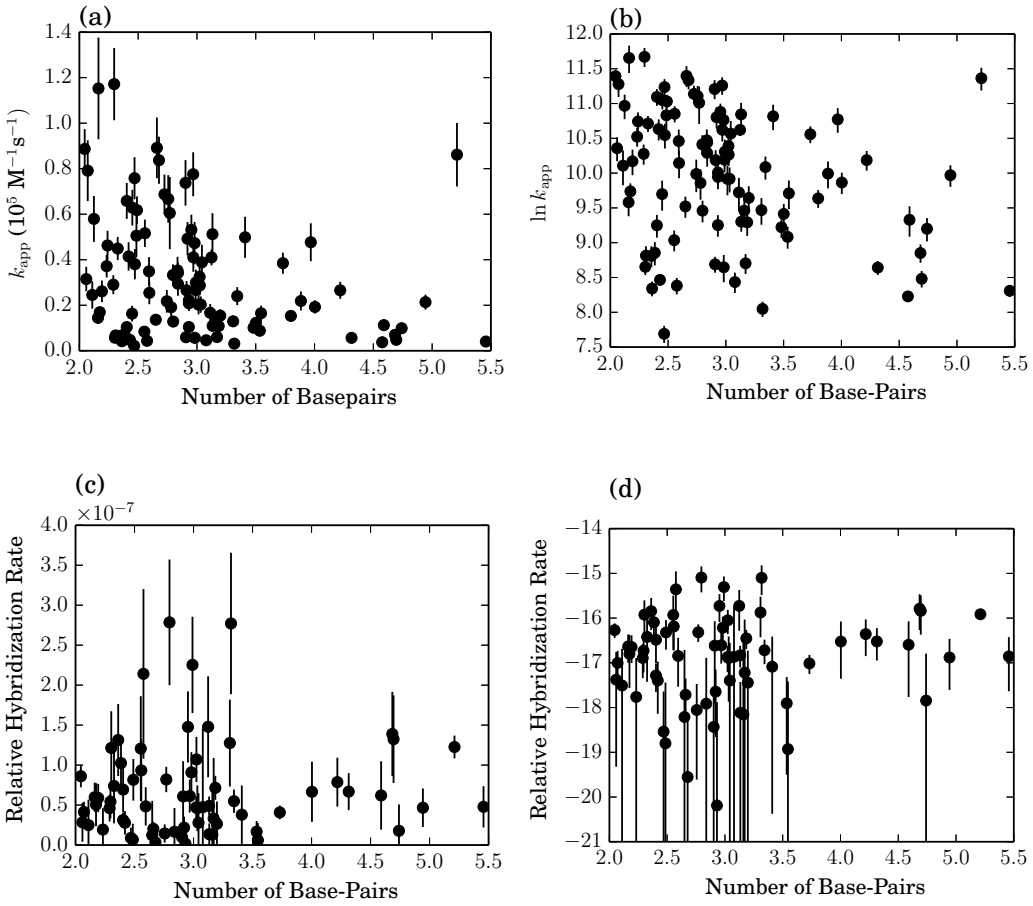

Figure 3.3.3: Comparison of the hybridization rate constant and the number of base-pairs in ssDNA strands. (a) The concentration of A strands was 1 μM. Error bars represent standard errors of the fitting parameter. The number of base-pairs was calculated for each ssDNA strand using secondary structure prediction, and averaged for each pair of strands. The correlation coefficient was -0.27 ($p = 0.006$). (b) The logarithms of hybridization rate constants were plotted against the number of base-pairs. The correlation coefficient was -0.28 ($p = 0.005$). (c) Same as in (a), but the A strand concentration used was 2 nM. The definition of the relative hybridization rate is described in the text. The correlation coefficient was 0.079 ($p = 0.5$). (d) The logarithms of hybridization rate constants at 2 nM A strand concentration were plotted against the number of base-pairs. The correlation coefficient was 0.11 ($p = 0.4$).

3.3.4 Comparison of the surface hybridization rate with the rate in solution

Similar to the study of hybridization kinetics in solution, an apparent influence of secondary structures on surface hybridization kinetics was observed when the concentration of the A strands was relatively high. There were nine sequences whose hybridization kinetics were measured both in solution and on the surface. Figure 3.3.4 shows a comparison of rate constants of surface and solution hybridization. No clear correlation was found between these two rate constants (Figure 3.3.4a) of the nine sequences compared. Hybridization rate constants of two sequences showed relatively large differences between the two hybridization environments. These two sequences (nos. 44 and 26) were found to have either the most or the least number of predicted base-pairs in ssDNA, respectively, among these nine sequences (Figure 3.3.4b). This suggests that the effect of ssDNA secondary structures on hybridization kinetics differs between the two hybridization environments. However, more number of sequences need to be investigated the differences in hybridization rates in the two environments, which may indicate any relationships between them.

3.4 Discussion

3.4.1 Concentration-dependent influence of ssDNA secondary structure on hybridization kinetics

In this chapter, the influence of thermodynamically unfavorable secondary structures on hybridization kinetics on a solid surface was investigated. Using the microarray method, hybridization kinetics on the surface was measured for one hundred distinct DNA sequences. A similar trend to that seen for hybridization in solution is observed when the concentration of A (free) strands is relatively high; the hybridization rate constant tend to decrease with the number of predicted base-pairs in the ssDNA (Figure 3.3.3a). On the other hand, when the concentration of A strands is relatively low, this trend is not observed (Figure 3.3.3b). In order to understand the mechanism causing the concentration dependence of influence arising from secondary structure, here, I propose a

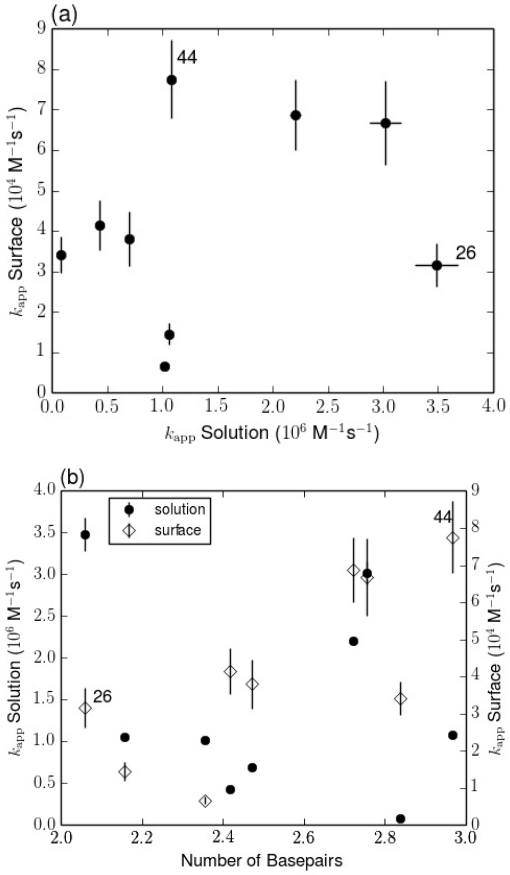

Figure 3.3.4: Comparison of rate constants of surface and solution hybridization. (a) A plot of k_{app} values of hybridization on the surface versus that of hybridization in solution. For hybridization on the surface, the concentration of A strands was 1 μM. There are nine sequences depicted, corresponding to those whose hybridization kinetics were measured both in solution and on the surface. (b) Comparison of the secondary structure dependence of k_{app} values of hybridization in solution and that of hybridization on the surface.

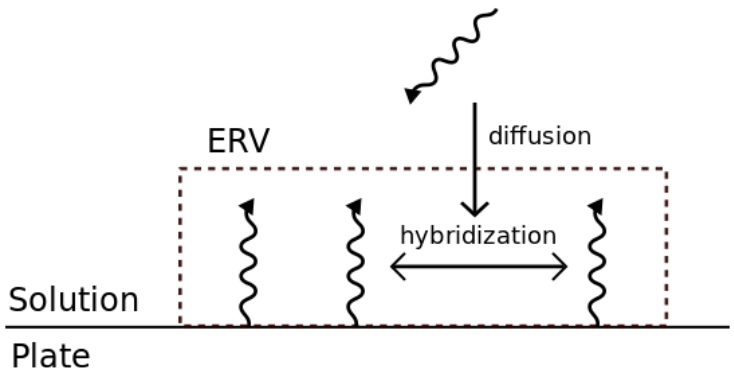

Figure 3.4.1: Schematic of the ERV and hybridization on the surface. Triangle-headed arrows represent ssDNA. The ERV is denoted by the broken line.

reaction model for the surface hybridization of ssDNA sequences with positive ΔG secondary structures.

According to previous research, in the case of hybridization on the surface, an effective reaction volume (ERV) can be adopted [50]. The ERV is defined as the product of the reactive surface area of the plate and the estimated height of DNA probes coupled to the surface. The ERV is schematically represented in Figure 3.4.1. When viewed through the lens of the ERV, hybridization on the surface can be divided into two processes; (i) diffusion from bulk solution to the ERV, (ii) hybridization of A and B strand within the ERV. The rate-limiting step of hybridization is considered to depend on these two processes, when the temperature and the concentration of A is held constant. When the A strand concentration is high, A strands can come into the ERV within a shorter time than that required for hybridization within the ERV. Therefore, the rate-limiting process is the latter process. In this case, the hybridization would be nucleation-limited, the same with hybridization in solution, and thus, the effect on the hybridization rate from ssDNA secondary structure may mainly be due to a reduction in available nucleation-sites. On the other hand, when the A strand concentration is low, it takes a longer time for A strand to come into the ERV. Therefore, the rate-limiting step switches to the diffusion-dependent process of encountering the ERV, and thus, the hybridization rate is not limited by secondary structures with the ssDNA.

3.4.2 Calculation of hybridization rate from the sequence of bases

Using the previously developed method and the parameters determined during the study of hybridization kinetics in solution, the hybridization rate constant k_{calc} was calculated from the sequence of bases. However, the obtained correlation coefficient (0.30) between the values of k_{calc} and k_{app} was much lower than that obtained in the solution study (0.92). The correlation plot is shown in Figure A21. This indicated that the hybridization model used for positive ΔG secondary structures is not applicable to hybridization kinetics on the surface. This could be because the influence of secondary structures differs between these two hybridization environments, as suggested in Section 3.3.4. Two factors that might possibly account for the decreased accuracy of calculation of hybridization rate on the surface are: (i) non-specific adsorption of B strands, and (ii) the fluorescent-labeling of A strands. The former factor has been known to decrease the amount of B strand which can hybridize to the complement (A strand) [52]. It is possible that the adsorption is affected by secondary structures in the ssDNA. As for the latter factor, labeling with Cy5-fluorescent dye has been known to increase the stability of base-pairing at the terminus [53]. When such differences between hybridization environments have been taken into account, the hybridization model for ssDNA molecules with positive ΔG structures remains applicable to hybridization on the surface.

Chapter 4

Effects of Single-Strand Base-Stacking on Hybridization Rates

4.1 Introduction

Single-strand base-stacking (SSBS) refers to intra-strand stacking between bases found in a single-stranded nucleic acid molecule. The SSBS changes the conformational stability of the nucleic acid molecule, which is often quantified by the change in Gibbs free energy (ΔG) [54, 55]. The stability change in the structure of the single-stranded nucleic acid that is due to the SSBS, can be calculated from the primary structure of the nucleic acid (i.e., the sequence of bases), because most SSBS occurs between two adjacent bases in the molecule. Therefore, compared with the stability change arising from the secondary structure by intramolecular base-pairing, the change due to SSBS is more fundamental. Therefore, the study of relationship between SSBS and hybridization kinetics should provide more fundamental insights into the influence of single-stranded structures on hybridization kinetics.

SSBS has been studied since the 1960s [7]. The formation and dissolution of SSBS has been monitored using UV absorbance spectroscopy [56], circular dichroism spectroscopy [57], fluorescence spectroscopy [58], and calorimetry [45]. These studies reported various properties for the formation/dissolution of SSBS, such as the time-scale of occurrence [58], the non-cooperative behavior [7], and the base-sequence dependence [45].

Alongside these experimental studies, SSBS has also been theoretically investigated since the 1970s [59], using molecular dynamics (MD) simulations [60], Monte Carlo simulations [61], and quantum chemical calculations [62]. These earlier studies succeeded in explaining the SSBS behavior observed experimentally at the dinucleotide level [63]. However, the simulation time was shorter (a few nanoseconds [63]) than the true time scale of the physical occurrence of formation/dissolution of SSBS (tens of nanoseconds [58]). Therefore, direct comparisons between results obtained from experiments and simulations were not reliable. Also, longer nucleic acid molecules (i.e., trimers or tetramers of nucleotides) had not been studied at a nanosecond time-scale due to the lack of computational power.

More recently, since the development of more powerful computational resources and the improvement of empirical force fields, researchers were enabled to perform microsecond-long simulations of several nucleotide tetramers [64]. Such simulations have provided results that can be directly compared with experimental results. Such comparisons are essential for further development of the theoretical study of SSBS, by which novel insights may be provided into the mechanisms causing the SSBS properties observed experimentally.

The influence of SSBS on the hybridization kinetics of nucleotide trimers was suggested, for the first time, in my own experimental study of hybridization kinetics in solution (Chapter 2). In this study, the strength of the SSBS influence was considered to be dependent on the sequence, which was quantified using the δ_{seq} parameters. The values of δ_{seq} parameters represent nucleation capabilities of nucleation sites, which are composed of three nucleotides.

In this chapter, in order to gain further insights into the influence of SSBS on hybridization kinetics, I modeled the sequence-dependence of SSBS stability using MD simulations. The SSBS stability of the nucleotide tetramers was studied for all 256 possible sequences. From the MD trajectories of these tetramers, the SSBS stabilities of trimers were extracted by averaging the stabilities of eight tetramers including each trimer. This method, inspired by the work of Dixit et al. [65], includes the influence of adjacent nucleotides to a trimer on the SSBS stability of that trimer, and the influence must be found in the SSBS of trimers involved in 23-mer nucleotides used in Chapter 2. The SSBS

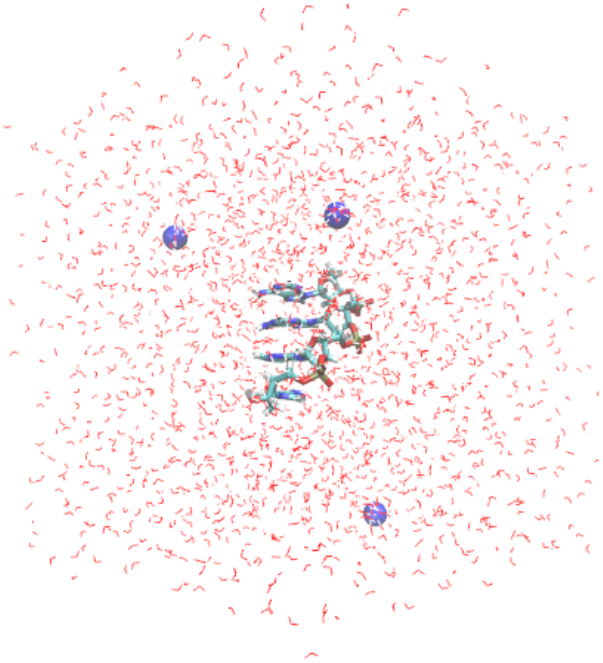

Figure 4.2.1: An image of the simulated system. A DNA segment (dApdApdApdA) is depicted at center. Bonds between atoms are represented by bars. Blue spheres represent Na^+ ions. Red lines represent water molecules.

stabilities obtained for trimers were suitable for informed comparison to the value of the δ_{seq} parameter.

4.2 Methods

4.2.1 The simulated system

A molecular model consisting of four deoxynucleosides linked by three bridging phosphodiester groups, was dropped into explicit water (TIP3P [66]) solvent with 3 Na^+ ions. A picture of the system is shown in Figure 4.2.1. Periodic boundary conditions were applied to a $40 \times 40 \times 40$ Å box. A total 256 of base-sequences were subjected to MD simulation. Initial structures of nucleotide tetramers were generated from the crystallographic structure of B-DNA using the

Stroud group's Make-NA server (http://structure.usc.edu/make-na/server.html).

4.2.2 Simulation conditions

Simulation conditions followed the approach used in previous modeling studies [54, 55]. All MD simulations were performed using GROMACS version 4.6.2. [67, 68]. The AMBER parm99 force field force [69] supplemented with the bsc0 parameters [70] that improve modeling of α and γ torsions and with improved parameters for the glycosidic torsions [71]. All systems were equilibrated at 1 bar and 300 K by 1 ns of NVT equilibration, followed by 1 ns of NPT equilibration. After the equilibration step, MD simulations were carried out for 100 ns under constant temperature and pressure using the Nosé-Hoover thermostat [72, 73] and Parrinello-Rahman barostat [74], respectively. The cutoff for short-range van der Waals and electrostatic interactions was 10 Å. The Particle Mesh Ewald (PME) method [75] was used for computation of longer-range electrostatic interactions. During the MD simulation, all solute coordinates were saved every 2 ps, giving a total of fifty-thousand snapshots from each simulation for subsequent analysis.

4.2.3 Criteria for stacking

The stacking parameter, ξ, which was defined by Jafilan et al. [54], was adopted; and it was used as the criterion for assessing the stacked conformation of two adjacent bases. This criterion considers three factors for base stacking: the distance between bases, the angle between base planes, and the spatial overlap of these bases. The value of ξ is calculated by the following formula:

$$\xi = \frac{R_{\mathrm{M}}}{S(\alpha)}, \tag{4.1}$$

where R_{M} is a distance between the center of masses of two bases, α is an angle (in radians) between base planes. An angular term $S(\alpha)$ is defined by the following:

$$S(\alpha) = e^{-\alpha^4} + e^{-(\alpha-\pi)^4} + 0.1e^{-(\alpha-0.5\pi)^4}. \tag{4.2}$$

The values of S are plotted as a function of α in Figure A22. This angular term eliminates T-shaped complexes and enforces near coplanarity of the stacked bases. With $\xi < 6.4$, the conformation was classified as having a stacked conformation. For all values greater than this, the bases were classified as adopting an unstacked conformation.

4.3 Results

4.3.1 Stacking dynamics

The MD trajectory was obtained for nucleotide tetramers (e.g., 5′-dApdGpdTpdC-3′) with all possible ($4^4 = 256$) base sequences. Figure 4.3.1 shows an example of the trajectories, which were obtained from the simulation for dApdCpdApdC. Transitions between the stacked and the unstacked conformation were observed for every pair of adjacent bases. Most of stacked conformations were found in the ξ range of 3.2–6. This was the same range reported in the study of nucleotide dimers [54]. This result indicated that extension from dimer to tetramer does not substantially affect the stacked conformation of each pair of adjacent bases. Therefore, the criterion to assess the presence of stacked conformations that was previous used in the study of nucleotide dimers, was found to be transferable to this study of tetramers.

Based on the definition of the ξ criterion, dissolution of the stacked conformation occurs due to two factors: (i) an increase in the angle between base planes, and (ii) an increase in the distance between bases. From snap-shots of MD conformations, sharp peaks in the trajectories (Figure 4.3.1) were mainly due to changes in the angle between base-pairs. In these peaks, both dissolution and re-formation occurred within a few nanoseconds. These peaks tended to appear more frequently in bases at termini than in those at the center of the DNA strand (Table B3), probably because motions of the terminal bases are less restricted by adjacent bases than those at the center of the DNA strand. On the other hand, unstacked conformations lasting for tens of nanoseconds were also found. These conformations arose mainly due to increased separation distances between adjacent base-pairs.

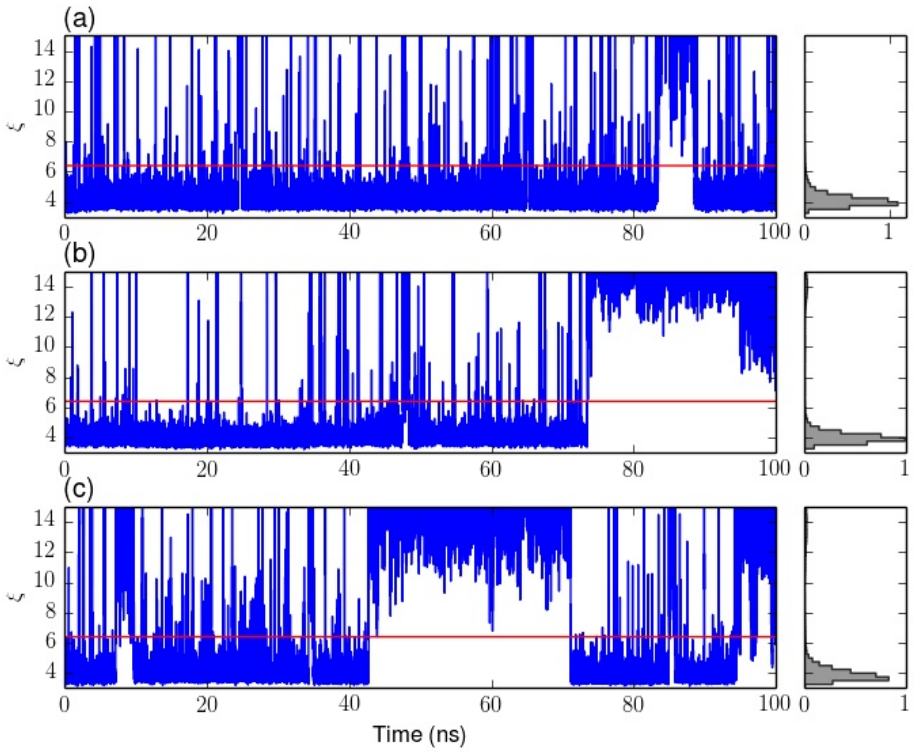

Figure 4.3.1: Example MD trajectories. The sequence was ACAC. (a) The time course of ξ (the stacking coordinate used in this study) for two bases from the $5'$-terminus (i.e., AC). The red line indicates the threshold between the stacked and the unstacked conformation ($\xi = 6.4$). The normalized distribution of conformations is shown to the right of the panel. (b, c) Same as above, but for the two bases at the center (CA), or at the $3'$-terminus (AC), respectively.

4.3.2 Sequence dependence

To evaluate the stability of SSBS of each nucleotide tetramer, the stacking number, which represents the number of adjacent bases in the stacked conformation, was counted in each tetramer. For example, the stacking number is '2' in the case of A=TG=C, where "=" denotes that AT and GC are in the stacked conformation. The stacking number was counted for all conformations found in all snapshots, and averaged for the entire simulation of each tetramer.

Figure 4.3.2 depicts the stacking number of all 256 tetramers. The stacking number was significantly dependent on the base sequence. The difference among sequences was 2.7 at a maximum. To understand basis for the sequence dependence, the stacking number was compared to the number of adjacent purine bases, which are known to stack more stably than the other configurations (i.e., purine-pyrimidine and pyrimidine-pyrimidine stacking) from the study of nucleotide dimers [54, 55]. However, there was no correlation between them (Figure A23). This result indicates that the predicted stability of SSBS within the tetramer is not determined by the stabilities of dimeric pairs in the tetramer.

4.3.3 Comparison with hybridization kinetics

Here, the sequence dependence of SSBS stability was compared with the value of the δ_{seq} parameter, which was determined in the study of hybridization in solution (Chapter 2). The stability of SSBS for nucleotide trimers was calculated by averaging the number of stacking of eight tetramers including each trimer, and compared with the value of δ_{seq}.

Figure 4.3.3 shows the plot of δ_{seq} values versus the stacking numbers. The lower-bound (Figure 4.3.3, broken lines) of the stacking numbers tended to rise as δ_{seq} values increase. This result was compatible with the discussion in Chapter 2 that the value of δ_{seq} is related to the decrease in entropy of the ssDNA strand due to the SSBS. In addition, the stacking number for trimers was also calculated from a direct simulation of nucleotide trimers. In this case, the correlation described above was not observed (Figure A24). Therefore, simulations of longer nucleotides (e.g., pentamers or hexamers) would improve

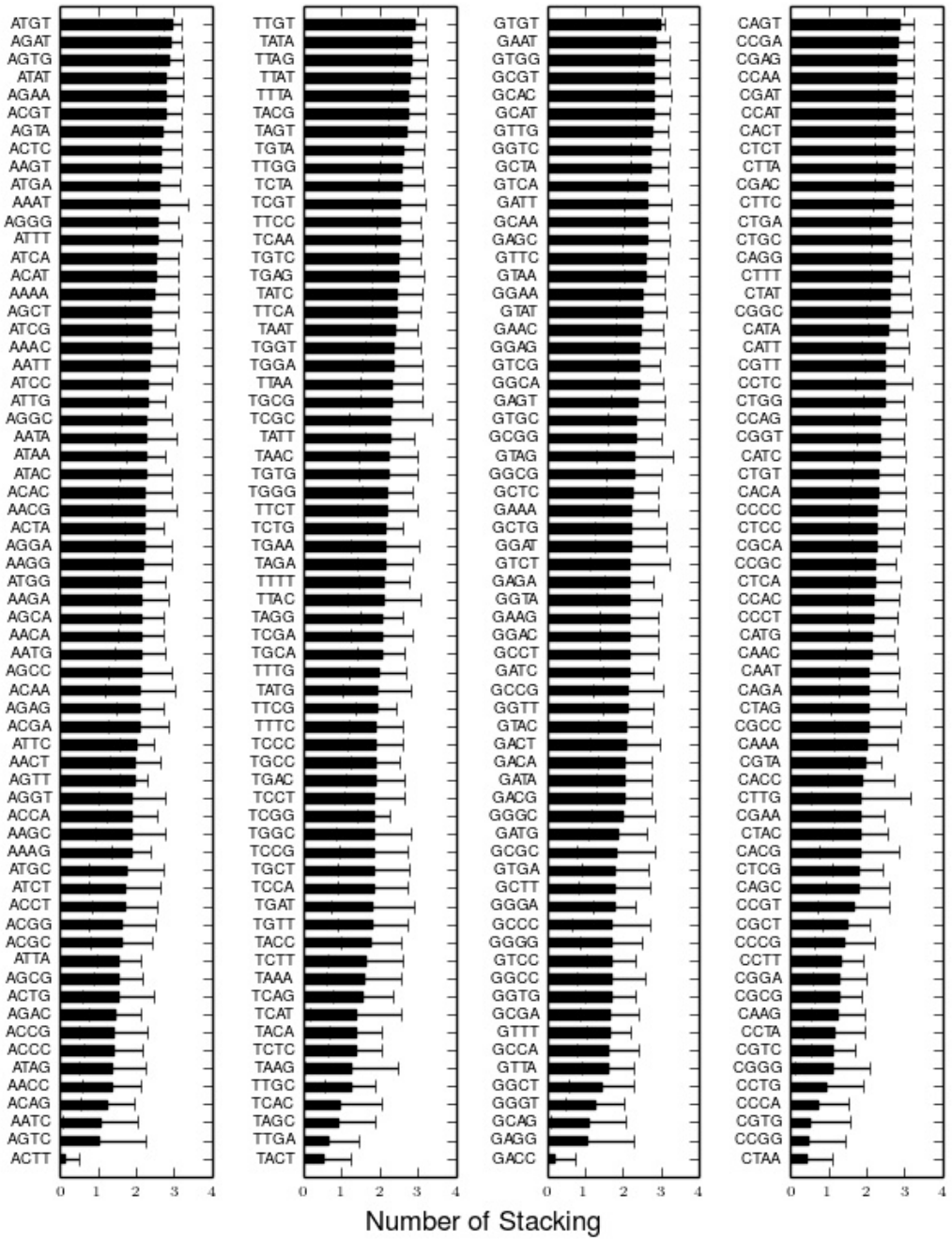

Figure 4.3.2: The stacking number for all 256 nucleotide tetramers, which was averaged over the entire simulation. Error bars represent standard deviations.

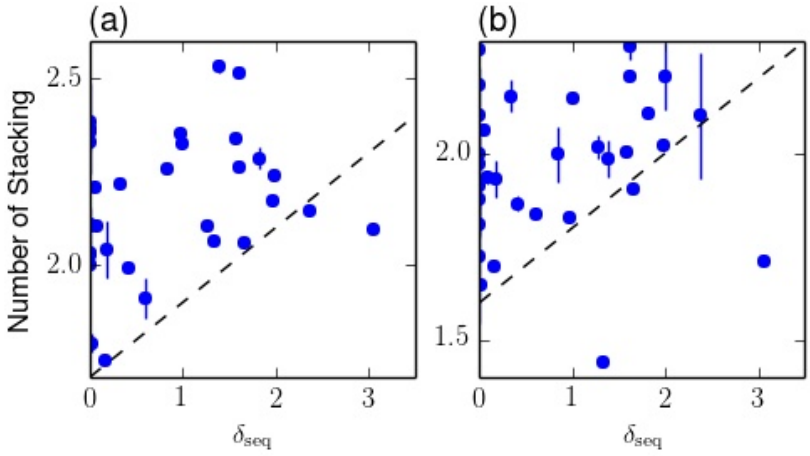

Figure 4.3.3: A comparison between the value of δ_{seq} and the stacking number. (a) Data for trimers having the larger stacking number between two complementary trimers. The value of δ_{seq} was determined for every pair of complementary base triplets in the study of hybridization in solution (Chapter 2). The stacking number was calculated for every nucleotide trimer by averaging the stacking number of eight tetramers including each trimer. The broken line, whose slope is 0.2, is simply included as a guide to the eye. (b) Data for the other trimers.

the correlation, because most trimers found in such nucleotides are connected to additional nucleotides at both terminals. If possible, simulations for base-stacking and base-pairing of 23-mer nucleotides should be performed, which would provide dynamics of base interactions for direct comparisons with the experimental results described in Chapter 2.

4.4 Discussion

4.4.1 Single-strand base-stacking and hybridization kinetics

In this study, SSBS (single-strand base-stacking) in DNA was investigated using MD simulations, from the point of view of the influence on hybridization kinetics. The simulation method and the criterion for evaluating stacked conformations followed those adopted in previous *in silico* models of the SSBS process. I accomplished the MD simulation of the nucleotide tetramers (all 256 possible base sequences) in explicit solvent, for as long as 100 ns using

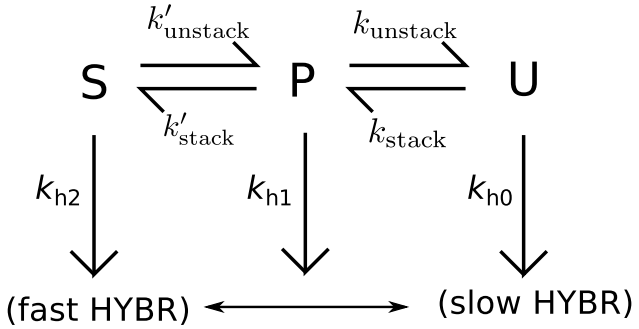

Figure 4.4.1: A schematic picture of the stacking-hybridization model. The completely stacked, and the completely unstacked conformation of a nucleotide trimer are represented by S and U, respectively. Partially stacked conformations are denoted by P. Rates of stacking and unstacking between adjacent bases are denoted by k_{stack} and k_{unstack}, respectively. The rate of hybridization is denoted by $k_{\mathrm{h}i}$, where i denotes the stacking number in the corresponding conformation.

state-of-the-art algorithms and computer systems.

The stacking number averaged over 100 ns was significantly dependent on the sequence (Figure 4.3.2). The stacking number for trimers was calculated from that for tetramers, which correlated with δ_{seq} values to a small extent (Figure 4.3.3). To describe the mechanism causing the influence of SSBS on hybridization kinetics, I proposed the multistate kinetic model shown in Figure 4.4.1. In a certain length of an oligonucleotide strand, a nucleotide trimer can adopt an unstacked conformation (U), a totally stacked conformation (S), and two partially stacked conformations (P). Transitions between these conformations occur through stacking/unstacking with rates depending not only on the bases involved in the stacking, but also on other bases in the strand. These transition rates determine, in part, the distribution of conformations.

From any conformation of S, P, and U, hybridization with the complementary trimer can occur. However, the rate of hybridization depends on the conformation of the trimer. For the conformation S, the entropy loss with hybridization should be relatively small, because base-stackings are formed both before and after hybridization. Therefore, the hybridization proceeds rapidly. On the other hand, for the conformation U, the entropy loss becomes larger, because base-stackings are not formed before the hybridization. Therefore, hybridization

is predicted to be less statistically favorable, and so to proceed more slowly. Thus, the following relationship is obtained among these hybridization rates: $k_{h2} > k_{h1} > k_{h0}$. Consequently, the apparent hybridization rate is predicted to increase with the concentration of conformation S.

4.4.2 Sampling adequacy

Owing to recent improvements in the computational force fields used to simulate highly charged systems, such as DNA, stable MD simulations are now available for long as ~one microsecond [55]. In this study, I simulated the behavior of nucleotide tetramers for 100 ns. However, this simulation time does not ensure that the MD trajectory covers the whole conformational landscape occupied by a nucleotide tetramer under the simulation conditions. It is possible that a longer simulation may find a more stable unstacked conformation, which would change the stability of the stacked conformation. This possibility had been considered in the previous work on SSBS in nucleotide dimers [54]. Most recently, a simulation of SSBS in dimers was performed [55] with a simulation time (1 μs) 25 times longer than that of the previous study (40 ns) [54]. This study reported the sequence dependence of SSBS stability, which is compatible with that reported in [54]. Therefore, for the simulation of tetrameric nucleotides, the sequence dependence of SSBS stability would not be expected to be substantially affected by the simulation time. However, more longer simulations need to be performed to confirm the above speculation, and to investigate the conformational landscape more accurately.

An experimental evaluation of the stability of SSBS was discussed in Appendix Chapter D. The SSBS of two tetranucleotides were investigated using UV absorbance. The absorbance change from low temperatures to high temperatures was observed, which has been known to be due to the destruction of SSBS. The absorbance change showed a sequence dependence to some extent, but the absolute amount of absorbance change was relatively small. Moreover, in order to quantitatively evaluate the relationship in the SSBS stabilities among base sequences, a systematic investigation to distinguish SSBS from intermolecular base-stacking and base-pair-induced base-stacking, would be

necessary.

Chapter 5

General Discussion and Conclusion

In this thesis, I described the influence of ssDNA structures on hybridization kinetics. The structures of ssDNA structures investigated, included thermodynamically unfavorable (positive ΔG) secondary structures and single-strand base-stacking (SSBS) events. The influence of these structures on hybridization kinetics was studied here for the first time.

Aiming to develop a method for quantitative prediction of hybridization rate from base sequence alone, the influence of these structures, especially on the rate-limiting step of hybridization, were studied here. The rate-limiting step was found to be the nucleation process for hybridization in solution with these ssDNA structures. The obstruction of nucleation was attributed to the influence of positive ΔG secondary structures, which were considered to reduce nucleation (or binding) sites by forming intramolecular base-pairs. SSBS was suggested to assist nucleation by decreasing entropic cost of nucleation.

For hybridization on surfaces, the rate-limiting step was suggested to depend on nucleation within the ERV on the surface, and the diffusion from bulk solution to the surface. For the diffusion-limited hybridization, values of the hybridization rate constant were not related to the stabilities of positive ΔG secondary structures. Therefore, the positive ΔG secondary structure would not be expected to have any substantial influence on the diffusion process.

The ssDNA structures can be classified in terms of their thermodynamic stability. The value of the Gibbs free energy change (ΔG) is associated with the favorability of structure formation. Two structures described in this thesis can

have ΔG values higher than ~ -2 kcal/mol, which is different from those of previous studies (lower than ~ -2 kcal/mol [11, 28]). Such unstable conformations are generally difficult to study experimentally, therefore *in silico* modeling can be called on to provide useful insights. The MD simulation of SSBS showed a significant sequence dependence of the stability of SSBS formation, which supported the mechanism described in this thesis for explaining the difference in hybridization rate among sequences having a similar stability of positive ΔG secondary structure.

This thesis clearly indicates the possibility that the kinetics of interactions between ssDNA strands is affected by such thermodynamically unfavorable ssDNA structures. Therefore, consideration of these ssDNA structures may permit the development of a method for quantitatively predicting kinetic rates of oligonucleotide hybridization. By helping researchers to avoid unfavorale nucleotide sequences, such predictions will improve the reliability and efficiency of a range of nucleotide-dependent technologies, including PCR, DNA microarrays, and DNA origami. In addition, for applications dependent on finely-tuned hybridization rates, such prediction would allow researchers to design nucleotides with control over the hybridization efficiency, exquisitely modulating the rate of interaction. Also, my observations for deoxyribo-oligonucleotides will almost certainly be applicable to hybridization of ribo-oligonucleotides. Therefore, this study has further implications for understanding the mechanisms of RNA-RNA interactions, such as RNA interference in gene expression.

Bibliography

[1] Wetmur, J. G. and Fresco, J. (1991) *Critical Reviews in Biochemistry and Molecular Biology* **26(3-4)**, 227–259.

[2] Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988) *Science* **239(4839)**, 487–491.

[3] Hutchison, C. a., Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., and Smith, M. (1978) *Journal of Biological Chemistry* **253(18)**, 6551–6560.

[4] Watson, J. D. and Crick, F. H. C. (1953) *Nature* **171(4356)**, 737–738.

[5] Mandel, M. and Marmur, J. (1968) *Methods in Enzymology* **12**, 195–206.

[6] Craig, M. E., Crothers, D. M., and Doty, P. (1971) *Journal of Molecular Biology* **62**, 383–401.

[7] Eigen, M. and Pörschke, D. (1970) *Journal of Molecular Biology* **53(1)**, 123–141.

[8] Gotoh, O. and Tagashira, Y. (1981) *Biopolymers* **20(5)**, 1033–1042.

[9] SantaLucia, J. (1998) *Proceedings of the National Academy of Sciences of the United States of America* **95(4)**, 1460–1465.

[10] Sekar, M. M. a., Bloch, W., and St John, P. M. (2005) *Nucleic Acids Research* **33(1)**, 366–375.

[11] Gao, Y., Wolf, L. K., Georgiadis, R. M. (2006) *Nucleic Acids Research* **34(11)**, 3370–3377.

[12] Doktycz, M. J., Goldstein, R. F., Paner, T. M., Gallo, F. J., and Benight, A. S. (1992) *Biopolymers* **32(7)**, 849–864.

[13] Manning, G. S. (1978) *Quarterly Reviews of Biophysics* **11(02)**, 179–246.

[14] Xia, T., SantaLucia, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998) *Biochemistry* **37(42)**, 14719–14735.

[15] Wetmur, J. G. and Davidson, N. (1968) *Journal of Molecular Biology* **31**, 349–370.

[16] Bloomfield, V. A., Crothers, D. M., and Tinoco, I. (2000) Chaper 8 In Nucleic Acids: Structures, Properties, and Functions chapter 8 University Science Books Sausalito, CA.

[17] Zhang, D. Y. and Winfree, E. (2009) *Journal of the American Chemical Society* **131(47)**, 17303–17314.

[18] Rauzan, B., McMichael, E., Cave, R., Sevcik, L. R., Ostrosky, K., Whitman, E., Stegemann, R., Sinclair, A. L., Serra, M. J., and Deckert, A. A. (2013) *Biochemistry* **52(5)**, 765–772.

[19] Sikora, J. R., Rauzan, B., Stegemann, R., and Deckert, A. (2013) *The Journal of Physical Chemistry B* **117**, 8966–8976.

[20] Dimitrov, R. a. and Zuker, M. (2004) *Biophysical Journal* **87(1)**, 215–226.

[21] McCaskill, J. S. (1990) *Biopolymers* **29(6-7)**, 1105–1119.

[22] Yin, Y. and Zhao, X. S. (2011) *Accounts of Chemical Research* **44(11)**, 1172–1181.

[23] Ouldridge, T. E., Šulc, P., Romano, F., Doye, J. P. K., and Louis, A. A. (2013) *Nucleic Acids Research* **41**, 8886–8895.

[24] Gerling, T., Wagenbauer, K. F., Neuner, A. M., and Dietz, H. (2015) *Science* **347(6229)**, 1446–1452.

[25] Amir, Y., Ben-Ishay, E., Levner, D., Ittah, S., Abu-Horowitz, A., and Bachelet, I. may 2014 *Nature Nanotechnology* **9(5)**, 353–357.

[26] Geary, C., Rothemund, P. W. K., and Andersen, E. S. (2014) *Science* **345(6198)**, 799–804.

[27] Shi, H. and Moore, P. B. (2000) *RNA* **6(8)**, 1091–1105.

[28] Chen, C., Wang, W., Wang, Z., Wei, F., and Zhao, X. S. (2007) *Nucleic Acids Research* **35(9)**, 2875–2884.

[29] Kitajima, T., Takinoue, M., Shohda, K., and Suyama, A. (2008) *Lecture Notes in Computer Science* **4848**, 119–129.

[30] Yin, Y. and Zhao, X. S. (2011) *Accounts of Chemical Research* **44(11)**, 1172–1181.

[31] Sobczak, J.-P. J., Martin, T. G., Gerling, T., and Dietz, H. (2012) *Science* **338(6113)**, 1458–1461.

[32] Seelig, G., Soloveichik, D., Zhang, D. Y., and Winfree, E. (2006) *Science* **314(5805)**, 1585–1588.

[33] Myhrvold, C., Dai, M., Silver, P. a., and Yin, P. (2013) *Nano Letters* **13(9)**, 4242–4248.

[34] Qian, L. and Winfree, E. (2011) *Science* **332(6034)**, 1196–1201.

[35] Genot, A. J., Zhang, D. Y., Bath, J., and Turberfield, A. J. (2011) *Journal of the American Chemical Society* **133(7)**, 2177–2182.

[36] Gotoh, O., Murakami, Y., and Suyama, A. (2011) *Nucleic Acids Research* **39(10)**, e70.

[37] Markham, N. R. and Zuker, M. (2008) UNAFold: Software for nucleic acid folding and hybridization volume **453**, of Methods in Molecular Biology chapter 1, pp. 3–31 Humana Press Totowa, NJ.

[38] Zuker, M., Mathews, D. H., and Turner, D. H. (1999) *NATO Science Series* **70**, 11–44.

[39] Singer, V. L., Jones, L. J., Yue, S. T., and Haugland, R. P. (1997) *Analytical Biochemistry* **249(2)**, 228–238.

[40] Dragan, A. I., Casas-Finet, J. R., Bishop, E. S., Strouse, R. J., Schenerman, M. A., and Geddes, C. D. (2010) *Biophysical Journal* **99(9)**, 3010–3019.

[41] McGhee, J. (1976) *Biopolymers* **15**, 1345–1375.

[42] Porschke, D., Uhlenbeck, O. C., and Martin, F. H. (1973) *Biopolymers* **12(6)**, 1313–1335.

[43] Morrison, L. E. and Stols, L. M. (1993) *Biochemistry* **32(12)**, 3095–3104.

[44] Seol, Y., Skinner, G., Visscher, K., Buhot, A., and Halperin, A. (2007) *Physical Review Letters* **98(15)**, 158103.

[45] Ramprakash, J., Lang, B., and Schwarz, F. P. (2008) *Biopolymers* **89(11)**, 969–979.

[46] Strezoska, Z., Paunesku, T., Radosavljevic, D., Labat, I., Drmanac, R., and Crkvenjakov, R. (1991) *Proceedings of the National Academy of Sciences* **88(22)**, 10089–10093.

[47] Southern, E. (1975) *Journal of Molecular Biology* **98(3)**, 503–517.

[48] Vercoutere, W. and Akeson, M. (2002) *Current Opinion in Chemical Biology* **6(6)**, 816–822.

[49] Chan, V., Graves, D. J., and McKenzie, S. E. (1995) *Biophysical Journal* **69(6)**, 2243–2255.

[50] Riccelli, P. V., Merante, F., Leung, K. T., Bortolin, S., Zastawny, R. L., Janeczko, R., and Benight, A. S. (2001) *Nucleic Acids Research* **29(4)**, 996–1004.

[51] Hayakawa, T. (2003) *Master's thesis, The University of Tokyo*.

[52] Zhang, J., Lang, H. P., Yoshikawa, G., and Gerber, C. (2012) *Langmuir : the ACS Journal of Surfaces and Colloids* **28(15)**, 6494–6501.

[53] Moreira, B. G., You, Y., and Owczarzy, R. (2015) *Biophysical Chemistry* **198**, 36–44.

[54] Jafilan, S., Klein, L., Hyun, C., and Florián, J. (2012) *Journal of Physical Chemistry B* **116(11)**, 3613–3618.

[55] Brown, R. F., Andrews, C. T., and Elcock, A. H. (2015) *Journal of Chemical Theory and Computation* **11**, 2315–2328.

[56] Simpkins, H. and Richards, E. G. (1967) *Biochemistry* **6(8)**, 2513–2520.

[57] Borer, P. N., Uhlenbeck, O. C., Dengler, B., and Tinoco, I. (1973) *Journal of Molecular Biology* **80(4)**, 759–771.

[58] Jean, J. M. and Hall, K. B. (2004) *Biochemistry* **43**, 10277–10284.

[59] Warshel, A. (1979) *The Journal of Physical Chemistry* **83(12)**, 1640–1652.

[60] Cieplak, P. and Kollman, P. A. (1988) *Journal of the American Chemical Society* **110(12)**, 3734–3739.

[61] Danilov, V. I. and Tolokh, I. S. (1984) *Journal of Biomolecular Structure & Dynamics* **2(1)**, 119–130.

[62] Sponer, J., Jurecka, P., Marchan, I., Luque, F. J., Orozco, M., and Hobza, P. (2006) *Chemistry (Weinheim an der Bergstrasse, Germany)* **12(10)**, 2854–2865.

[63] Norberg, J. and Nilsson, L. (1995) *Biophysical Journal* **69(6)**, 2277–2285.

[64] Schrodt, M. V., Andrews, C. T., and Elcock, A. H. (2015) *Journal of Chemical Theory and Computation* **11(12)**, 5906–5917.

[65] Dixit, S. B., Beveridge, D. L., Case, D. a., Cheatham, T. E., Giudice, E., Lankas, F., Lavery, R., Maddocks, J. H., Osman, R., Sklenar, H., Thayer, K. M., and Varnai, P. (2005) *Biophysical Journal* **89(6)**, 3721–3740.

[66] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) *The Journal of Chemical Physics* **79(2)**, 926–935.

[67] Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005) *Journal of Computational Chemistry* **26(16)**, 1701–1718.

[68] Hess, B., Kutzner, C., van derSpoel, D., and Lindahl, E. (2008) *Journal of Chemical Theory and Computation* **4(3)**, 435–447.

[69] Cheatham, T. E., Cieplak, P., and Kollman, P. A. (1999) *Journal of Biomolecular Structure & Dynamics* **16(4)**, 845–862.

[70] Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A., and Orozco, M. (2007) *Biophysical Journal* **92(11)**, 3817–3829.

[71] Krepl, M., Zgarbová, M., Stadlbauer, P., Otyepka, M., Banáš, P., Koča, J., Cheatham, T. E., Jurečka, P., and Sponer, J. (2012) *Journal of Chemical Theory and Computation* **8(7)**, 2506–2520.

[72] Nosé, S. (1984) *The Journal of Chemical Physics* **81(1)**, 511–519.

[73] Hoover, W. (1985) *Physical Review A* **31(3)**, 1695–1697.

[74] Parrinello, M. (1981) *Journal of Applied Physics* **52(12)**, 7182–7190.

[75] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) *The Journal of Chemical Physics* **103(19)**, 8577–8593.

[76] Murai, R. (2016) *Master's thesis, The University of Tokyo*.

Acknowledgements

I am most grateful to Prof. Akira Suyama for supervising my doctoral dissertation. I am deeply grateful to him and to all the members of his laboratory, especially, Dr. Koichiro Shohda and Ms. Maasa Yokomori, without whose help I could have done little.

I greatly appreciate Mr. Tetsuro Kitajima for his support for the stopped-flow experiments described in Chapter 2. I would like to thank Ms. Yoko Sakai for her support for the thermal melting experiments described in Chapter 2, and the microarray experiments described in Chapter 3. I thank a lot Mr. Ryosuke Murai for his technical supports and insightful discussion relating to the MD simulations described in Chapter 4.

The research described in Chapter 4 was partially supported by MEXT SPIRE Supercomputational Life Science. I was supported by the Japan Society for the Promotion of Science through the Program for Leading Graduate Schools (MERIT).

I am grateful for stimulating discussions with many scientists in the wider community, including Prof. Takashi Kato, Mr. Junya Yamada, Ms. Eriko Watanabe, and many others. I would like to thank Shinnosuke Seki for helping me on a once-in-a-lifetime trip to Finland, where I was able to share ideas with many people who think about DNA computing/nanotechnology; I would like to name in particular Mr. Abdulmelik Mohammed, Dr. Jugen Czeizler, Dr. Pierre-Étienne Meunier, Prof. Juho Rousu, and Prof. Pekka Orponen.

Finally, I thank my parents and my family for all the ways they have nurtured my growth.

Chapter A

Figures

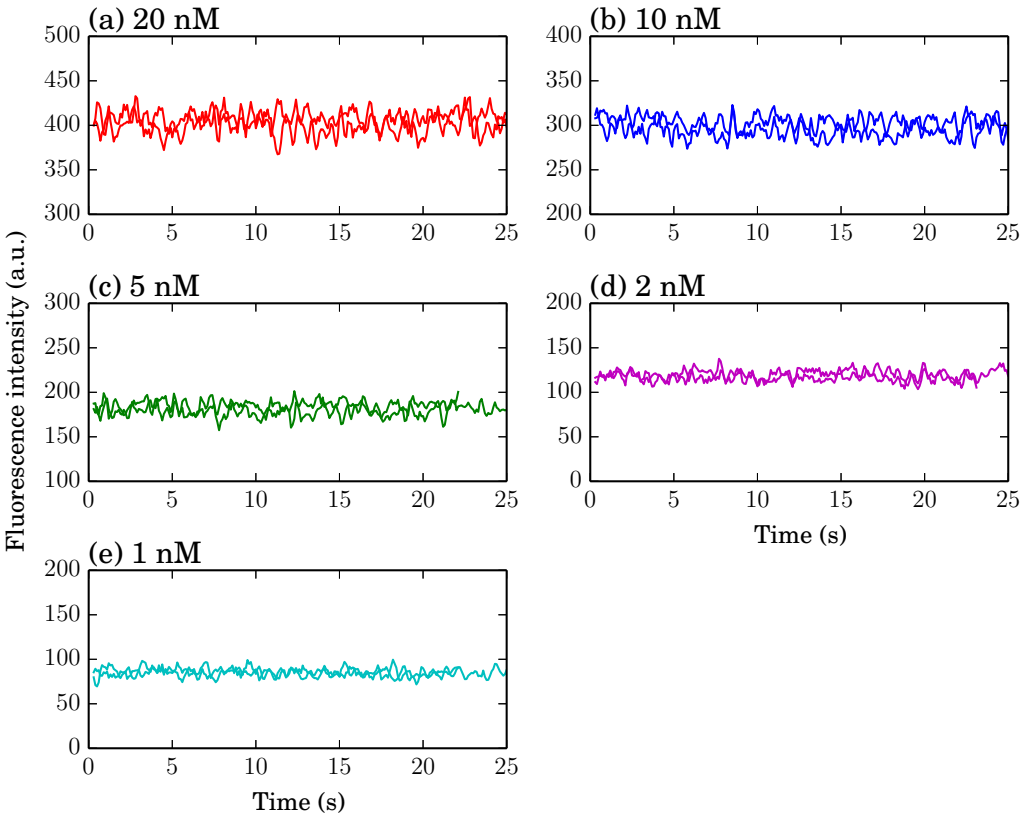

Figure A1: Fluorescence intensities after mixing PicoGreen®dye and dsDNA solutions using a stopped-flow apparatus. (a) Time courses of fluorescence intensities for 20 nM dsDNA at the same dye concentration as that used for the hybridization kinetic measurements. Fluorescence intensities were measured every 0.1 s. Data obtained from three independent experiments are shown here. (b–e) Time courses of fluorescence intensities as in (a), but with dsDNA concentrations of 10, 5, 2, and 1 nM, respectively.

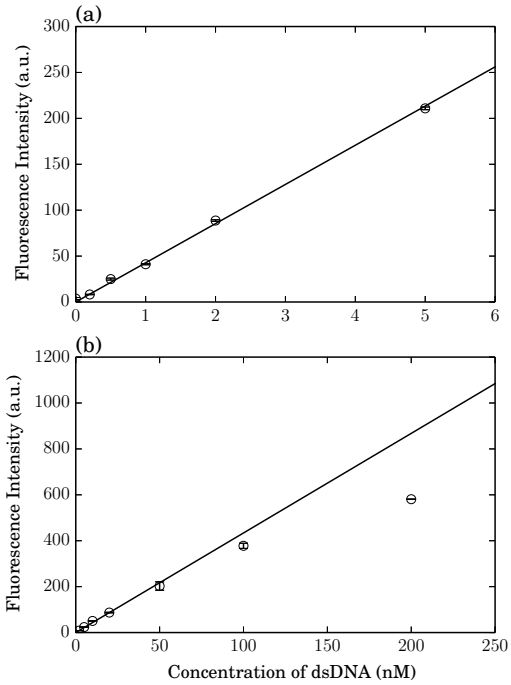

Figure A2: Linearity between the fluorescence intensity from PicoGreen®dye and the concentration of dsDNA. (a) Fluorescence intensities were measured every 0.5 s for 30 s at dsDNA concentrations of 0, 0.2, 0.5, 1, 2, and 5 nM. The plotted values for the fluorescence intensity are averages over the 30 s for each measurement. Error bars indicate standard deviations obtained by three independent measurements. A solid line is the best fit to the all data points. The DNA sequence used was AGTCGGTTGCTATTAACAGACGG. (b) Fluorescence intensities as in (a), but the dsDNA concentrations used were 2, 5, 10, 20, 50, 100, and 200 nM. A solid line is the best fit to the data points at the dsDNA concentrations of 2, 5, 10, 20, 50, and 100 nM.

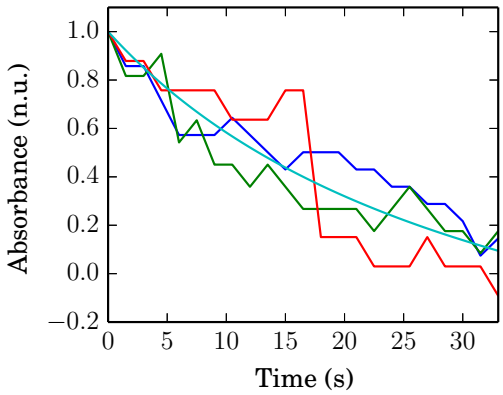

Figure A3: Typical kinetic DNA hybridization data obtained by the absorbance method. Hybridization was observed by monitoring the decrease in UV absorbance at 260 nm upon mixing equal molar (0.1 μM) concentrations of complimentary strand solutions. The DNA sequence used was no. 23 in Table 2.2.1. These solutions were mixed via hand pipetting into a cell. The cell was pre-equilibrated at 25°C and hybridization was monitored at this fixed temperature. Red, blue, and green lines represent data up to 95% hybridization. A light blue line is the best-fit curve to data within the time range shown. The hybridization rate constant calculated by non-linear fit was $3 \pm 2 \times 10^5$ $M^{-1}s^{-1}$, which was consistent with that determined using the fluorescence method ($1.73 \pm 0.03 \times 10^5$ $M^{-1}s^{-1}$). The normalized absorbance was obtained in such a way that one normalized unit (n.u.) of absorbance corresponded to the maximum absorbance data point of each trace, and $t = 0$ also corresponded to this point. An absorbance of 0 n.u. corresponded to the average of the final 200 data points of absorbance.

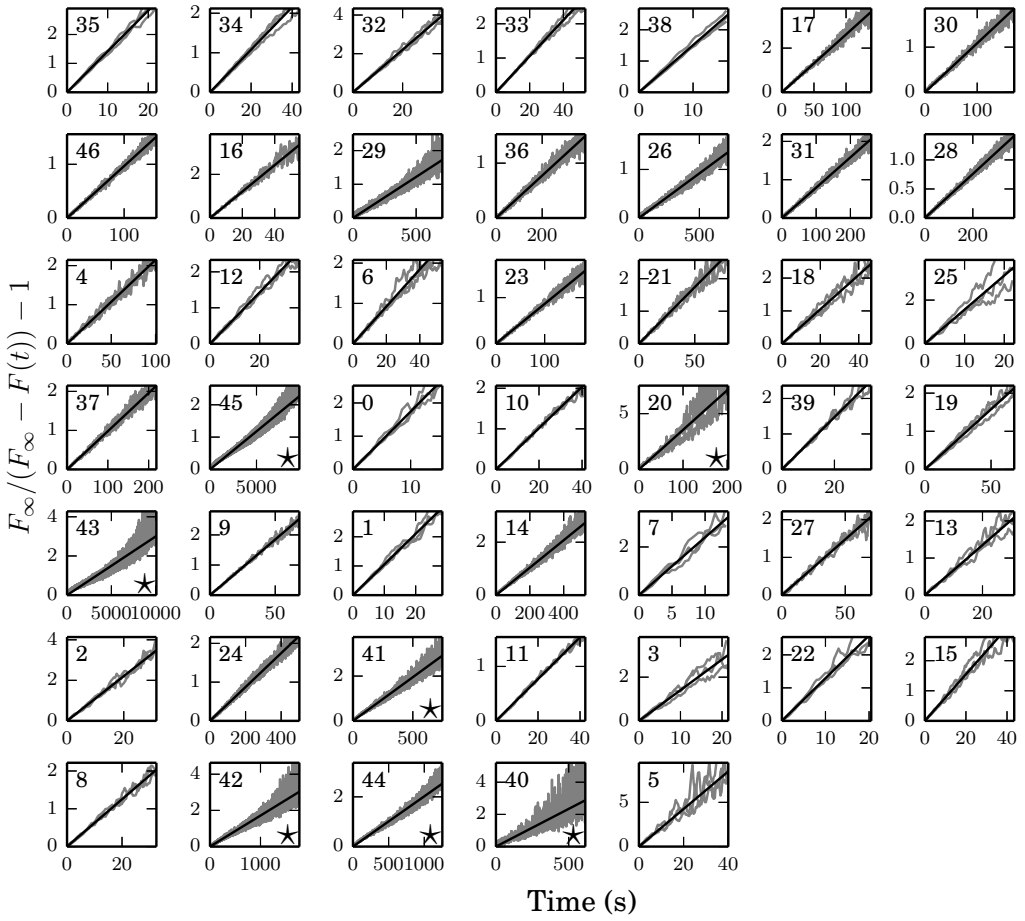

Figure A4: Second-order rate plots for the data in Figure 2.3.1. The sequence number is shown in the upper left of each panel. Black lines represent the best fit of the data obtained from three independent experiments (gray lines). The value of F_∞ was determined by non-linear regression using Equation 2.3. Stars indicate data from sequences for which hybridization could not be described by second-order reaction kinetics.

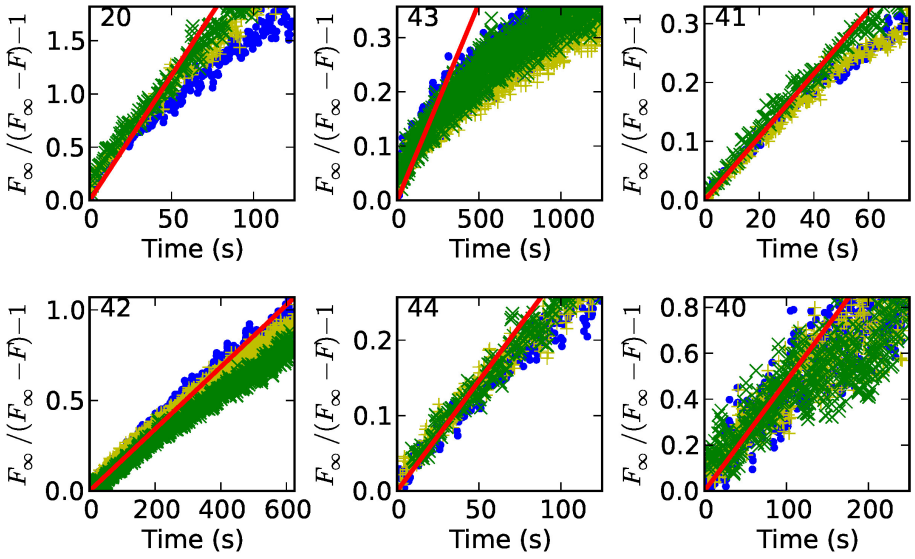

Figure A5: The linearized fluorescence data showing the hybridization kinetics for the six sequences for which hybridization could not be described by second-order reaction kinetics. The sequence number is shown in the upper left of each panel. The solid red line is fit to the data by linear regression using Equation 2.2, and the slope corresponds to the second-order hybridization rate constant.

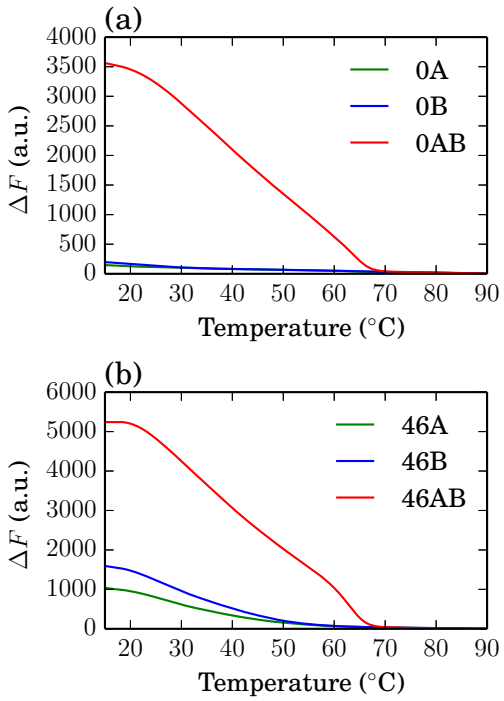

Figure A6: Melting curves for ssDNA and dsDNAs for typical DNA sequences monitored by fluorescence spectroscopy. (a) Thermal melting of two complementary ssDNA strands (0A and 0B) and dsDNA strands (0AB) for a sequence with no known negative ΔG secondary structure. The sequence is shown in Table 2.2.1 (sequence no. 0). Samples of 50 nM ssDNA in 1× SSC and 0.5× SYBR® Green I were heated at a ramp rate of 0.5 °C/min. (b) Thermal melting of two complementary ssDNA/dsDNA, as in (a), but for a sequence having negative ΔG ssDNA secondary structures (sequence no. 46).

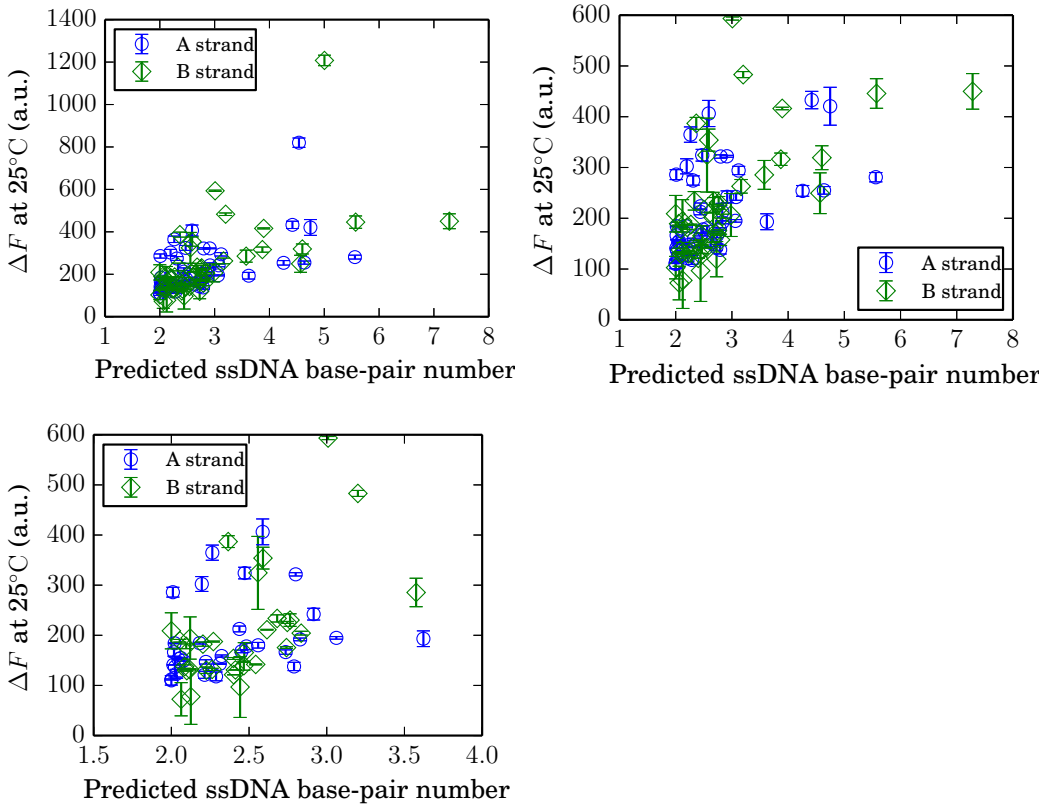

Figure A7: Graph of the value of ΔF at 25°C versus the predicted number of intramolecular base-pairs. (Top-left) Plotted values for the predicted number of base-pairs for each single-strand were calculated using UNAFold under the same conditions employed in the melting-curve measurements, from which the values of fluorescence intensity were obtained. Error bars were calculated from duplicate experiments ($N = 2$). The correlation coefficient was 0.61 ($p = 5 \times 10^{-6}$). (Top-right) Same as in (Top-left), but the two data points which showed the maximum ΔF values for A and B strands were removed to test a robustness of the correlation. The correlation coefficient of values of ΔF at 25°C and predicted base-pair numbers was 0.58 ($p = 1 \times 10^{-9}$). (Bottom-left) Same as in (Top-left), but data for sequences which were predicted to have only positive ΔG secondary structures were plotted. The correlation coefficient of values of ΔF at 25°C and predicted base-pair numbers was 0.45 ($p = 2 \times 10^{-4}$).

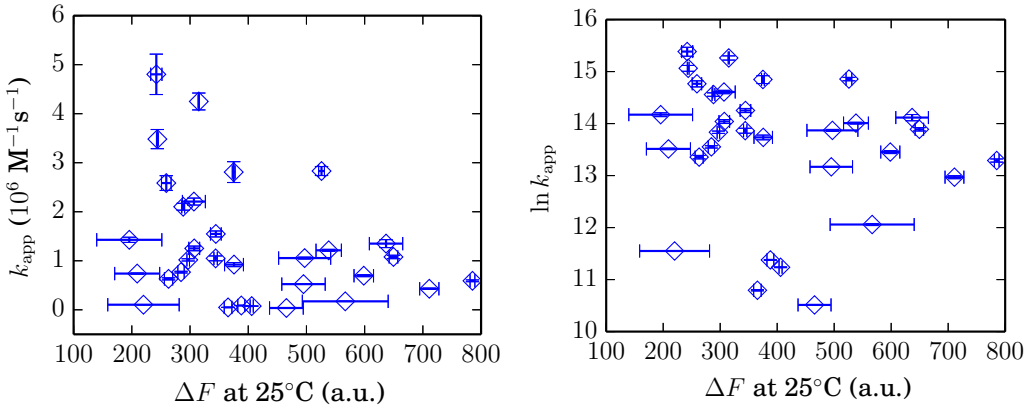

Figure A8: Plots of k_{app} versus ΔF at 25°C obtained from the ssDNA melting experiments. (Left) The data of sequences having only positive ΔG secondary structures were plotted. Vertical error bars correspond to standard deviations calculated from three independent experiments. Horizontal error bars were calculated from duplicate experiments ($N = 2$). The correlation coefficient was -0.32 ($p = 0.08$). (Right) The logarithms of hybridization rate constants were plotted against ΔF values. The correlation coefficients was -0.19 ($p = 0.3$).

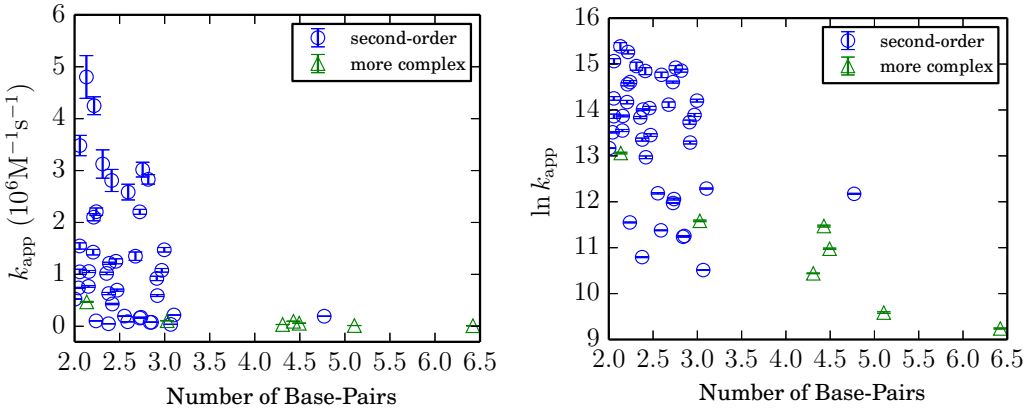

Figure A9: Plots of the observed hybridization rate constant, k_{app}, versus the predicted number of intramolecular base-pairs of ssDNA. (Left) The hybridization rate constants were plotted in a linear scale. The prediction was performed using UNAFold. The predicted numbers of base-pairs for each single-strand involved in the duplex were summed to obtain the number plotted. The correlation coefficients between the two values for the second-order sequences, for the more complex sequences, and for all of the sequences were -0.31 ($p = 0.05$), -0.80 ($p = 0.03$), and -0.4 ($p = 0.005$), respectively. (Right) The logarithms of hybridization rate constants were plotted against the predicted number of intramolecular base-pairs. The correlation coefficients between the two values for the second-order sequences, for the more complex sequences, and for all of the sequences were -0.35 ($p = 0.03$), -0.93 ($p = 0.002$), and -0.67 ($p = 3 \times 10^{-7}$), respectively.

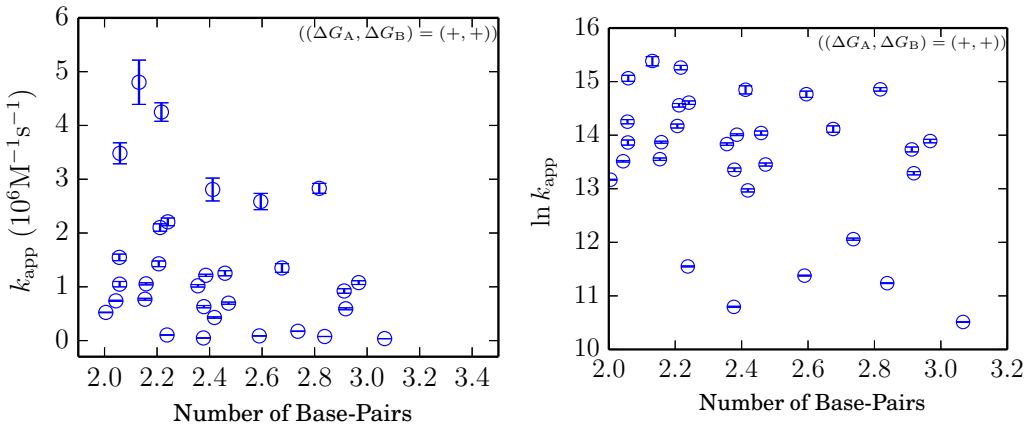

Figure A10: Plots of the observed hybridization rate constant, k_{app}, versus the predicted number of intramolecular base-pairs of ssDNA for sequences having only positive ΔG secondary structures. (Left) The hybridization rate constants were plotted in a linear scale. The correlation coefficient was -0.30 ($p = 0.1$). (Right) The logarithms of hybridization rate constants were plotted against the predicted number of intramolecular base pairs. The correlation coefficients was -0.37 ($p = 0.04$).

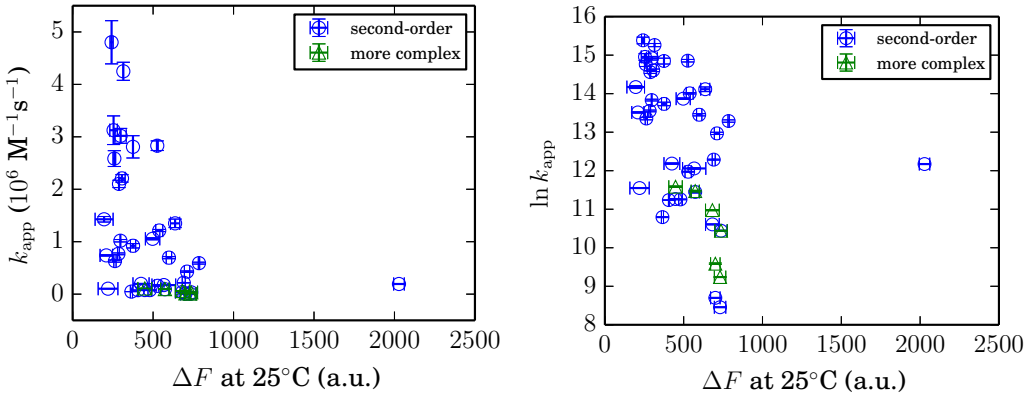

Figure A11: For sequences having a T_{m} of 63.5 ± 1.5°C, the dependence of k_{app} on the value of ΔF at 25°C obtained from ssDNA melting experiments. (Left) The hybridization rate constants were plotted in a linear scale. The correlation coefficients between the two values for the second-order sequences, for the more complex sequences, and for all of the sequences were -0.38 ($p = 0.02$), -0.90 ($p = 0.01$), and -0.40 ($p = 0.008$), respectively. (Right) The logarithms of hybridization rate constants were plotted against the values of ΔF. The correlation coefficients between the two values for the second-order sequences, for the more complex sequences, and for all of the sequences were -0.37 ($p = 0.03$), -0.78 ($p = 0.07$), and -0.41 ($p = 0.007$), respectively.

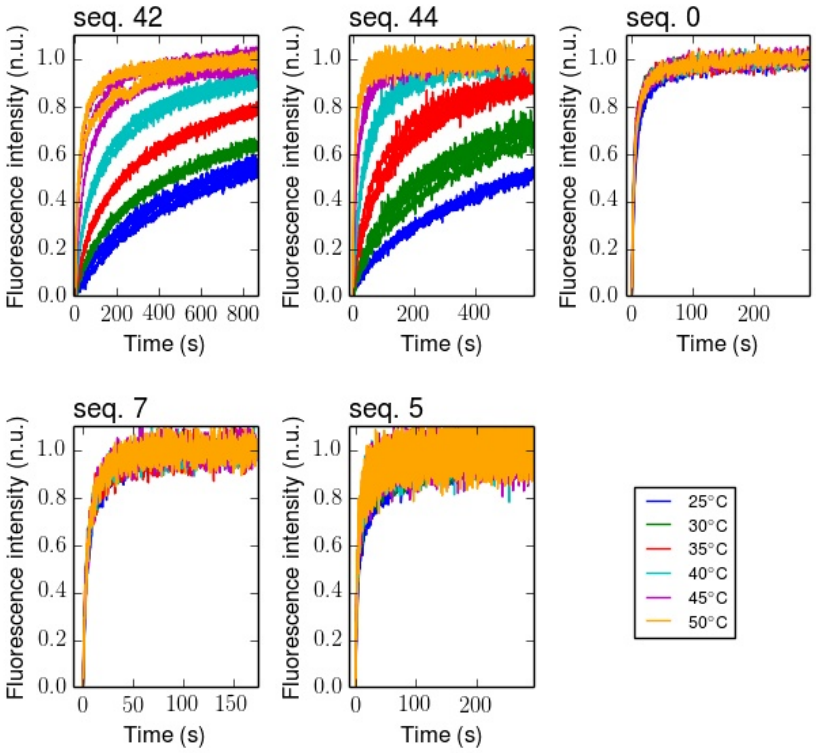

Figure A12: Time courses of hybridization at 25, 30, 35, 40, 45, and 50°C for five sequences. Sequence numbers are shown above panels. Colors correspond to temperatures. Hybridization was measured three times for each temperature, and the three traces are plotted. Fluorescence intensities were normalized at the intensity at equilibrium.

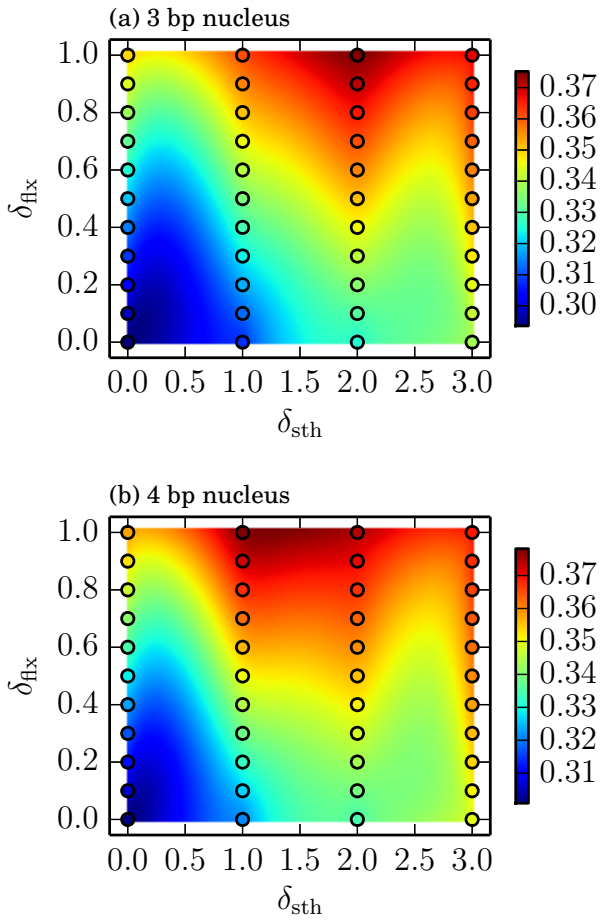

Figure A13: Dependence of the correlation coefficient between k_{app} and k_{calc} on δ_{flx} and δ_{sth}. (a) Correlation in the case of a nucleus composed of three successive base-pairs ($L_{\mathrm{nuc}} = 3$). (b) Correlation in the case of a nucleus composed of four successive base-pairs ($L_{\mathrm{nuc}} = 4$).

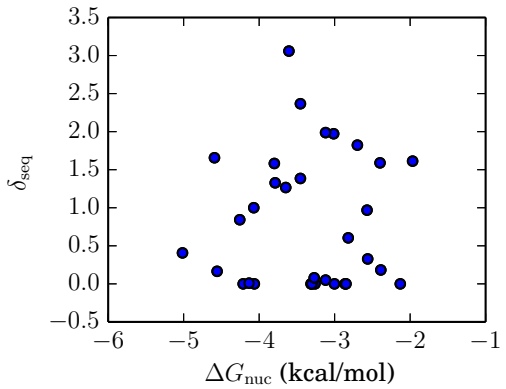

Figure A14: A plot of the δ_{seq} value versus the nucleus duplex stability, ΔG_{nuc}. The ΔG_{nuc} value was calculated using Nearest-Neighbor parameters at 25°C.

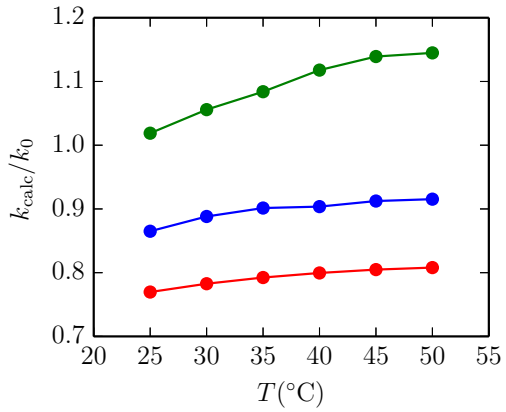

Figure A15: Temperature dependence of k_{calc}/k_0, which is the ratio of the calculated hybridization rate constant, k_{calc}, and the rate constant of hybridization between the unstructured coils, k_0. Red, blue, and green circles correspond to data from sequence no. 0, 5, and 7, respectively.

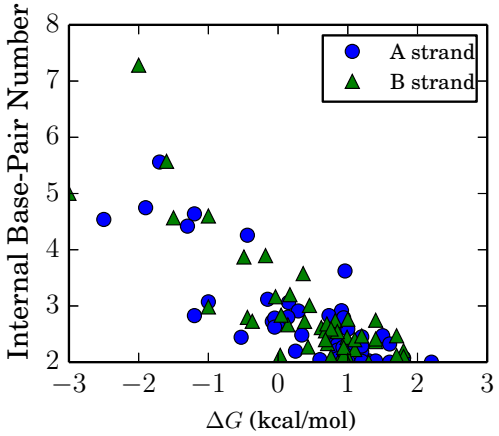

Figure A16: A plot of the predicted ΔG values versus the predicted internal base-pair numbers. The correlation coefficient was -0.83 ($p = 2 \times 10^{-13}$).

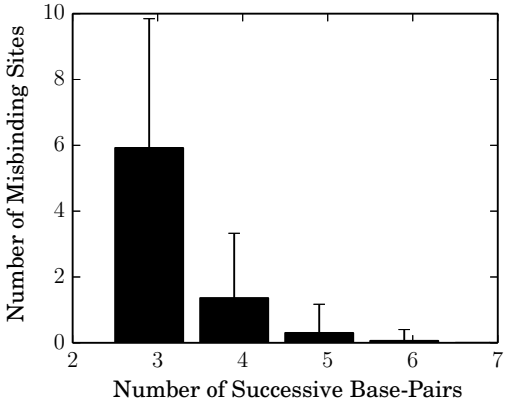

Figure A17: The number of misbinding sites composed of a certain number of successive base-pairs. Error bars represent standard deviations of 47 DNA sequences used in the study of solution hybridization.

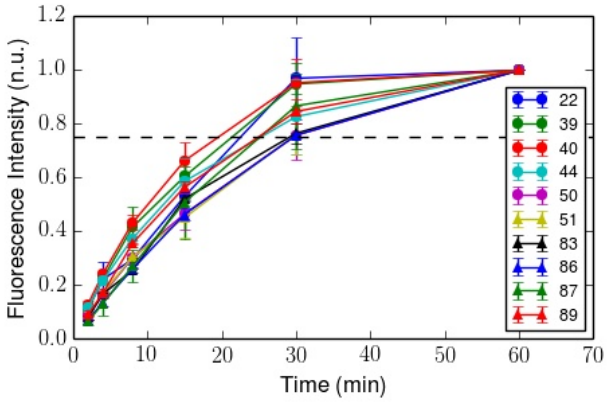

Figure A18: Time courses of hybridization of the ten fast-hybridizing sequences at an A strand concentration of 10 nM. The error bars were calculated from two independent experiments. The fluorescence intensity was measured at 2, 4, 8, 15, 30, and 60 min, and normalized at 60 min. For all of the sequences, the value of the fluorescence intensity at 30 min was higher than 0.75 n.u. (the broken line). Therefore, the hybridization for these sequences was considered to have reached equilibrium under these hybridization conditions.

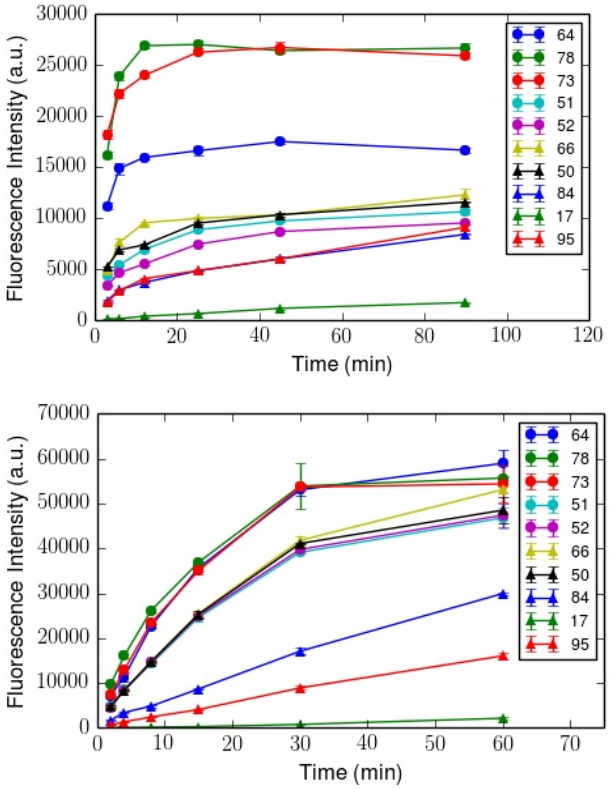

Figure A19: Observed surface hybridization kinetics. (Top) Time courses of hybridization of the ten sequences at an A strand concentration of 100 nM. The error bar represents the standard deviation calculated from fluorescence intensities of three array spots for each sequence. The fluorescence intensity was measured at 3, 6, 12, 25, 45, and 90 min. (Bottom) Hybridization kinetics related to the same sequences as in the top panel, but the concentration of A strands was only 10 nM. The fluorescence intensity was measured at 2, 4, 8, 15, 30, and 60 min.

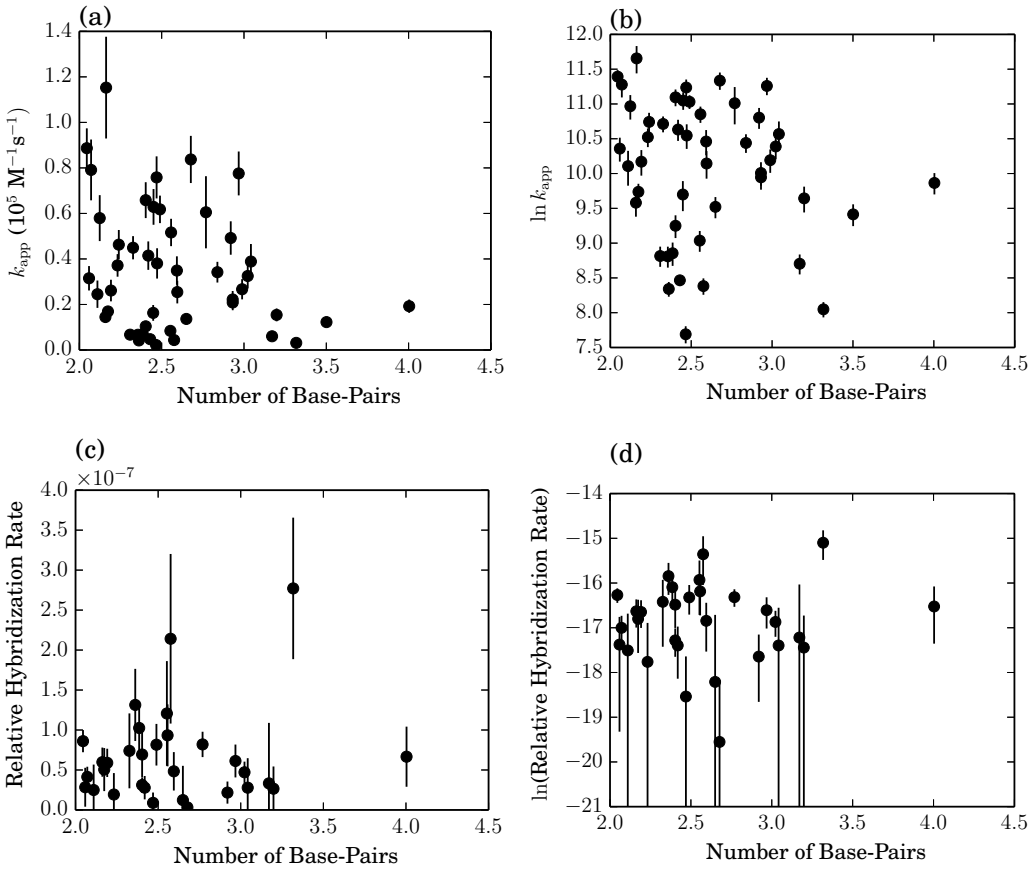

Figure A20: Relationships between the hybridization rate constant and the number of predicted base-pairs in ssDNA strands for sequences without negative ΔG secondary structures. (a) The concentration of the A strands was 1 μM. Error bars represent standard errors of the fitting parameter. The number of base-pairs was calculated for each ssDNA strand using secondary structure prediction, and averaged for each pair of strands. The correlation coefficient was -0.25 ($p = 0.09$). (b) The logarithms of hybridization rate constants were plotted against the predicted numbers of intramolecular base-pairs. The correlation coefficients was -0.18 ($p = 0.2$). (c) Same as in (a), but the A strand concentration was 2 nM. The definition of the relative hybridization rate is described in the text. The correlation coefficient was 0.16 ($p = 0.4$). (d) The logarithms of hybridization rate constants at 2 nM A strand concentration were plotted against the predicted numbers of intramolecular base-pairs. The correlation coefficients was 0.06 ($p = 0.8$).

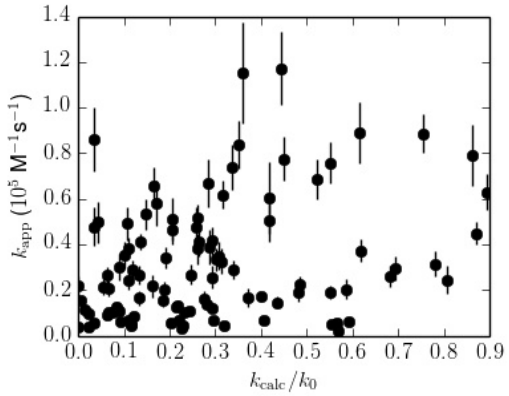

Figure A21: A comparison of the observed hybridization rate, k_{app}, and the calculated hybridization rate k_{calc}. The value of k_{calc}/k_0 calculated according to the method described in Chapter 2. Values of parameters, which were determined in Chapter 2, were used for the calculation.

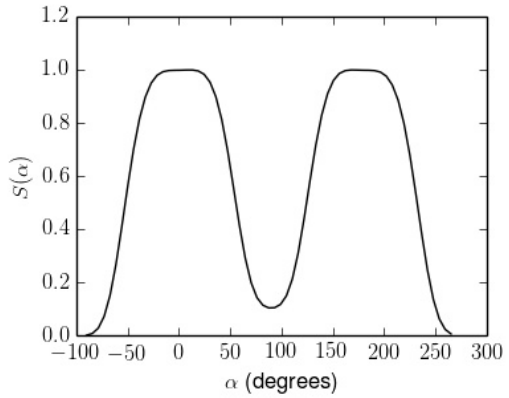

Figure A22: Values of S are plotted as a function of α.

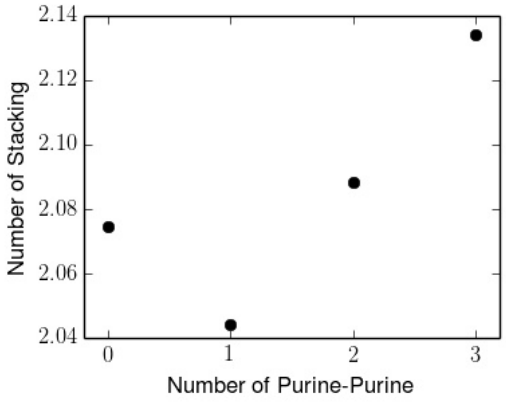

Figure A23: A plot of the stacking number versus the 'number of purine-purine', which refers to the number of adjacent purine pairs in a nucleotide tetramer. (For example, the 'number of purine-purine' for AAAT and ATAT are 2 and 0, respectively). Values of the stacking number were calculated by averaging those of tetramers having the same value of the 'number of purine-purine'.

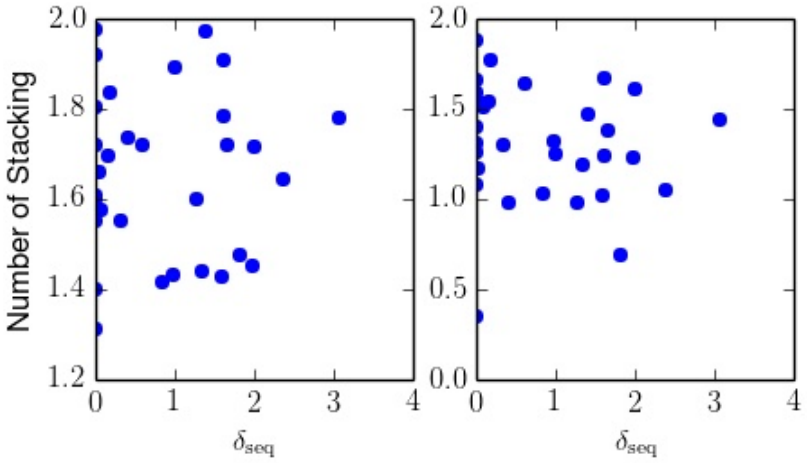

Figure A24: A comparison between the value of δ_{seq} and the stacking number obtained from simulations on nucleotide trimers. (Left) Data for trimers having the largest stacking number between two complementary trimers. The simulations for all 64 possible trimers ($4^4 = 64$) were performed using the same method as that for tetramers. The value of δ_{seq} was determined for every pair of complementary base triplets in the study of hybridization in solution (Chapter 2). (Right) Data for the other trimers.

Chapter B

Tables

Table B1: List of the frequency of base triplets in sequences used in this study.

First 2 bases	Last base in a triplet			
	A	T	G	C
AA	42	31	25	31
AT	37	31	25	27
AG	29	31	26	23
AC	31	31	22	19
TA	46	37	23	20
TT	46	42	21	23
TG	27	31	20	19
TC	27	29	21	33
GA	23	27	34	26
GT	20	31	24	26
GG	33	19	25	24
GC	19	23	16	24
CA	21	25	19	24
CT	23	25	19	34
CG	21	22	20	16
CC	20	26	20	25

Table B2: Determination of δ_{seq} parameters with different numbers of learning samples.

Number of samples		r_{test}[b]	Δr_{test}[c]
Learning	Test[a]		
10	30	0.44	0.21
20	20	0.68	0.13
30	10	0.86	0.22

[a] Test samples used to evaluate the utility of the determined δ_{seq} parameters.

[b] The correlation coefficient of k_{app} and k_{calc} for test samples after δ_{seq} determination is denoted by r_{test}.

[c] The difference in the r_{test} value before and after δ_{seq} determination is denoted by Δr_{test}.

Table B3: The frequency of transitions between the stacked and the unstacked conformation.

Position of stacking[a]	1	2	3
Frequency of transition[b]	1±1	0.8±0.7	1±1

[a] The position of stacking represents the number of stacking from the 5′-terminus of the nucleotide tetramer. For example, in the case of ATGC, the position of stacking of AT, TG, and GC are 1, 2, and 3, respectively.

[b] 10^3 times. The average and the SD for all 256 tetramers.

Chapter C

Determination of δ_{seq} parameters

Values of δ_{seq} parameters were determined to maximize the correlation coefficient between k_{app} and k_{calc} using the constrained Newton-CG algorithm (Truncated Newton Constrained method) implemented in the scipy.optimize.minimize function in SciPy. The initial value for each parameter was unity. All parameters were given the same upper and lower bounds, 10 and 0, respectively.

Here, the uncertainties of determined δ_{seq} parameters were evaluated using Monte Carlo simulations. From the hybridization model with the determined δ_{seq} values (described in Chapter 2), hypothetical data sets were generated. In the data generation, first, k_{calc}/k_0 values were generated from DNA sequences used in the stopped-flow experiments. Next, by adding random values with a normal distribution (with $\mu = 0$, $\sigma = 0.03 k_{calc}/k_0$) to the k_{calc}/k_0 values, a hypothetical data set was generated. The σ value was chosen according to the average of CV (coefficient of variation) values of k_{app} (0.03). Then, δ_{seq} values were recalculated using the parameter optimization described above. The redetermined δ_{seq} values were denoted as $\delta^{\dagger}_{seq}$. The $\delta^{\dagger}_{seq}$ calculation was repeated for 100 times. The average and the standard deviation of $\delta^{\dagger}_{seq}$ values were listed in Table C1. A plot of δ_{seq} values and $\delta^{\dagger}_{seq}$ values are shown in Figure C1.

Table C1: List of $\delta^{\dagger}_{\mathrm{seq}}$ parameters.

Sequence	$\delta^{\dagger}_{\mathrm{seq}}$	Sequence	$\delta^{\dagger}_{\mathrm{seq}}$
AGG/CCT	3.5 ± 0.6	GCG/CGC	0.7 ± 0.3
GTG/CAC	2.2 ± 0.4	AAA/TTT	0.5 ± 0.1
GAG/CTC	2.1 ± 0.4	TAG/CTA	0.4 ± 0.2
TTG/CAA	2.3 ± 0.5	CGG/CCG	0.2 ± 0.3
ATC/GAT	1.7 ± 0.4	CAG/CTG	0.1 ± 0.2
GGC/GCC	1.7 ± 0.3	AGA/TCT	0.1 ± 0.1
AAT/ATT	1.6 ± 0.3	GGG/CCC	0.3 ± 0.2
ATA/TAT	1.6 ± 0.3	AAC/GTT	0.0 ± 0.2
TGG/CCA	1.7 ± 0.4	ATG/CAT	0.0 ± 0.1
ACA/TGT	1.8 ± 0.3	AGT/ACT	0.5 ± 0.2
ACC/GGT	1.5 ± 0.4	AGC/GCT	0.00 ± 0.09
TCC/GGA	1.2 ± 0.3	ACG/CGT	0.00 ± 0.03
TCG/CGA	1.1 ± 0.3	TAA/TTA	0.0 ± 0.1
TAC/GTA	0.9 ± 0.2	TTC/GAA	0.00 ± 0.05
TGC/GCA	0.9 ± 0.3	TGA/TCA	0.0 ± 0.2
AAG/CTT	0.4 ± 0.2	GAC/GTC	0.0 ± 0.2

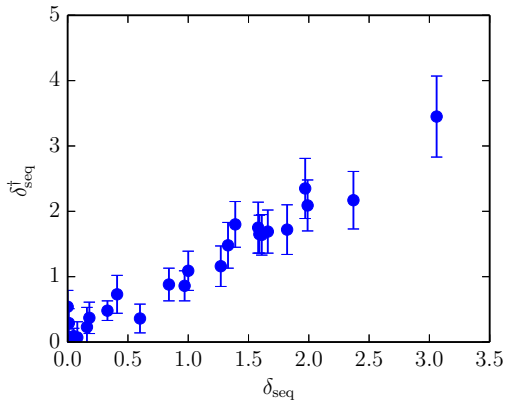

Figure C1: Comparison of δ_{seq} values and $\delta^{\dagger}_{\mathrm{seq}}$ values. The correlation coefficient was 0.98 ($p = 8 \times 10^{-23}$).

Chapter D

Experiments of single-strand base-stacking

Here, experimental observations of single-strand base-stacking (SSBS) in oligonucleotides are described. For SSBS measurements, two base sequences of GTTA and ACTT were chosen. The former contains the base triplet of GTT which was found to have the δ_{seq} value of zero, and the latter was shown to hardly form SSBS by the MD simulations in Chapter 4. DNA oligonucleotides used were commercially synthesized (SIGMA-Genosys, Hokkaido, Japan), and stored in 1× TE buffer (10 mM Tris·HCl, 1 mM EDTA, pH 8.0) at −20°C until use.

First, UV absorbance spectra were measured for the two base sequences at 25°C. The UV absorbance was recorded using DU-800 spectrophotometer (Beckman Coulter, Inc., Brea, CA, USA). A total solution volume of 325 μL and DNA concentrations of 25 μM were used in the experiments. The buffer solution was saline-sodium citrate (SSC) (150 mM NaCl, 15 mM sodium citrate). For GTTA, the concentration dependence was also investigated by measuring UV spectra at two different concentrations (25 and 12.5 μM) of DNA strands. Figure D1a shows the observed UV spectra. The two sequences at the strand concentration of 25 μM showed similar absorbance peaks around 260 nm, but the wavelengths at the maximum of absorbance were slightly different. For GTTA, the maximum appeared at 255 nm; for ACTT, the maximum appeared at 262 nm. The absorbance spectrum of GTTA at 12.5 μM showed the similar trend to that at 25 μM.

Next, the UV absorbance was measured at 80°C to study the effect of destruction of SSBS on absorbance spectra. Figure D1b shows the observed UV

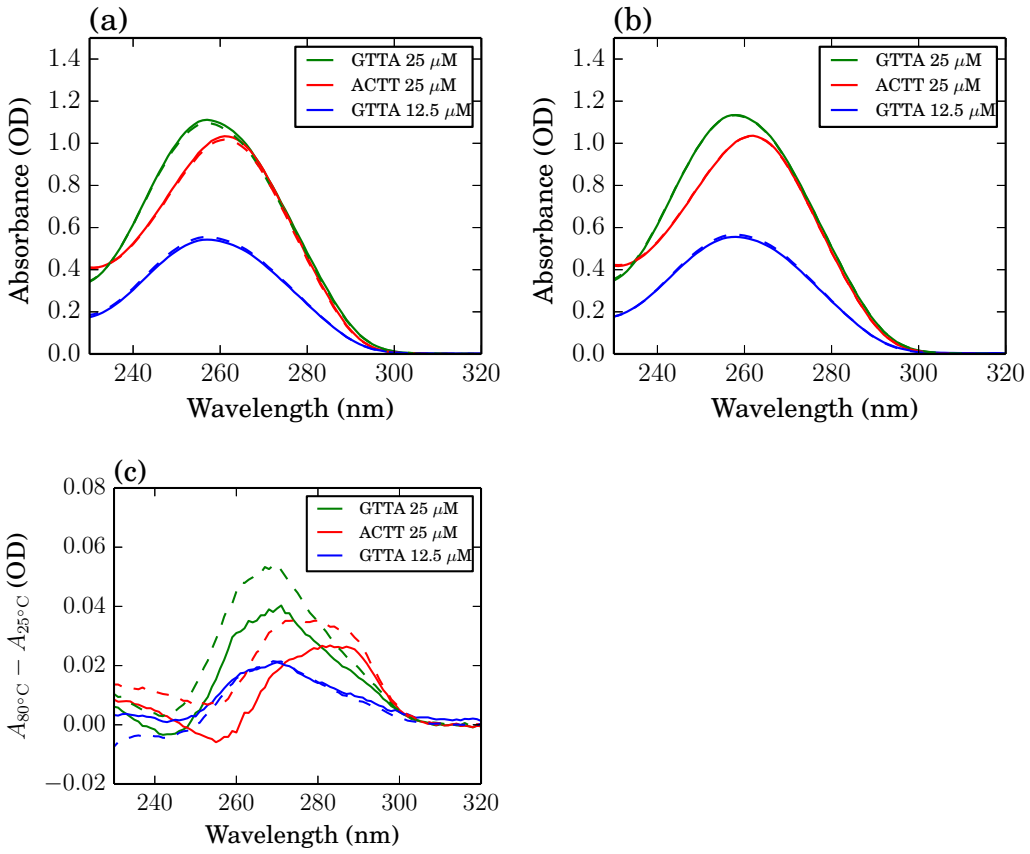

Figure D1: UV absorbance spectra for tetranucleotides. (a) UV absorbance spectra at 25°C for tetranucleotides with sequences of GTTA and ACTT. For GTTA, the absorbance spectra were measured at two different concentrations of oligonucleotides (12.5 and 25 μM). Solid and broken lines represent data obtained from two independent experiments. (b) UV absorbance spectra at 80°C. (c) Differential spectra between (a) and (b).

spectra, and Figure D1c shows the differential spectra between that at 80°C and that at 25°C. For GTTA, a peak in differential spectra was observed around 270 nm. The difference in absorbance was about 0.04 OD. This difference in absorbance spectra of oligonucleotides from low temperatures to high temperatures has been known to arise from destruction of SSBS [56]. For ACTT, the peak was found around 280 nm. Such sequence dependence of the shift of absorbance peaks was also reported previously [56]. The maximum difference in absorbance for ACTT was about 0.03 OD. On the other hand, the difference in the strand concentration of GTTA did not significantly differ the wavelength producing the maximum absorbance.

Then, melting curves for tetranucleotides were monitored by UV absorbance. From the differential UV spectra of GTTA, the absorbance at 270 nm was monitored. The reaction mixtures were the same as described above. The reaction mixtures were heated from 25 to 78°C at a rate of 1°C/min. The absorbance was recorded every 1 min. Figure D2 shows the obtained melting curves. Values of absorbance were normalized as following. The 0 n.u. corresponds to the absorbance at 25°C, and the 1 n.u. corresponds to the absorbance at 78°C. The absorbance changes in OD are shown in Table D1. For GTTA, the value of absorbance constantly increased as temperature increases, and the melting curve was not sigmoidal. This trend shows that the destruction of SSBS is not cooperative [7]. The same trend was observed for ACTT, but the value of absorbance reached unity around 55°C, and after that, the value did not increase constantly. This may show that SSBS in ACTT may totally be broken at higher temperatures than 55°C, and so the SSBS stability of ACTT is lower than that of GTTA. This relationship in SSBS stability between GTTA and ACTT was compatible with the results obtained from MD simulations described in Chapter 4. GTTA was also reported to frequently form non-adjacent base-stackings [76]. The effect of such non-native SSBS in oligonucleotides on hybridization kinetics might be related to the low δ_{seq} value. On the other hand, the melting curves of GTTA did not show any substantial difference with changing strand concentrations (Figure D2b). Therefore, for both of melting curves and absorbance spectra of GTTA did not show any concentration dependence. From this result, the base-stacking can be considered to be

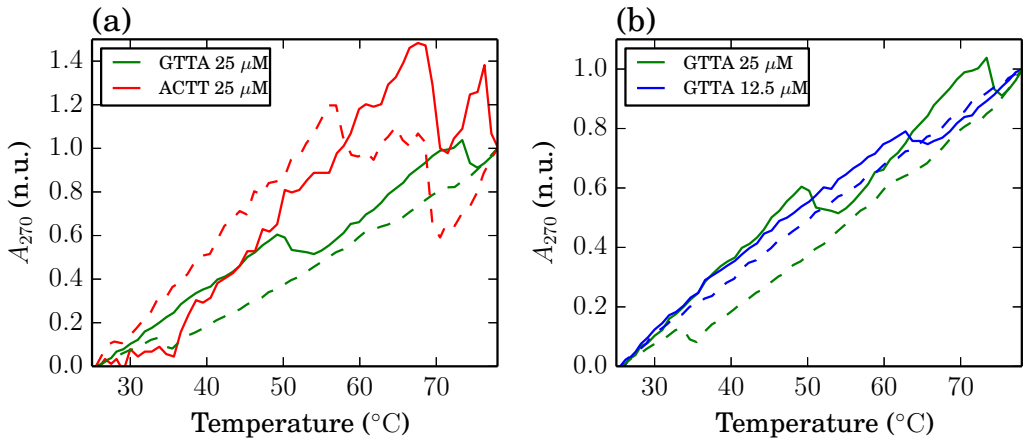

Figure D2: Melting curves for tetranucleotides monitored by UV absorbance at 270 nm. (a) Thermal melting of tetranucleotides with base sequences of GTTA and ACTT. Samples of 25 and 12.5 nM oligonucleotides in 1× SSC were heated at a ramp rate of 1°C/min. Solid and broken lines represent data obtained from two independent experiments. (b) Thermal melting of tetranucleotides, as in (a), but measured at two different concentrations of oligonucleotides with GTTA.

Table D1: Absorbance changes at 270 nm from 25 to 78°C.

	GTTA 25 μM	ACTT 25 μM	GTTA 12.5 μM
$A_{78°C} - A_{25°C}$[a] (OD)	0.046 ± 0.009	0.011 ± 0.002	0.020 ± 0.002

[a] Calculated from data obtained by two independent experiments.

unimolecular rather than bimolecular.

The sequence dependencies in UV absorbance of oligonucleotide described above may shed lights into the sequence dependence of SSBS stability. However, there are some difficulties to study the sequence dependence of SSBS. For example, there are not only intramolecular base-stackings, but also intermolecular base-stackings. To accurately distinguish these two types of base-stackings, more detailed studies about the concentration dependence in UV absorbance of oligonucleotides need to be performed. Furthermore, in the case of more longer oligonucleotides (e.g., pentamers and hexamers), intramolecular base-pairings can be formed. Such base-pairing stabilizes adjacent base-stackings. Thus, a systematic investigation to distinguish SSBS from such

base-pair-induced base-stacking would be necessary.