ABSTRACT

Title of Dissertation: THERMODYNAMICS AND APPLICATION OF A PAIR OF SYNTHETIC NUCLEOBASES FROM ARTIFICIALLY EXPANDED GENETIC INFORMATION SYSTEM

Xiaoyu Wang, Doctor of Philosophy, 2016

Dissertation Directed by: Associate Professor Jason Kahn, Department of Chemistry & Biochemistry

UV-melting experiments were performed on 9-mer duplexes containing a pair of synthetic nucleobases P·Z, two members of Expanded Genetic Information System (AEGIS), or P, Z containing mismatches. Enthalpy, entropy and free energy change were derived from simulation using two-state transition model. Nearest neighbor thermodynamic parameters of trimers or tetramers containing P·Z pair or P, Z containing mismatches were derived based on known nearest neighbor parameters. Proposed structures based on thermodynamic parameters are discussed. An application using P·Z pair as reverse selection tool of desired nucleic acid secondary structure is described.

THERMODYNAMICS AND APPLICATION OF A PAIR OF SYNTHETIC NUCLEOBASES FROM ARTIFICIALLY EXPANDED GENETIC INFORMATION SYSTEM

by

Xiaoyu Wang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Associate Professor Jason D. Kahn, Chair/Advisor
Professor Dorothy Beckett
Associate Professor Douglas Julin
Assistant Professor Paul Paukstelis
Professor Jeffrey DeStefano, Dean's Representative

Dedication

To my parents, Haiyan&Huhua Wang,

and my parents-in-law, Furong&Lang Zhang,

for their steadfast love and unceasingly support.

Acknowledgements

I would like to thank my advisor Jason Kahn, for taking me as his student, and for having been incredibly patient with me throughout the journey. He has been providing surprisingly huge academic freedom all the time, which I had no idea know how to appreciate at first. He has been willing to provide guidance and directions, at the same time insisting on leaving the joy of exploration to the student, which I had no idea how to appreciate at first either, until realizing how much faster the work could be done if I were simply made to obey orders, yet would have been deprived the chance to think, try, fail, and grow on my own. The fact that Jason has chosen to sacrifice the efficiency of getting results for the sake of my growth made him an advisor rather than a boss. Jason has also pulled me out from withdrawing crisis, twice, and I am glad that I was stopped.

The department has been providing the working opportunity to me. I am very grateful to have had the chance to work with experienced lecturers as well as to interact with students. Very importantly, this working opportunity lessened my financial burden and enabled me to support myself for years. I would like to thank all the department staffs for offering helps.

I would like to thank my current lab mates, Jason Hustedt, Iowis Zhu, and Kenneth Sharp, for tolerating my depressed moments, and still being my accompanies. I would like to thank my previous lab mates, Aaron Huesler, Kathy Goodson, Lucas Tricoli, Sarah Sucayan, Daniel Gowetski and Jeffery Leupp, who all helped me in settling down into the lab, learning experimental techniques, understanding individual projects, and offering help.

My committee members are my teachers who introduced me to fundamental knowledge when I entered into graduate school. It is my regret that I did not take as much advantage of them as I could.

I would like to thank the university for providing counseling services and writing workshops.

My parents Haiyan&Huhua Wang, have been unceasingly supportive to me. I cannot make thus far without their encouragement and help. I am deeply grateful having them as my parents.

My husband Hao Zhang is currently living in Liverpool UK, and has been taking on the role of bread-earner since I started writing. The first draft of Chapter 2 to 4 and part of Chapter 1 were completed in UK.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Synthetic nucleobases 6-amino-5-nitropyridin-2-one (Z), a purine analogue, and 2-aminoimidazo[1,2a]-1,3,5-triazin-4(8H)-one (P), a pyrimidine analogue, form a base pair via a non-Watson-Crick hydrogen bonding pattern (Figure 1-1) (Yang et al., 2011). UV-melting studies on oligonucleotides containing P·Z nucleobase pairs and P and Z containing mismatches produces quantitative thermodynamics, enabling evaluation of the stability of P, Z containing pairs and provides insight into their structures. They have led to new proposed structures for P and Z containing mismatches. The biochemical properties of P and Z make them great candidates for building blocks in nucleic acid based probes and they provide a solution to the problem of selecting a desired secondary structure in probe design.

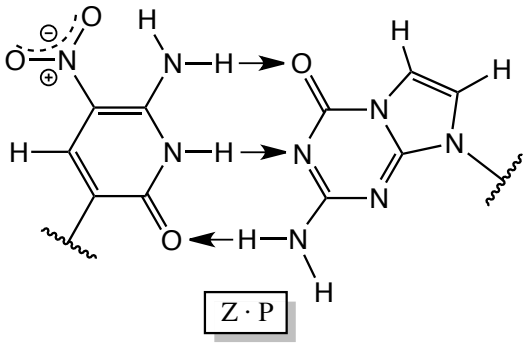


Figure 1-1: Z·P base pair

1.1 Synthetic Nucleobase Pairs as Additions to the Genetic Alphabet

Effort on seeking a third pair of nucleobases has been made for over 30 years. Adding to the adenine (A)·thymine (T), and guanine (G)·cytosine (C) pair will allow more efficient site-specific incorporation of unnatural amino acids into proteins and

eventually changing protein functions. This idea has driven synthetic biologists to explore possible artificial nucleobases forming hydrogen bonds different from Watson-Crick patterns.

Rappaport brought up the idea of adding an unnatural nucleobase into nucleic acids in 1988, and measured the thermal stability of the 6-thioguanine·5-methyl-2-pyrimidinone base pair, which was close to A·T (Figure 1-2). Synthesis and physical characterization of the mutagenic base analogue 2-amino purine (AP) found that it forms a Watson-Crick base pair with thymine, forms a Wobble pair with adenine, and its protonated structure forms a Wobble pair with C (Eritja et al., 1986) (Figure 1-3). The two tautomeric forms of isoguanine (iso-G) could pair with isocytosine (iso-C) and uracil (U), respectively (Switzer, Moroney and Benner, 1989) (Figure 1-4). And more artificial nucleobase pairs forming hydrogen bonds are listed: 5-(2,4-diaminopyridimine) (κ) and a purine analogue bearing either deoxyxanthosine (X) or N'-methyloxoformycin B (π) (Piccirilli et al., 1990; Horlacher et al., 1995) (Figure 1-5), 2-amino-6-(N,N-dimethylamino)purine (x) and pyridin-2-one (y) (Figure 1-6).

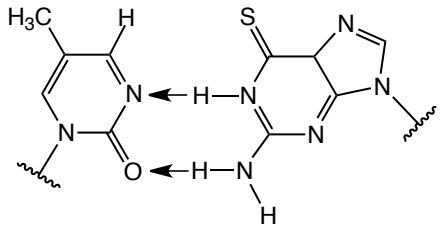


Figure 1-2: 6-thioguanine·5-methyl-2-pyrimidionone base pair

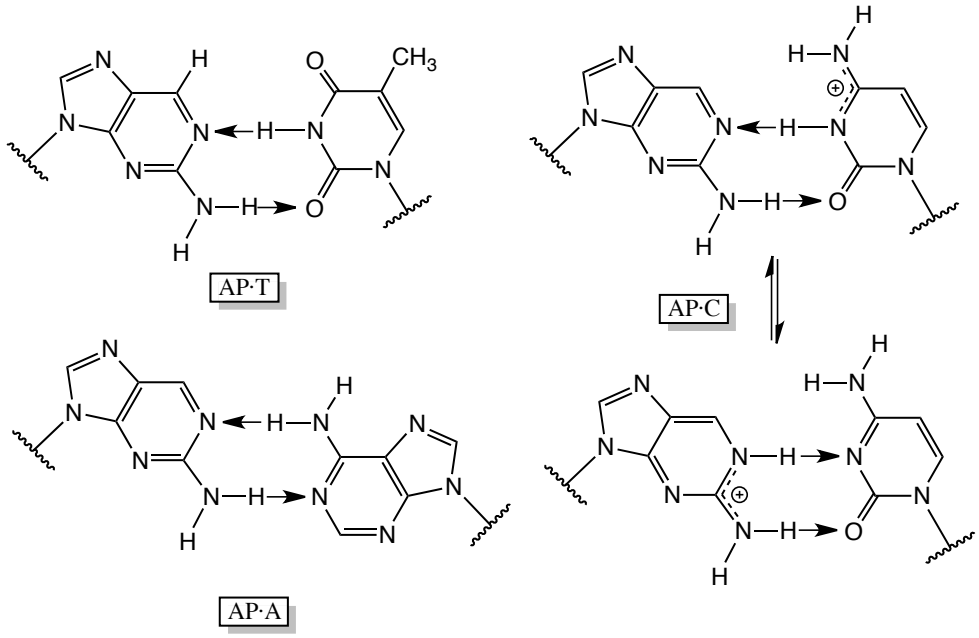


Figure 1-3: 2-amino purine form base pair or wobble pairs with A, T, and C.

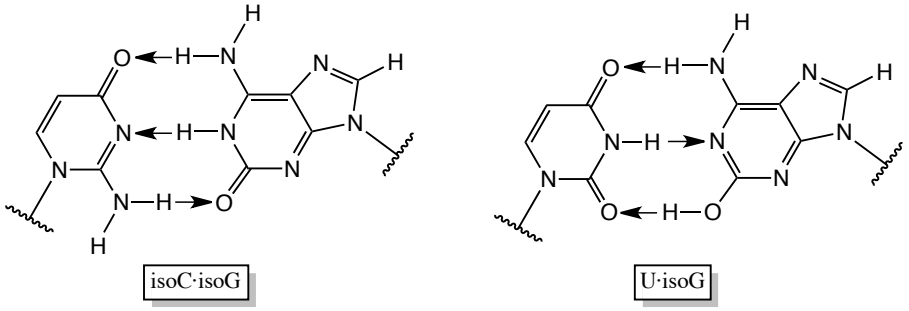


Figure 1-4: Tautomeric isomers of iso-G form base pairs with iso-C (left) and uracil (right).

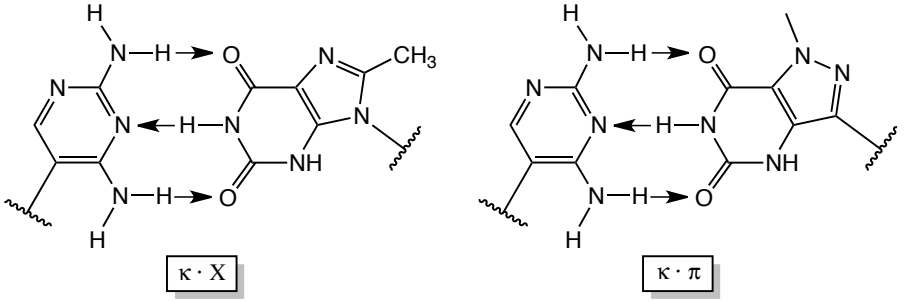


Figure 1-5: κ · X and κ · π base pair

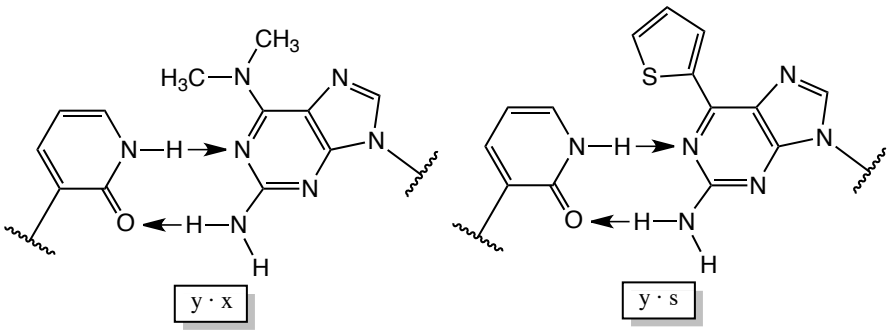


Figure 1-6: y · x and y · s base pair

The exploration of alternative nucleotides has been focused on making artificial nucleobase pairs compatible with well-established genetic systems, i.e. the nucleobases need to be complementary to each other, be able to be accepted by DNA and RNA polymerases, and to be incorporated exclusively opposite to each other. In early work, the iso-G and iso-C pair was tested against T7 RNA polymerase, avian myeloblastosis virus (AMV) reverse transcriptase and the Klenow fragment. The two analogues were incorporated during replication and transcription opposite to each other (Switzer, Moroney and Benner, 1993; Tor and Dervan, 1993). An unnatural amino acid iodotyrosine esterified to $tRNA_{CU(iso\text{-}dG)}$ was incorporated into a peptide when the (iso-C)AG codon was present in mRNA (Bain et al., 1992). However when iso-G was placed in a DNA template, both T and iso-C were incorporated in product upon extention of a primer by Klenow fragment, and only uracil (U) was incorporated opposite iso-G by T7 RNA polymerase (Switzer, Moroney and Benner, 1993). The paring between a minor tautomeric form of iso-G and U (Figure 1-4) explain those results. Inevitable confusion exists in A, T, G and C composed genetic system if iso-G and iso-C were used as the third pair.

P and Z were synthesized by Steven Benner's Lab, as members of an Artificially Expanded Genetic Information System (AEGIS) (Yang et al., 2006, 2011, 2013; Laos et al., 2013; Zhang et al., 2015). Based on their chemical structures, P·Z shall form a purine donor-donor-acceptor and pyrimidine acceptor-acceptor-donor hydrogen bonding pattern. In vitro experiments reported that P and Z have performed well in known molecular biology systems: they were accepted by DNA polymerases and were amplified in PCR (Yang et al., 2009; 2011), and they were accepted by T7 RNA polymerase and reverse transcriptase (Leal et al., 2015).

Although P and Z are believed to pair with each other orthogonally, restriction of PCR products show that interconversion between P·Z and G·C pairs occurred, indicating that P or Z might have formed mismatches with natural nucleobases (Yang et al., 2011). Thermodynamic study on P·Z as well as P-pyrimidine, Z-purine mismatches should provide insight into mismatching.

1.2 Selecting Desired Secondary Structures in Nucleic Acid Based Probe Design

Nucleic acid based probes are widely used in monitoring PCR reactions on a real time basis. Linear shaped probes can be engineered by incorporation of backbone and sugar modified nucleic acid analogues in order to increase specificity and affinity. Peptide nucleic acid (PNA) is a nucleic acid analogue with the sugar phosphate backbone substituted by an uncharged pseudopeptide backbone composed of N-(2-amino-ethyl)-glycine units (Nielsen et al, 1991). Nanomolar to femtomolar sensitivity has been achieved in vitro by combining PNA based linear nucleic acid probe with a variety of detection methods (Wang et al., 1997; Gao, Lei and Ju, 2013; Hu et al., 2015). PNA hybridizes to DNA and RNA by Watson-Crick

hydrogen-bonding and increases duplex stability. A single mismatch destabilizes a PNA-DNA duplex more than it does a DNA-DNA duplex (Ratilainen et al., 2000). Locked nucleic acid (LNA) was first synthesized for the purpose of increasing RNA binding affinity based on model building (Herdewijn, 1999): O2' and C4' of ribose was linked by -CH2-, forcing the sugar ring to take a 3'-endo conformation and facilitating nucleobase stacking (Petersen and Wengel, 2003). Sequence-dependent thermodynamic parameters for LNA-DNA duplex formation were obtained and can be used as guidance for elaborate probe design (McTigue, Peterson and Kahn, 2004).

Probes folding into secondary structure are more complicated to design, but at the same time offers more possibilities to increase probe sensitivity (Nguyen et al., 2011). An example of nucleic acid based probes, molecular beacon (MB) probe and its structure, application and design challenge is described below.

1.2.1 Structure of Molecular Beacon (MB) Probe

MB probes are single stranded oligonucleotide probes that can report the presence of specific nucleic acids in homogeneous solutions (Kessler, 2000). The classic secondary structure of MB in the "off" state is a hairpin stem-loop (Figure 1-7). The loop part binds specifically to the target sequence (analyte), and the stem part is stabilized by complementary base pair hydrogen bonding. The 5' and 3' ends are labeled by a fluorescent group and a quencher group respectively. In the "off" state, the fluorescent group is quenched by the proximal quencher group. In the presence of the analyte, the MB probe undergoes a conformational change that shifts to the "on" state: the hairpin shape opens up upon hybridizing to a target sequence, and the fluorescent group gives out signal as the quencher group is pulled away.

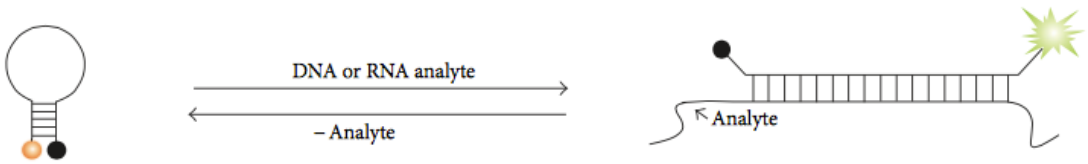


Figure 1-7:Schematic of operation principle of MB probe (Kolpashchikov, 2012)

1.2.2 Application of MB Probe

MB probes have been applied in detection of many pathogens in vitro including adenovirus, Hepatitis B virus, HIV-1 and others (Goel et al., 2005). A single nucleobase difference at the same locus is considered to be single nucleotide polymorphisms (SNP) if a nucleobase variant appears in more than 1% of the population, and SNP is the most common difference between two individuals' genome. SNP may appear in noncoding and coding regions, and they may play a role in gene expression when present in regulatory and coding areas. Therefore they may be associated with disease and mapping SNPs can provide valuable guidance for clinical diagnosis. Also SNP's can be linked to known disease gene, which is the principle of genome-wide associated studies (GWAS). MB probes can also be used for detection of, SNPs.

1.2.3 Strategies for Selecting a Desired Secondary Structure for MB probes

The loop part sequence of a MB probe is constrained by the target sequence. The stem part of the sequence is normally G·C rich to increase stability. Challenges arise when a MB probe containing desired sequences can fold into multiple secondary structures. An example is shown below (Figure 1-8). Swapping G and C sequence in stem sequence to avoid the second and the third structures may be feasible, while

consideration of quadruplex formation between consecutive GGGG sequences (Kim, Cheong and Moore, 1991) should be taken. Introducing LNA into the intended stem-forming sequence may stabilize the desired structure. Another strategy utilizing P·Z pair is described in Chapter 4.

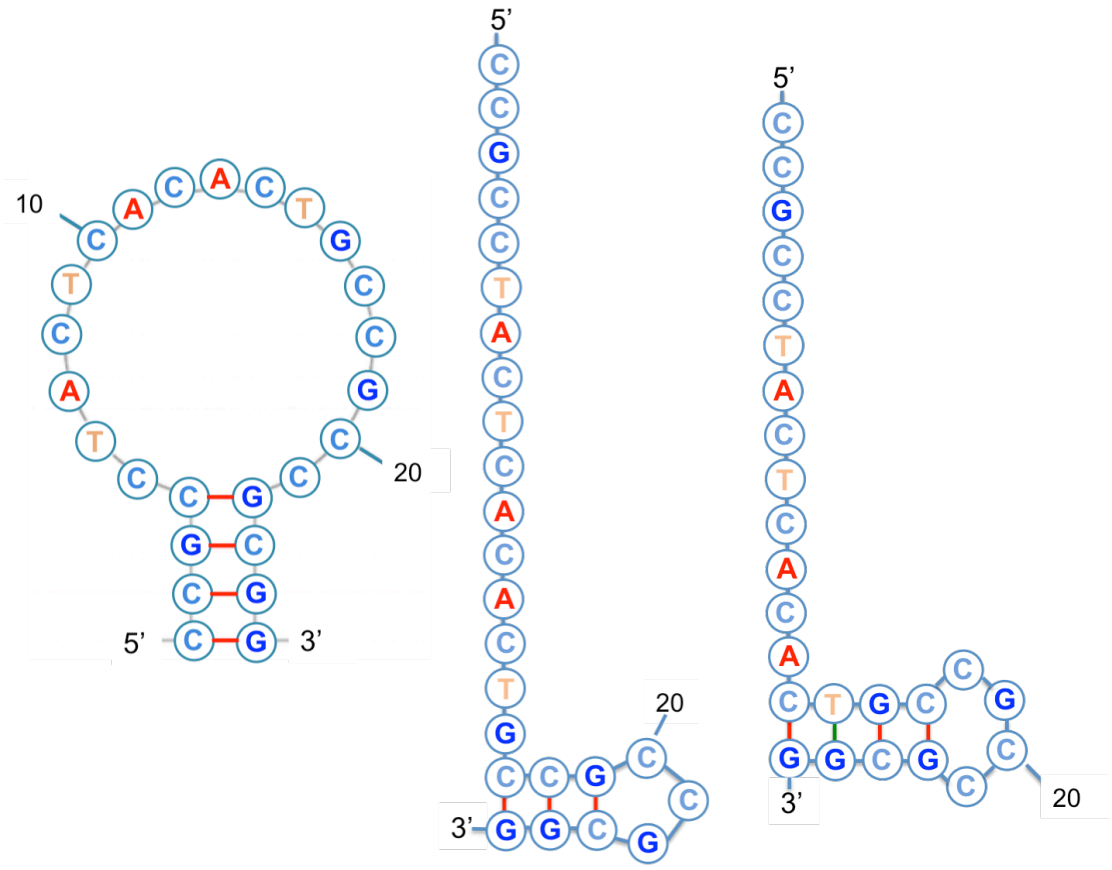


Figure 1-8: Single stranded 15-mer 5' CCGCCTACTCACACTGCCGCCGCGG folding into three secondary structures: from the left to the right, the folding free energy changes are -1.85 kcal/mole, -1.23 kcal/mole, and -0.99 kcal/mole. Secondary structures and folding free energy change are provided by Mfold.

Chapter 2: Materials and Methods

Natural nucleobases are classified into two categories, purine (adenine and guanine) and pyrimidine (thymine and cytosine). Conjugated double bonds in purine and pyrimidine rings contain π electrons, which absorb UV light at 260 nm. In UV-melting experiment, absorbance at 260 nm is collected over a broad temperature range for the purpose of evaluating the amounts of single stranded and double stranded DNA. Thermodynamic information is derived accordingly. The experimental part of this work collected absorbance versus temperature data for 29 ds9-mers.

2.1 Sequence Design

The P, Z containing sequences were synthesized by Steven Benners' Lab, and the duplex contain P·Z pair at position 2, position 3, position 5, or both position 2 and 3 (Table 2.1, 5PZ, 3PZ, 2PZ and 23PZ). We were interested in comparing the stability contribution of P·Z and G·C, so we designed a reference sequence that contains G·C at these three positions.

```
5′ G-C-C-A-G-T-T-A-A 3′
3′ C-G-G-T-C-A-A-T-T 5′
```

Thermodynamics of P, Z containing mismatches were of interest as well, and we designed sequences that would give G·Z, P·C, G·T, P·T, A·Z or A·C at these three positions, and the 2,3 doublet as well. So totally there are 4 position variants for each of the 7 base pair or mismatches, plus one reference sequence. All the single stranded and double stranded oligonucleotide sequences are listed in Table 2.1.

Table 2.1 Single-stranded and Double-stranded Oligomers used in UV-Melting Experiments

Single Stranded Oligomers			Double Stranded Oligomers with P·Z Pair or Mismatch				
ss-No	ss-Sequence	ss-Name	ds-No	ds-Sequence	ds-Name	Combination	
ss01	5'GCCAPTTAA	t5P	ds01	5'GCCAPTTAA 3'CGGTZAATT	5PZ	t5P	b5Z
ss02	5'TTAAZTGGC	b5Z	ds02	5'GCZAGTTAA 3'CGPTCAATT	3PZ	t3Z	b3P
ss03	5'GCZAGTTAA	t3Z	ds03	5'GZCAGTTAA 3'CPGTCAATT	2PZ	t2Z	b2P
ss04	5'TTAACTPGC	b3P	ds04	5'GCCAGTTAA 3'CGGTZAATT	5GZ	tRef	b5Z
ss05	5'GZCAGTTAA	t2Z	ds05	5'GCZAGTTAA 3'CGGTCAATT	3GZ	t3Z	bRef
ss06	5'TTAACTGPC	b2P	ds06	5'GZCAGTTAA 3'CGGTCAATT	2GZ	t2Z	bRef
ss07	5'GZZAGTTAA	t23Z	ds07	5'GCCAPTTAA 3'CGGTCAATT	5PC	t5P	bRef
ss08	5'TTAACTPPC	b23P	ds08	5'GCCAGTTAA 3'CGPTCAATT	3PC	tRef	b3P
ss09	5'GCCAATTAA	t5A	ds09	5'GCCAGTTAA 3'CPGTCAATT	2PC	tRef	b2P
ss10	5'TTAATTGGC	b5T	ds10	5'GCCAGTTAA 3'CGGTTAATT	5GT	tRef	b5T
ss11	5'GCTAGTTAA	t3T	ds11	5'GCTAGTTAA 3'CGGTCAATT	3GT	t3T	bRef
ss12	5'TTAACTAGC	b3A	ds12	5'GTCAGTTAA 3'CGGTCAATT	2GT	t2T	bRef
ss13	5'GTCAGTTAA	t2T	ds13	5'GCCAPTTAA 3'CGGTTAATT	5PT	t5P	b5T
ss14	5'TTAACTGAC	b2A	ds14	5'GCTAGTTAA 3'CGPTCAATT	3PT	t3T	b3P
ss15	5'GTTAGTTAA	t23T	ds15	5'GTCAGTTAA 3'CPGTCAATT	2PT	t2T	b2P
ss16	5'TTAACTAAC	b23A	ds16	5'GCCAATTAA 3'CGGTZAATT	5AZ	t5A	b5Z
ss17	5'GCCAGTTAA	tRef	ds17	5'GCZAGTTAA 3'CGATCAATT	3AZ	t3Z	b3A
ss18	5'TTAACTGGC	bRef	ds18	5'GZCAGTTAA 3'CAGTCAATT	2AZ	t2Z	b2A
			ds19	5'GCCAATTAA 3'CGGTCAATT	5AC	t5A	bRef
			ds20	5'GCCAGTTAA 3'CGATCAATT	3AC	tRef	b3A
			ds21	5'GCCAGTTAA 3'CAGTCAATT	2AC	tRef	b2A
			ds22	5'GCCAGTTAA 3'CGGTCAATT	GC	tRef	bRef
			ds23	5'GZZAGTTAA 3'CPPTCAATT	23PZ	t23Z	b23P
			ds24	5'GZZAGTTAA 3'CGGTCAATT	23GZ	t23Z	bRef
			ds25	5'GCCAGTTAA 3'CPPTCAATT	23PC	tRef	b23P
			ds26	5'GTTAGTTAA 3'CGGTCAATT	23GT	t23T	bRef

	ds27	5'GTTAGTTAA 3'CPPTCAATT	23PT	t23T	b23P
	ds28	5'GZZAGTTAA 3'CAATCAATT	23AZ	t23Z	b23A
	ds29	5'GCCAGTTAA 3'CAATCAATT	23AC	tRef	b23A

2.2 Sample Preparation

2.2.1 Stock solution preparation.

Eight ss-9mer oligonucleotides containing P or Z (ss01 to ss08, Table 2.1) were synthesized by Steve Benner's Lab, and were provided HPLC purified and lyophilized. The other eight ss-9mer containing natural bases only (ss09 to ss16, Table 2.1) were purchased from Integrated DNA Technologies®. Single stranded oligonucleotides were dissolved in 1X TE buffer (10 mM TrisHCl, 1 mM EDTA, pH 7.4) to make 100 µM stock solution, and stored in -20 °C freezer.

2.2.2 UV-melting sample preparation.

Stock solutions were thawed diluted in a 1.5 mL microcentrifuge tube with 1X cacodylate buffer (1 M NaCl, 10 mM Na cacodylate, and 0.5 mM Na_2EDTA, pH 7.0; total Na^+ concentration equals 1011 mM.). Melting samples that contain two fully or partially complementary ss-oligomers at total strand concentration (C_T) of 6 µM or 15 µM were made, as well as corresponding reference samples that contain only one ss oligomer at a concentration of 3 µM or 7.5 µM. Samples dissolved in cacodylate buffer were transferred to CARY self-masking quartz cuvettes for UV absorbance measurement. The cuvette cap was wrapped with Teflon tape 2 to 3 times and then used to seal the cuvette, in order to prevent evaporation during heating process.

2.3 Equipment Setup

UV melting profiles were collected on a Cary®100 Bio UV-visible spectrophotometer paired with a multi-cell sample charger and a Peltier heating/cooling system. The spectrophotometer and the cooling system were switched on and given several minutes for initiation before data collection. The spectrophotometer was controlled by the Cary® 'Thermal' software running under Windows operating system. Setup of 'Thermal' is described in the next paragraph. Baseline correction of spectrophotometer was performed with a sample containing cacodylate buffer only. TE buffer (10 mM Tris-Cl, 1 mM EDTA) was diluted in cacodylate buffer when melting samples or reference samples were made, the maximum EDTA concentration 0.15 mM when melting samples of 15 μM were prepared, therefore the effect of 260 nm absorbance caused by diluted TE buffer was considered negligible. The thermo probe was placed into the same cacodylate buffer sample to monitor temperature over the range of 0 °C to 85 °C. Up to 12 samples were placed into sample cells in the sample compartment for a single batch of measurements. The inlet port below the sample compartment was connected to a high-pressure nitrogen gas tank with plastic tubing. Nitrogen gas was turned on to purge the compartment whenever the temperature was below 10 °C in order to prevent moisture in the air from condensing on the cuvette surface, which could distort the absorbance reading.

Set up of 'Thermal' software: Open 'Setup' dialog box. Under 'Cary' tab input 260 nm as the value for 'Wavelength'. Click 'Advanced Collect', input details of cooling-heating ramps. The standard set up is one fast heating ramp that increases

temperature from 25 °C to 85 °C at a speed of 10 °C/min, and one slow cooling ramp that decreases temperature from 85 °C to 0 °C at the speed of 1 °C /min, and maintained at 85°C for 1 min; data was collected every 1 °C change. Under 'Accessory' tab check 'Use Cell Change', which will activate 'Select Cells' option, and check the numbers of sample cells that are used in the experiment. If the 'Multi Zero' option is selected, the absorbance reading obtained from a designated reference sample is subtracted from all the other samples. Check 'Show Status Display' option to open the real-time reading window. Under 'Reports' tab choose 'Select for ASCII (csv)' in 'Autoconvert' column to automatically save absorbance at 260 nm versus temperature profiles in csv file format.

Chapter 3: Data Analysis

Symbols and annotations mentioned in this chapter are listed below (Table 3.1).

Table 3.1 Annotation of Symbols Used in Data Analysis

Symbol	Annotation
A_{260}	Absorbance at 260 nm
$A_{260\,W}$	Absorbance at 260 nm of reference Watson strand
$A_{260\,ds}$	Absorbance at 260 nm produced by double-stranded DNA
$A_{260\,ss}$	Absorbance at 260 nm produced by single-stranded DNA
m_{ds}	The doubled slope of complete double stranded area
m_{ss}	The doubled slope of complete single stranded area
A_{min}	Absorbance at lowest temperature (all the strands in double-stranded form)
A_{max}	Absorbance at highest temperature (all the strands in single-stranded form)
Watson	The single strand in excess
Crick	The single strand complementary to *Watson*
WC	The double strand composed of *Watson* and *Crick*
$[W]$	The concentration of *Watson* strand
$[C]$	The concentration of *Crick* strand
E	The concentration of excess *Watson* strand
Keq	The equilibrium constant of hybridization
C_T	The total strand concentration
$C_{T\,melt}$	The concentration of meltable strand
α	Fraction of DNA in double-stranded form out of the total strand concentrations
β	Fraction of DNA in double-stranded form out of the meltable strand concentrations
$C_{T\,sum}$	The total strand concentration calculated from the two reference sample concentrations
$C_{T\,est}$	The total strand concentration calculated from the melt sample
f	Correction factor, the ratio of [*ExW*] to [*W*]
$Der(i)$	The derivative of absorbance at point *i*
$A_{min_{stand}}$, $A_{max_{stand}}$, $m_{ss_{stand}}$, $m_{ds_{stand}}$, $\Delta H°_{stand}$, $\Delta S°_{stand}$	Parameters determined in the standard baseline fitting.
$A_{min_{opt}}$, $A_{max_{opt}}$, $m_{ss_{opt}}$, $m_{ds_{opt}}$, $\Delta H°_{opt}$, $\Delta S°_{opt}$	Parameters determined in the optimized baseline fitting.
$m_{ss_{unrest}}$, $m_{ds_{unrest}}$, $\Delta H°_{unrest}$, $\Delta S°_{unrest}$	Parameters determined in the unrestrained baseline fitting.

$\Delta H°_{unrest(\max)}$, $\Delta H°_{unrest(\min)}$ $\Delta S°_{unrest(\max)}$, $\Delta S°_{unrest(\min)}$	Maximum and minimum enthalpy and entropy determined in the unrestrained baseline fitting.
A_{pred}, $A_{pred_{hi}}$, $A_{pred_{lo}}$, D_{pred}, $D_{pred_{hi}}$, $D_{pred_{lo}}$	Regenerated maximum and minimum absorbances and absorbance derivatives, both for meltable strands

3.1 Overview

The UV absorbance at 260 nm versus temperature curve was analyzed by a Matlab program "ofitwithE_LM" utilizing a two-state model, which assumes that the oligonucleotides are either in the helical duplex form, or in the random coil form. The parameters and fitting follows the model below:

$$A_{260} = \alpha A_{260ds} + (1-\alpha) A_{260ss} \tag{3-1}$$

where A_{260} is the total absorbance generated by dsDNA and ssDNA at 260 nm, $A_{260\,ds}$ and $A_{260\,ss}$ are absorbance when all the strands are in ds form or in ss form. Both are assumed to change linearly with temperature, and they reach the minimum or the maximum at the lowest or highest temperature:

$$A_{260\,ds} = A_{min} + \frac{1}{2} m_{ds}(T - T_{min}) \tag{3-2}$$

$$A_{260\,ss} = A_{max} + \frac{1}{2} m_{ss}(T - T_{max}) \tag{3-3}$$

Where $A_{\min}$ and $A_{\max}$ are the minimum and maximum absorbances at lowest and highest temperature, and m_{ds} and m_{ss} are twice as much as the slopes of complete double-stranded area and complete single-stranded area respectively. More details of how the values of $A_{\min}$, $A_{\max}$, m_{ds}, and m_{ss} are determined are discussed in Chapter 3.1.1.

α is the fraction of dsDNA. Consider the hybridization equilibrium

$$W + C \rightleftharpoons WC \tag{3-4}$$

According to the definition of α

$$\alpha = \frac{2[WC]}{C_T} \tag{3-5}$$

$[WC]$ can expressed in terms of α and C_T

$$[WC] = \frac{\alpha C_T}{2} \tag{3-6}$$

Where C_T is the total concentration of all strands

$$C_T = [W] + [C] + 2[WC] \tag{3-7}$$

In an equal molar reaction, $[W] = [C]$, therefore

$$[W] = \frac{1}{2}(C_T - 2[WC]) \tag{3-8}$$

Substituting $[WC]$ with α and C_T

$$C_T = [W] + [C] + 2[WC] \tag{3-9}$$

$$[W] = \frac{(1-\alpha)C_T}{2} \tag{3-10}$$

K_{eq} can be expressed in terms of α and C_T

$$Keq = \frac{[WC]}{[W][C]} = \frac{2\alpha}{(1-\alpha)^2 C_T} \tag{3-11}$$

Solving the quadratic equation to express α in terms of C_T and K_{eq}

$$\alpha = \frac{K_{eq}C_T + 1 - \sqrt{2K_{eq}C_T + 1}}{K_{eq}C_T} \tag{3-12}$$

And K_{eq} is a function of enthalpy $\Delta H°$ and entropy $\Delta S°$:

$$K_{eq} = e^{-\frac{1}{R}(\Delta H° \cdot \frac{1}{T} - \Delta S°)} \tag{3-13}$$

Therefore α can be expressed in terms of C_T, $\Delta H°$, $\Delta S°$ and T:

$$\alpha = \frac{C_T e^{-\frac{1}{R}(\Delta H^\circ \cdot \frac{1}{T} - \Delta S^\circ)} + 1 - \sqrt{2C_T e^{-\frac{1}{R}(\Delta H^\circ \cdot \frac{1}{T} - \Delta S^\circ)} + 1 + e^{-\frac{2}{R}(\Delta H^\circ \cdot \frac{1}{T} - \Delta S^\circ)}}}{C_T e^{-\frac{1}{R}(\Delta H^\circ \cdot \frac{1}{T} - \Delta S^\circ)}} \quad (3\text{-}14)$$

The two strands in a DNA melting experiment are at different concentrations. The excess of one single strand drives duplex formation at low concentrations of the other strand. The apparent fraction in double stranded form does not go to 1 at low temperature, because only one of the ssDNA concentrations goes to zero. At the same time there is a contribution to the observed absorbance from the excess single strand. To deal with this issue, absorbance resulting from the excess single strand was subtracted from the total absorbance, and the fraction double stranded form out the meltable total strands was introduced.

The existence of excess Watson strand complicates the relationship between α, C_T and K_{eq}. The definition of α remains the same, therefore Equation (3-5) and (3-6) are still validate. In Equation (3-7), substituting $[C]$ with $[W] - E$:

$$C_T = 2[W] - E + 2[WC] \quad (3\text{-}15)$$

Now $[W]$ and $[C]$ can be expressed in terms of C_T, E and α:

$$[W] = \frac{(1-\alpha)C_T + E}{2} \quad (3\text{-}16)$$

$$[C] = \frac{(1-\alpha)C_T - E}{2} \quad (3\text{-}17)$$

The expression of K_{eq} is

$$K_{eq} = \frac{[WC]}{[W][C]} = \frac{2\alpha}{(1-\alpha)^2 C_T - E^2} \quad (3\text{-}18)$$

Solving the quadratic equation to express α in terms of C_T, E, and K_{eq}

$$\alpha = \frac{K_{eq}C_T + 1 - \sqrt{2K_{eq}C_T + 1 + {K_{eq}}^2 E^2}}{K_{eq}C_T} \quad (3\text{-}19)$$

α is a function of E, and at infinite K_{eq} when all Crick strands are in the ds from, α equals $1-\frac{E}{C_T}$. It is convenient to define βas the fraction of the "meltable" DNA that is actually duplex.

$$\beta = \alpha(1-\frac{E}{C_T})^{-1} = \frac{K_{eq}C_T + 1 - \sqrt{2K_{eq}C_T + 1 + {K_{eq}}^2E^2}}{K_{eq}(C_T - E)} \tag{3-20}$$

$\beta$from 0 to 1 as K_{eq} ranges from 0 to 1. The relationship between $\Delta H°$, $\Delta S°$ and K_{eq} is the same as (3-13), βcan be expressed in terms of C_T, E, $\Delta H°$, and $\Delta S°$:

$$\beta = \frac{e^{-\frac{1}{R}(\Delta H°\cdot\frac{1}{T}-\Delta S°)}C_T + 1 - \sqrt{2e^{-\frac{1}{R}(\Delta H°\cdot\frac{1}{T}-\Delta S°)}C_T + 1 + e^{-\frac{2}{R}(\Delta H°\cdot\frac{1}{T}-\Delta S°)}E^2}}{K_{eq}(C_T - E)} \tag{3-21}$$

In order to relate β to the absorbance, the absorbance resulting from the excess *Watson* strand needs to be subtracted from the total absorbance. The spectrum of *Watson* strand is measured in experiments. This spectrum is interpolated to give absorbances at the T values of the melt spectrum by Matlab. The ratio of excess *Watson* strand to the total *Watson* strand is defined as the correction factor f

$$f = \frac{E}{\frac{1}{2}(C_T + E)} \tag{3-22}$$

The corrected absorbance A_{corr} and its relation to β is therefore

$$\begin{aligned} A_{corr}(T) &= A_{260}(T) - \frac{E}{\frac{1}{2}(C_T + E)}A_{260\,\mathrm{W}}(T) \\ &= \left[A_{max,corr} + \frac{1}{2}m_{ss}(T - T_{max})\right](1-\beta) \\ &\quad + \left[A_{min,corr} + \frac{1}{2}m_{ds}(T - T_{min})\right]\beta \end{aligned} \tag{3-23}$$

The derivative of the corrected absorbance, just referred to as the derivative, was calculated at each temperature, and the derivative versus temperature curve was

fitted. On the corrected absorbance versus temperature curve, certain data point was located, and then the second data point behind it was located as well. The difference between their corrected absorbances was divided by the difference between their temperatures to give the derivative at the temperature of the first data point:

$$Der(i) = \frac{dA_{melt}(i)}{dT(i)} = \frac{A_{melt}(i+2) - A_{melt}(i)}{T(i+2) - T(i)} \quad (3\text{-}24)$$

The derivatives are calculated from the first data point to the third from the last data point. The point following the peak point in the derivative plot has the largest tangent in the absorbance plot.

Accurate definitions of completely double-stranded area and completely single-stranded area are crucial to obtaining correct values of m_{ds} and m_{ss}, both affect the accuracy of $\Delta H°$ and $\Delta S°$. Because it is unknown how large these two areas are to each melt, the program first set the boundaries of both arbitrarily in the standard-baseline fitting round; in the optimized-baseline fitting round, the boundaries were refined by using data points with fraction of strands in double-stranded form (calculated based on parameter values obtained from standard-baseline fitting) larger than 0.99; in the final unrestrained-baseline fitting round, no definite numbers of data points were assigned to these two area, instead, the two slopes were allowed to float, and were globally fitted with $\Delta H°$ and $\Delta S°$. More details of the three rounds of fitting are described in Chapter 3.3.

3.2 Concentrations: Watson and Crick strand concentrations, total strand concentration C_T and excess Watson concentration E

The total strand concentration was estimated from the two reference samples each containing one single stranded oligonucleotide, as well as from the melt sample

that contained both single stranded oligonucleotides. Each trial (including data collected from melting sample and reference samples) was isolated from the original csv file output by the instrument, and data from the melting sample, top-strand reference sample, and bottom-strand reference sample was copied into an Excel spreadsheet as the trial's UV file. The top and bottom reference sequences 5' GCCAGTTAA (tRef) and 5' TTAACTGGC (bRef) were provided to OligoAnalyzer 3.1 (Integrated DNA Technologies®). The extinction coefficients of tRef and bRef were 91500 $L \cdot mol^{-1} \cdot cm^{-1}$ and 83600 $L \cdot mol^{-1} \cdot cm^{-1}$ according to the program. These two values were used as the universal estimated top and bottom extinction coefficients. Concentrations were calculated using Beer-Lambert Law $A = \varepsilon c l$, where A is the absorbance, ε is the extinction coefficient, and l is the path length (1 cm). The single strand with higher concentration was designated as *Watson*, and the other was designated as *Crick*. Absorbance at 20 °C was used for calculation of single strand concentrations. More than 20 randomly chosen reference samples' UV files were examined for the difference between absorbance at 20 °C and 85 °C. The differences were from 3% to 5%, while for the melting sample the difference were about 23%. Concentrations of *Watson* and *Crick* strands were added up and designated as $C_{T\,\mathrm{sum}}$. The total strand concentration was estimated from the melt sample as well. Assuming that the melt sample contained 50% each top and bottom strands, and that all strands are in single-stranded form at 85 °C, the absorbance of the melting sample at 85 °C was divided by the average of the extinction coefficients of Watson and Crick strands (87550 $L \cdot mol^{-1} \cdot cm^{-1}$) to give the estimated total strand concentration $C_{T\,\mathrm{est}}$ again using Beer-Lambert Law. Usually $C_{T\,\mathrm{sum}}$ and $C_{T\,\mathrm{est}}$ was

close to each other, in the case that the difference between $C_{T\,\mathrm{sum}}$ and $C_{T\,\mathrm{est}}$ was larger than 15%, the UV profile was abandoned. $C_{T\ \mathrm{sum}}$ was used as the total strand concentration C_T.

The excess *Watson* concentration E was calculated as the concentration difference between *Watson* and *Crick* strands.

3.3 Details of Three Rounds of Fitting

3.3.1 Fitting with standard baseline

Boundaries for double-stranded and single-stranded regions were initially set arbitrarily: the first 11 data points were assigned as the completely double-stranded region of the melt, and the last 20 data points were for assigned as the purely single-stranded region.

The four parameters that depend on the shaped of the curve were estimated based on the definition of the two regions. $m_{ds_{stand}}$ and $m_{ss_{stand}}$ were calculated as twice as the average of the derivatives in the double-stranded region and twice the average of derivative in the single-stranded region:

$$m_{ds_{stand}} = \frac{2}{11}\sum_{i=1}^{n} Ader(i) \tag{3-25}$$

$$m_{ss_{stand}} = \frac{1}{10}\sum_{i=n-19}^{n} Ader(i) \tag{3-26}$$

$A_{min_{stand}}$ was calculated as the average of the corrected absorbances of the first five data points minus the product of $m_{ds_{stand}}$ and temperature difference between the third and the first data points:

$$A_{min_{stand}} = \frac{1}{5}\sum_{i=1}^{5} A(i) - m_{ds_{stand}}[T(3) - T(1)] \tag{3-27}$$

$A_{max_{stand}}$ was calculated as the average of the corrected absorbances of the last six data points plus the product of $m_{ss_{stand}}$ and temperature difference between the last and the third from the last data points:

$$A_{max_{stand}} = \frac{1}{6}\sum_{i=n-5}^{n} A(i) - m_{ss_{stand}}[T(n) - T(n-2)] \tag{3-28}$$

Starting values were assigned for $\Delta H°$ and $\Delta S°$. Two arbitrarily chosen values, -55.0 kcal/mol and -150 eu were assigned to $\Delta H°$ and $\Delta S°$.

The lowest and highest temperatures were assigned for T_{min} and T_{max}.

At this stage eight parameters in (3-21) have been assigned values: C_T, E, T_{min}, T_{max}, m_{ds}, m_{ss}, A_{min} and A_{max}, and the two queried parameters $\Delta H°$ and $\Delta S°$ were assigned starting. Least square method was used to fit the derivative versus temperature curve. The best-fit values of $\Delta H°$ and $\Delta S°$ from standard baseline fitting were denoted as $\Delta H°_{stand}$ and $\Delta S°_{stand}$. For 5PZ_1, $\Delta H°_{stand}$ and $\Delta S°_{stand}$ are -54.588 kcal/mole and 146.934 eu.

The derivative residual was the difference between the optimal-solution-derived derivative and the experimental-data-derived derivative. The sum of squared residuals, *rss*, was the sum of squared residuals. The *residual* and *rss* from standard baseline fitting were denoted as $residual_{stand}$ and rss_{stand}. Both were calculated as references for goodness of fit.

The maximum and minimum possible values for $\Delta H°$ and $\Delta S°$ were obtained from the 95% confidence interval. A Matlab built-in function using non-linear least

squares method was applied. The minimum and maximum values were denoted as $\Delta H°_{stand_{min}}$, $\Delta H°_{stand_{max}}$, $\Delta S°_{stand_{min}}$, and $\Delta S°_{stand_{max}}$.

3.3.2 Fitting with optimized baseline

New boundaries for double-stranded and single-stranded regions were refined using fraction in double-stranded form β as reference, more specifically the predicted fraction in double-stranded form β_{pred}. Best-fit values from standard-baseline fitting were used to calculate β_{pred} according to (3-21), and therefore was denoted as $\beta_{pred_{stand}}$. Every temperature corresponded to a $\beta_{pred_{stand}}$ value. The double-stranded region was defined as the set of points with $\beta_{pred_{stand}}$ larger than 0.99; the single-stranded region was defined as the aggregate of points with $\beta_{pred_{stand}}$ smaller than 0.01.

For the second time the four shape-depending parameters were calculated, and denoted as $m_{ds_{opt}}$, $m_{ss_{opt}}$, $A_{min_{opt}}$, and $A_{max_{opt}}$.

When the number of data points in double-stranded region was smaller than eight, it was considered that the double-stranded baseline was too short, and $m_{ds_{opt}}$ was calculated as twice as the average of the derivatives in the first five data points, and $A_{min_{opt}}$ was an extrapolation calculated as the average of the first five corrected absorbances minus the product of $m_{ds_{opt}}$ and the temperature difference between the fifth and the first data points:

$$m_{ds_{opt}} = \frac{2}{5}\sum_{i=1}^{5} Ader(i) \tag{3-29}$$

$$A_{min_{opt}} = \frac{1}{5}\sum_{i=1}^{5} A(i) - m_{ds_{opt}}[T(5) - T(1)] \tag{3-30}$$

Otherwise, $m_{ds_{opt}}$ and $m_{ss_{opt}}$ were calculated as twice as the average derivatives in the single-stranded region and double-stranded region, both redefined according to β_{pred} as described before:

$$m_{ds_{opt}} = \frac{2}{j}\sum_{i=1}^{j} Ader(i) \tag{3-31}$$

$$m_{ss_{opt}} = \frac{2}{k}\sum_{i=n-k+1}^{n} Ader(i) \tag{3-32}$$

where j and k are numbers of data points in double-stranded and single-stranded regions. A_{min} and A_{max} were calculated in the same way as they were calculated in the standard baseline fitting process, with renewed m_{ds}, m_{ss} and i values:

$$A_{min_{opt}} = \frac{1}{5}\sum_{i=1}^{5} A(i) - m_{ds_{opt}}[T(3) - T(1)] \tag{3-33}$$

$$A_{max_{opt}} = \frac{1}{6}\sum_{i=n-5}^{n} A(i) - m_{ss_{opt}}[T(n) - T(n-2)] \tag{3-34}$$

Starting values were again assigned for $\Delta H°$ and $\Delta S°$. The best-fit values of $\Delta H°$ and $\Delta S°$ from standard baseline fitting $\Delta H°_{stand}$ and $\Delta S°_{stand}$ were used.

At this stage the 6 parameters just discussed have been assigned new values or new starting values, and the values of C_T, E, T_{min} and T_{max} remain the same as those calculated in the standard baseline fitting process. A Matlab built-in least square method was used to fit the derivative versus temperature curve. The best-fit values of $\Delta H°$ and $\Delta S°$ from optimized baseline fitting were denoted as $\Delta H°_{opt}$ and $\Delta S°_{opt}$. Accordingly the values of rss_{opt} and $residual_{opt}$ were calculated.

The maximum and minimum values for ΔH°_{opt} and ΔS°_{opt} were again obtained as a 95% confidence interval. The minimum and maximum values were denoted as $\Delta H^\circ_{opt_{min}}$, $\Delta H^\circ_{opt_{max}}$, $\Delta S^\circ_{opt_{min}}$, and $\Delta S^\circ_{opt_{max}}$.

3.3.3 Fitting with unrestrained baseline

The third round and final round of fitting did not restrain the boundaries of single-stranded or double-stranded region, that is the values of m_{ds} and m_{ss} were not pre-calculated. Instead, m_{ds}, m_{ss} were treated like ΔH° and ΔS° and were to be queried and solved during this last round of fitting. A_{min} and A_{max} were restrained however, because fitting that allows all six parameters to float does not work very well – the regenerated curves some times failed to converge, and the program was slowed down. The starting values of these four parameters were assigned as $m_{ds_{opt}}$, $m_{ss_{opt}}$, ΔH°_{opt}, and ΔS°_{opt}. The values of A_{min} and A_{max} were given as the minimum and maximum absorbances calculated in the optimized baseline fitting process, $A_{min_{opt}}$ and $A_{max_{opt}}$. The values of C_T, E, T_{min} and T_{max} remain the same as those calculated in standard baseline fitting process.

At this stage the 10 parameters in (3-21) were assigned values or starting values. The Matlab built-in least square method was used to fit the derivative versus temperature curve. The optimal solution returned a group of best-fit values for m_{ds}, m_{ss}, ΔH° and ΔS°, denoted as $m_{ds_{unrest}}$, $m_{ss_{unrest}}$, ΔH°_{unrest} and ΔS°_{unrest} respectively. Accordingly the values of rss_{unrest} and $residual_{unrest}$ were calculated. The minimum and maximum possible values for $m_{ds_{unrest}}$, $m_{ss_{unrest}}$, ΔH°_{unrest} and ΔS°_{unrest} were obtained within 95% confidence interval.

3.3.4 Enthalpy, entropy, free energy change and experimental melting temperature

Enthalpy and entropy were the best-fit values given by the unrestrained baseline fitting process.

Free energy change at 37 °C was calculated using best-fit values of $\Delta H°$ and $\Delta S°$ according to the definition of Gibbs free energy

$$\Delta G° = \Delta H° - T\Delta S° \tag{3-35}$$

The T_M in terms of β is the temperature at which half of the meltable duplex is melted, so the temperature at which β=1/2.

$[WC]$, $[W]$ and $[C]$ can be expressed in terms of β, E and C_T

$$[WC] = \frac{\beta}{2}(C_T - E) \tag{3-36}$$

$$[W] = \frac{(1-\beta)C_T + (1+\beta)E}{2} \tag{3-37}$$

$$[C] = \frac{(1-\beta)C_T - (1-\beta)E}{2} \tag{3-38}$$

At β=1/2, the concentrations of the three species are

$$[WC] = \frac{1}{4}(C_T - E) \tag{3-39}$$

$$[W] = \frac{1}{4}(C_T + 3E) \tag{3-40}$$

$$[C] = \frac{1}{4}(C_T - E) \tag{3-41}$$

The equilibrium constant therefore is

$$K_{eq} = \frac{\frac{1}{4}(C_T - E)}{\frac{1}{4}(C_T + 3E)\frac{1}{4}(C_T - E)} = \frac{4}{C_T + 3E} \tag{3-42}$$

Accordingly the T_M is expressed in terms of $\Delta H°$, $\Delta S°$, C_T and E:

$$T_M = \frac{\Delta H°}{\Delta S° - R\ln K_{eq}} = \frac{\Delta H°}{\Delta S° - R\ln[(C_T + 3E)/4]} \tag{3-43}$$

3.4 Estimated uncertainties in thermodynamic parameters

The absolute associated uncertainties of $\Delta H°$ and $\Delta S°$ are defined as half of the difference between the minimum and maximum possible obtained with 95% confidence interval during unrestrained baseline fitting.

The uncertainties in $\Delta H°$ and $\Delta S°$ are correlated. The minimum and maximum possible values for $\Delta G°_{37}$ and T_M were obtained by using either the combination of the minimum possible values of $\Delta H°$ and $\Delta S°$, or the maximum possible values of them.

The relative associated uncertainty was calculated by dividing the absolute associated uncertainty by the best-fit value. For T_M only the absolute associated uncertainty was reported.

3.5 Melting Temperature at 1 μM C_T

T_M is a function of C_T. Because C_T varies from melt trial to melt trial, a tabulated *Tm* at a uniform C_T is desired. T_M at 1 μM C_T was calculated. Two methods are used to calculate this melting temperature. The first method first calculated equilibrium constant by plugging E=0 and C_T =1 μM C_T into equation (3-43):

$$T_M = \frac{\Delta H°}{\Delta S° - R ln \frac{4}{1\ \mu M}} \tag{3-44}$$

where $\Delta H°$ and $\Delta S°$ are derived from fitting.

The second method combined (3-43) and Vant' Hoff equation

$$\ln K_{eq1} - \ln K_{eq2} = -\frac{\Delta H°}{R} \cdot (\frac{1}{T_1} - \frac{1}{T_2}) \tag{3-45}$$

and derived the relation between the queried Tm and experimental T_M:

$$T_M = \frac{1}{\frac{1}{T_{M\,\beta=0.5}} - \frac{R}{\Delta H^\circ} \cdot \ln \frac{C_T + 3E}{1\ \mu\text{M}}} \tag{3-46}$$

where $T_{M\,\beta=0.5}$ and C_T are the experimental values.

The two methods reported the same results.

3.6 Regeneration of predicted curves

By the end of unrestrained fitting, best-fit values and estimated maximum and minimum values were obtained for m_{ds}, m_{ds}, ΔH° and ΔS°; the other six parameters had their values calculated either in standard baseline fitting (C_T, E, T_{min} and T_{max}), or in the optimized baseline fitting (A_{min} and A_{max}). Plugging these values back into corresponding equations completed predictions of theoretical values. Specifically absorbances versus temperature, absorbance derivatives versus temperature, and fraction in double-stranded form versus temperature were of interest and regenerated.

3.6.1 Regeneration of absorbance and absorbance derivative curves

The predicted absorbances were calculated according to (3-23), and the predicted absorbance derivatives were calculated according to (3-23) and (3-24). The low-limit, high-limit and the most possible predictions to absorbances and the derivatives were calculated using variant combinations of six parameters. The names of predicted absorbances and the derivatives, and their corresponding combinations of parameters are listed in Table 3.2.

Table 3.2 Parameters Used for Regeneration of Three Categories of Predicted Absorbances: the Most Possible (Apred), the High-limit (Apred_hi) and the Low-limit (Apred_lo); and the Corresponding Predicted Derivatives: Der, Der_hi, and Der_lo

	ΔH_{unrest}, ΔS_{unrest}	$m_{ds_{unrest}}$, $m_{ss_{unrest}}$	A_{min}, A_{max}
Apred; Dpred	Best-fit		
Apred_hi; Dpred_hi	Maximum	Best-fit	$A_{min_{opt}}$, $A_{max_{opt}}$
Apred_lo; Dpred_lo	Minimum		

3.6.2 Regeneration of fraction in double-stranded form curves

Fraction in double-stranded form can be derived in two ways. The predicted fraction in double-stranded form was calculated using thermodynamic parameters ($\Delta H°$ and $\Delta S°$) and reaction-related parameters (C_T and E) (3-21). The experimental fraction in double-stranded form was calculated using parameters related to melting-curve-shape (A_{min}, m_{ds}, A_{max} and m_{ss}) by rearranging to (3-23):

$$\beta_{expt} = \frac{A_{corr}(T) - [A_{max,corr} + 0.5m_{ss}(T - T_{max})]}{[A_{min,corr} + 0.5m_{ds}(T - T_{min})] - [A_{max,corr} + 0.5m_{ss}(T - T_{max})]} \quad (3\text{-}47)$$

The corresponding parameter values obtained from the unrestrained fitting were plugged into (3-21) and (3-47) to generate two β curves.

3.7 Parameters and Graphs in Fitting Process, using Trial 5PZ_1 as An Illustration

Table 3.3 Concentrations Parameters

Parameter	Values	Units
$C_{T\,\mathrm{sum}}$	8.765	μM
$C_{T\,\mathrm{est}}$	8.343	μM
C_T	8.765	μM
E	0.688	μM
f	0.1455	

Table 3.4 Fitting Parameters Determined in Standard Baseline Fitting, Optimized Baseline Fitting, and Unrestrained Baseline Fitting

Parameter	Values	Units
$m_{ds_{stand}}$	1.993198E-04	
$m_{ss_{stand}}$	7.276612E-04	
$A_{min_{stand}}$	5.331571E-01	
$A_{max_{stand}}$	6.656142E-01	
$\Delta H°_{stand}$	-54.588	kcal/mol
$\Delta S°_{stand}$	-146.934	e.u.
$\Delta H°_{stand_{min}}$	-53.345	kcal/mole
$\Delta H°_{stand_{max}}$	-55.830	kcal/mol
$\Delta S°_{stand_{min}}$	-143.026	e.u.
$\Delta S°_{stand_{max}}$	-150.842	e.u.
rss_{stand}	2.491827E-06	
$m_{ds_{opt}}$	4.552944E-04	
$m_{ss_{opt}}$	6.593096E-04	
$A_{min_{opt}}$	5.325683E-01	
$A_{max_{opt}}$	6.654603E-01	
$\Delta H°_{opt}$	-56.256	kcal/mole
$\Delta S°_{opt}$	-152.035	e.u.
$\Delta H°_{opt_{min}}$	-55.084	kcal/mole
$\Delta H°_{opt_{max}}$	-57.427	kcal/mol
$\Delta S°_{opt_{min}}$	-148.353	e.u.
$\Delta S°_{opt_{max}}$	-155.717	e.u.
rss_{opt}	2.0087E-06	
$m_{ds_{unrest}}$	4.780347E-04	
$m_{ss_{unrest}}$	5.626011E-04	
$\Delta H°_{unrest}$	-55.800	kcal/mole
$\Delta S°_{unrest}$	-150.548	e.u.
$\Delta H°_{unrest_{min}}$	-54.112	kcal/mole
$\Delta H°_{unrest_{max}}$	-57.488	kcal/mol
$\Delta S°_{unrest_{min}}$	-145.248	e.u.
$\Delta S°_{unrest_{max}}$	-155.848	e.u.
rss_{unrest}	1.9468E-06	

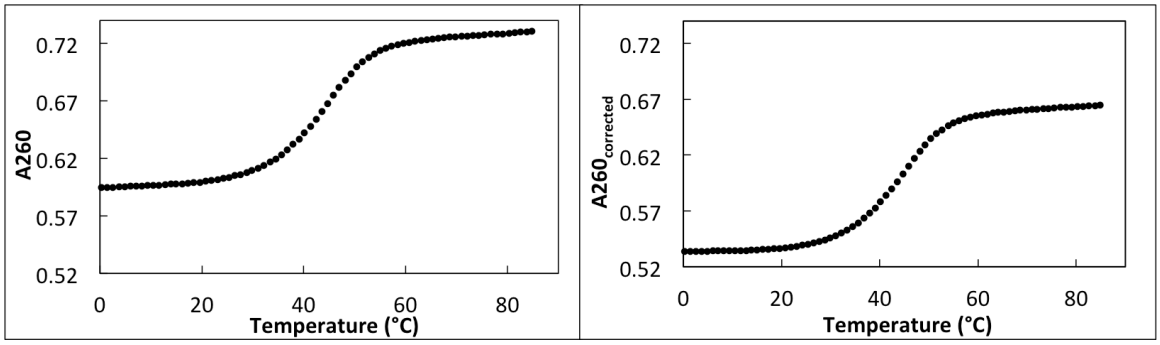


Figure 3-1. The melting sample absorbance measured experimentally (left) and the corrected absorbance produced by meltable strands (right), 5PZ_1.

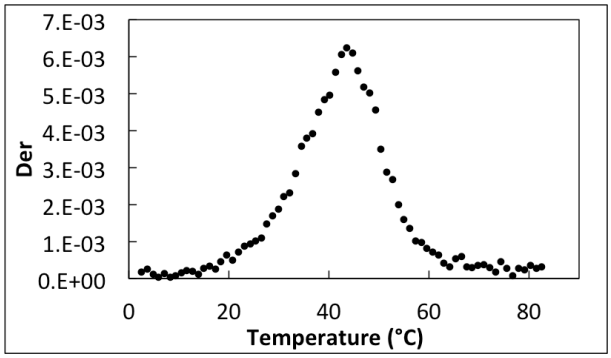


Figure 3-2. Corrected absorbance derivative versus temperature, 5PZ_1.

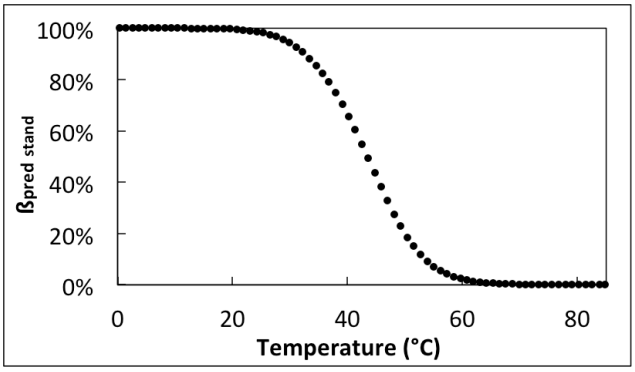


Figure 3-3. Predicted fraction in double-strand from calculated based on the best-fit values obtained from the standard baseline fitting, 5PZ_1.

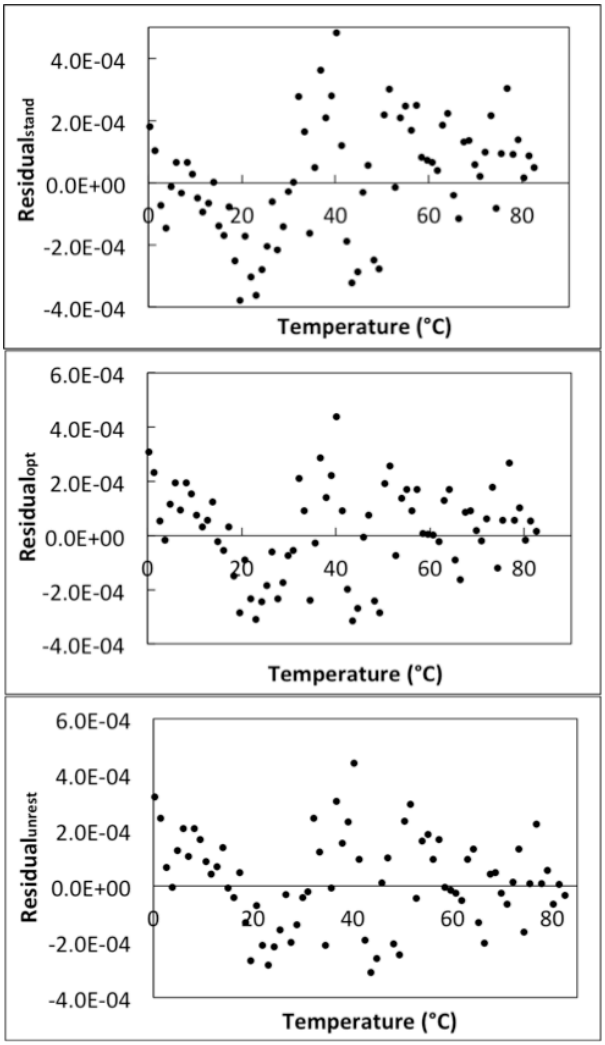


Figure 3-4. Derivative residuals from standard baseline fitting, 5PZ_1.

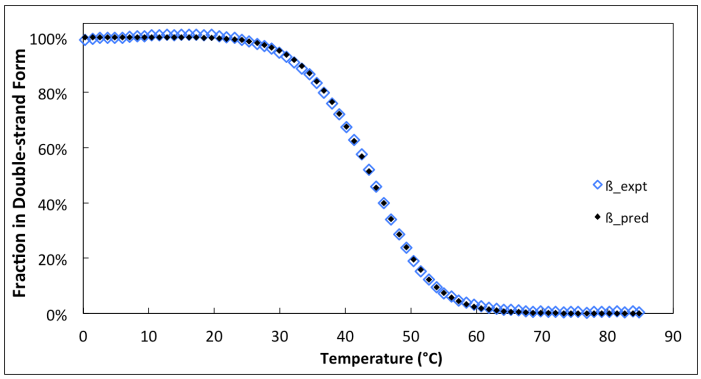


Figure 3-5. Experimental and predicted fraction in double-stranded form change as a function of temperature, 5ZP_1.

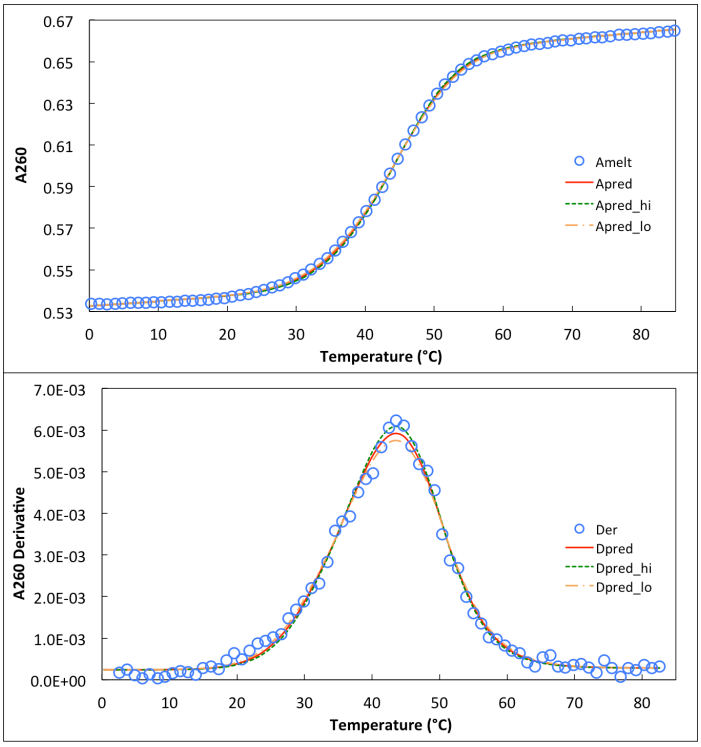


Figure 3-6. Experimental and predicted absorbances versus temperature (top), and absorbance derivatives versus temperature (bottom), 5ZP_1

3.8 Results

Enthalpy, entropy, free energy and melting temperature and their uncertainties of all trials are reported in Table 3.5.

Table 3.5: Best-fit Values and Associated Estimated Uncertainties of $\Delta H°$, $\Delta S°$ and $\Delta G°_{37}$, and Calculated *Tm* at 1 µM C_T of Melting Trials

Sequence	Trial	$\Delta H°$ (kcal/mol)			$\Delta S°$ (eu)			$\Delta G°_{37}$ (kcal/mol)			*Tm* (°C)		
		best-fit	est. uncert.		best-fit	est. uncert.		best-fit	est. uncert.		best-fit	est. uncert.	C_T=1 µM
			absolute	percent		absolute	percent		absolute	percent		absolute	
5PZ	5PZ _1	-55.80	1.69	3.03	-150.55	5.30	3.52	-9.107	0.044	0.48	43.86	0.04	35.55
	5PZ _2	-53.52	2.52	4.71	-143.40	7.94	5.54	-9.046	0.057	0.63	43.87	0.01	35.15
	5PZ _5	-59.55	2.48	4.16	-161.82	7.77	4.80	-9.360	0.071	0.76	43.40	0.11	36.96
3PZ	3PZ _1	-57.81	2.42	4.19	-156.13	7.59	4.86	-9.389	0.069	0.73	44.29	0.08	37.11
	3PZ _5	-59.09	2.39	4.04	-160.30	7.48	4.66	-9.371	0.070	0.75	43.63	0.11	37.01
2PZ	2PZ _1	-56.12	2.12	3.78	-151.28	6.66	4.40	-9.198	0.052	0.57	43.63	0.04	36.06
	2PZ_2	-63.18	3.41	5.40	-173.70	10.69	6.16	-9.310	0.089	0.96	42.54	0.15	36.71
	2PZ_5	-59.72	3.07	5.14	-162.69	9.60	5.90	-9.260	0.098	1.06	42.37	0.25	36.44
5GZ	5GZ_1	-49.74	2.52	5.07	-135.42	8.13	6.00	-7.740	0.004	0.05	34.98	0.07	27.17
	5GZ_2	-46.06	2.60	5.64	-122.77	8.35	6.80	-7.978	0.006	0.08	37.08	0.04	27.91
	5GZ_5	-47.94	2.09	4.36	-129.08	6.72	5.21	-7.900	0.005	0.06	35.94	0.08	27.78
3GZ	3GZ_5	-49.57	3.36	6.78	-137.05	10.89	7.94	-7.067	0.015	0.21	30.43	0.35	23.24
2GZ	2GZ_2	-47.22	3.69	7.81	-127.40	11.87	9.31	-7.705	0.012	0.16	33.62	0.34	26.45
	2GZ_3	-54.32	1.90	3.50	-151.24	6.06	4.01	-7.416	0.017	0.23	35.79	0.14	26.24
5PC	5PC_1	-57.37	2.53	4.41	-166.74	8.39	5.03	-5.652	0.072	1.27	23.58	0.21	18.13
	5PC_5	-59.92	2.70	4.51	-175.79	8.95	5.09	-5.398	0.080	1.48	23.38	0.21	17.72
3PC	3PC_2	-47.98	2.44	5.09	-136.57	8.14	5.96	-5.626	0.096	1.71	24.04	0.10	14.56
	3PC_3	-53.39	1.47	2.75	-154.11	4.86	3.15	-5.592	0.037	0.66	27.59	0.05	16.51
	3PC_4	-55.76	2.54	4.56	-161.48	8.46	5.24	-5.677	0.082	1.44	24.37	0.13	17.74
	3PC_5	-51.72	3.09	5.97	-148.18	10.27	6.93	-5.768	0.100	1.73	24.21	0.18	16.81
2PC	2PC_3	-48.38	1.32	2.73	-137.31	4.34	3.16	-5.797	0.030	0.52	27.86	0.06	15.68
	2PC_4	-54.36	3.90	7.17	-156.29	12.92	8.26	-5.887	0.103	1.75	24.53	0.32	18.33
	2PC_6	-50.08	2.62	5.23	-142.61	8.67	6.08	-5.847	0.073	1.25	23.25	0.28	16.62
5GT	5GT_1	-47.51	4.70	9.89	-135.47	15.61	11.52	-5.492	0.143	2.60	20.53	0.71	13.60

	5GT_5	-49.79	3.48	6.99	-143.04	11.60	8.11	-5.430	0.121	2.23	21.42	0.35	14.27
3GT	3GT_1	-45.47	4.62	10.16	-129.06	15.35	11.89	-5.444	0.137	2.52	19.96	0.81	12.36
	3GT_5	-45.76	4.44	9.70	-130.33	14.77	11.33	-5.335	0.139	2.61	18.98	0.82	11.87
2GT	2GT_1	-48.64	4.09	8.41	-137.83	13.52	9.81	-5.892	0.102	1.73	23.50	0.50	16.31
	2GT_6	-46.76	3.21	6.86	-131.86	10.66	8.08	-5.869	0.092	1.57	23.00	0.36	15.41
5PT	5PT_1	-50.30	3.66	7.28	-146.21	12.30	8.41	-4.950	0.151	3.05	19.77	0.36	11.95
	5PT_2	-62.61	4.58	7.32	-187.34	1.53	0.82	-4.503	0.016	0.36	19.47	0.05	14.64
	5PT_5	-60.47	3.98	6.58	-180.49	1.32	0.73	-4.493	0.013	0.29	19.28	0.05	13.86
3PT	3PT_1	-38.58	3.10	8.04	-109.17	10.39	9.52	-4.726	0.127	2.69	12.17*	0.97	3.69
	3PT_6	-40.28	2.09	5.19	-115.39	7.02	6.08	-4.494	0.091	2.02	11.75*	0.61	3.52
2PT	2PT_1	-48.04	4.35	9.05	-136.30	14.36	10.54	-5.767	0.106	1.84	22.40	0.64	15.37
	2PT_5	-44.85	1.97	4.39	-126.36	6.68	5.29	-5.654	0.101	1.79	20.50	0.06	13.28
5AZ	5AZ_1	-38.65	3.82	9.88	-109.66	12.90	11.76	-4.640	0.176	3.79	11.18*	1.16	3.19
	5AZ_5	-36.91	3.08	8.34	-103.52	10.36	10.01	-4.801	0.132	2.75	11.50*	1.02	2.84
3AZ	3AZ_6	-38.12	2.06	5.40	-108.54	6.99	6.44	-4.452	0.112	2.52	9.57**	0.59	1.56
2AZ	2AZ_1	-40.65	4.25	10.46	-113.53	14.12	12.43	-5.435	0.127	2.34	17.38	1.08	9.63
	2AZ_6	-41.48	2.21	5.33	-116.26	7.34	6.31	-5.423	0.069	1.27	17.01	0.54	10.06
5AC	5AC_1	-50.90	0.67	1.32	-152.91	2.24	1.47	-3.473	0.029	0.84	10.85*	0.16	4.80
	5AC_5	-37.89	1.57	4.14	-108.47	5.41	4.98	-4.244	0.107	2.52	8.21**	0.36	0.05
3CA	3AC_1†	-55.30	0.51	0.92	-166.67	1.35	0.81	-3.604	0.095	2.64	13.92*	0.66	7.72
	3AC_6†	-53.12	0.39	0.73	-162.25	1.08	0.67	-2.804	0.050	1.78	9.07**	0.43	2.89
2AC	2AC_1	-34.64	3.58	10.33	-94.84	11.95	12.61	-5.226	0.127	2.43	12.85*	1.35	3.87
	2AC_6	-34.99	2.09	5.97	-95.83	6.99	7.30	-5.273	0.074	1.40	13.04*	0.76	4.51
23PZ	23PZ_1	-60.46	1.67	2.76	-163.29	5.20	3.18	-9.812	0.057	0.58	46.14	0.05	39.29
	23PZ_2	-52.85	2.61	4.94	-140.55	8.20	5.83	-9.258	0.070	0.76	45.31	0.01	36.35
23GZ	23GZ_1	-42.87	2.30	5.37	-116.84	7.54	6.45	-6.633	0.041	0.62	27.59	0.21	18.40
	23GZ_3	-48.49	1.82	3.75	-135.43	5.91	4.36	-6.491	0.013	0.20	30.60	0.16	19.63
23PC	23PC_1†	-53.99	0.24	0.44	-165.72	0.87	0.52	-2.592	0.023	0.89	7.37**	0.01	2.42
	23PC_5†	-54.14	0.40	0.74	-164.77	1.21	0.73	-3.041	0.029	0.95	11.12*	0.32	4.55

23GT	23GT_1†	-54.97	0.24	0.44	-169.49	0.73	0.43	-2.401	0.018	0.75	8.01**	0.20	2.11
	23GT_5†	-52.80	0.05	0.09	-164.55	0.17	0.10	-1.762	0.003	0.17	3.46**	0.02	-2.06
23PT	23PT_1†	-53.13	0.02	0.04	-167.18	0.11	0.06	-1.282	0.008	0.62	3.01**	0.02	-3.96
	23PT_5†	-52.78	0.03	0.06	-165.82	0.18	0.11	-1.350	0.024	1.78	2.08**	0.09	-3.91
23AZ	23AZ_1†	-51.77	0.09	0.17	-160.60	0.31	0.19	-1.959	0.003	0.15	4.44**	0.04	-1.83
23AC	23AC_1‡	-52.18	0.15	0.29	-163.22	0.31	0.19	-1.558	0.053	3.40	3.18**	0.33	-3.38
GC	GC_1	-56.93	2.39	4.20	-156.56	7.60	4.85	-8.371	0.028	0.33	38.36	0.09	31.66
	GC_5	-60.98	2.94	4.82	-169.46	9.40	5.55	-8.426	0.027	0.32	38.79	0.05	32.28
Short-tails	5AC_1s	-42.64	3.56	8.34	-125.01	12.30	9.84	-3.869	0.257	6.64	8.64	0.61	1.57
	3AC_1s	-37.74	3.22	8.53	-109.39	11.60	10.60	-3.815	0.392	10.28	5.70	0.21	-2.78
	3AC_6s	-36.29	2.26	6.23	-104.93	8.30	7.91	-3.751	0.311	8.29	4.08	0.29	-4.57
	23PC_1s	-41.67	1.49	3.58	-123.29	5.40	4.38	-3.434	0.184	5.36	4.60	0.07	-1.66
	23PC_5s	-38.79	4.87	12.55	-113.12	17.60	15.56	-3.709	0.590	15.91	6.29	0.36	-2.49
	23GT_1s	-38.06	3.09	8.12	-111.23	11.30	10.16	-3.557	0.425	11.95	4.12	0.38	-4.09
	23GT_5s	-40.89	4.12	10.08	-121.96	15.20	12.46	-3.062	0.587	19.17	2.59	0.44	-4.44
	23PT_1s	-33.88	10.28	30.34	-97.26	37.80	38.86	-3.716	1.444	38.86	3.48	1.59	-7.34
	23PT_5s	-38.61	6.62	17.15	-114.22	24.40	21.36	-3.187	0.934	29.31	2.34	0.66	-5.80
	23AZ_1s	-37.93	4.92	12.97	-111.77	18.20	16.28	-3.267	0.706	21.61	2.39	0.58	-5.97

* Best-fit value of Tm lower than 15°C;

** Best-fit value of Tm lower than 5°C (more discussion in Chapter 3.3);

‡ Fitting failed.

3.9 Enthalpy, Entropy, Free Energy and Melting Temperature of Sequences

3.9.1 Weighted average

A sequence's enthalpy, entropy, and free energy were calculated as the weighted average of corresponding thermodynamic parameters derived from all the trials performed to this sequence. Each trial's weight was the reciprocal of the squared associated uncertainty (σ_i) obtained from the 95% confidence interval:

$$\omega_i = \frac{1}{{\sigma_i}^2} \tag{3-48}$$

The weighted average $\bar{x}$ is the sum of products of individual best-fit parameter and corresponding weight divided by the sum of individual weights:

$$\bar{x} = \frac{\sum x_i \omega_i}{\sum \omega_i} \tag{3-49}$$

All digits of best-fit values and estimated uncertainties obtained from fitting process were used in calculating ω_i and $\bar{x}$.

3.9.2 Propagated uncertainty and standard error of the mean (SEM)

Each trial produced a set of estimated uncertainties of $\Delta H°$, $\Delta S°$ and $\Delta G_{37}°$ (Table 3.1). Two methods were applied to calculate the estimated uncertainty for a sequence $\bar{\sigma}$: calculating the propagated uncertainty $\overline{\sigma_A}$, which is the square root of the quotient of the sum of squared individual estimated uncertainties divided by the number of trials:

$$\overline{\sigma_A} = \sqrt{\frac{\sum_{i=1}^{n} {\sigma_i}^2}{n}} \tag{3-50}$$

and calculating the standard error of the mean $\overline{\sigma_B}$, which is the quotient of standard deviation of best-fit values divided by the square root of number of trials

$$\overline{\sigma_B} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_{avg})^2}{n(n-1)}} \tag{3-51}$$

where x_{avg} is the arithmetic mean of trials' individual best-fit values.

The minimum and maximum uncertainties in enthalpies, entropies and free energies are 0.028 kcal/mol and 4.533 kcal/mole, 0.14 eu and 15.06 eu, and 0.003 kcal/mol and 0.155 kcal/mol. (Table 3.6).

There are limitations for both two methods. The propagated uncertainty ignores the differences between individual best-fit values. Under the circumstance that two or more trials produce very different best-fit values with similar estimated uncertainties, i.e. when individual trial was fitted with good precision while the consistency among multiple trials were poor presumably due to random errors, the propagated uncertainty underestimates the real error, and the SEM is a better estimation. The standard error of the mean ignores the differences between individual estimated uncertainties. Under the circumstance that two or more trials produce similar best-fit values however individual trials have very large estimated uncertainty, i.e. when fittings of individual trials report poor precision thus makes the seemingly good consistency among multiple trials meaningless, the SEM underestimates the real error, and the propagated uncertainty is a better estimation. Therefore the larger one of propagated uncertainty and SEM was reported as the final error.

3.9.3 Chi-square test

Chi-square (χ^2) test result suggested that choosing the larger one of propagated uncertainty and SEM as the final report was a reasonable decision.

$$\chi^2 = \sum_{i=1}^{m} \frac{(x_i - \bar{x})^2}{\sigma_i^2} \tag{3-52}$$

where m is the number of trials, x_i is the best-fit of individual trial, $\bar{x}$ is the weighted average, and σ_i is individual estimated uncertainty.

χ^2-distribution was calculated using a Microsoft Excel built-in function CHIDIST. χ^2 value and degree of freedom, which in this case is the number of trials minus one, were provided to CHIDIST, and corresponding χ^2-distribution value was returned.

χ^2-distribution was calculated for enthalpy, entropy and free energy of each sequence, unless only one trial was performed and χ^2-distribution could not be tabulated. Results are reported in Table 3.6.

Every trial with χ^2-distribution larger than 0.1 demonstrates a smaller SEM than propagated uncertainty. The fact that most trials fall into this category indicates that the uncertainty from fitting process is the main source of error. It was expected that trials with lower Tm would show larger estimated uncertainty from fitting, because the indefinite shape of the initial part of melting curve should increase the uncertainty when the data is fit within 95% confidence interval. The expectation was seen on results of short-tail version data, which is an evidence that the short-tail treatment is more proper for low Tm trials.

For enthalpy and entropy, every trial with χ^2-distribution smaller than 0.1 demonstrates a larger SEM than propagated uncertainty. Very few trials fall into this

category, which indicated that only a few trials had random error as the main source of the final error. Pipetting caused concentration error is one source of random error. Others errors could come from different UV light bulb status, composition of buffer, or differences between cuvettes.

3.9.4 Sequence melting temperature at 1μM C_T

Two methods were applied to calculate each sequence's Tm at 1 μM C_T. One was averaging each trial's Tm at 1 μM C_T (Table 3.5); the other was using sequence enthalpy, entropy and equation (3-43). The two methods produced very close values, and *Tm* calculated from sequence enthalpy and entropy was reported. Maximum and minimum possible sequence enthalpy and entropy values were calculated according to error range. Maximum and minimum *Tm* at 1 μM were calculated by plugging in the maximum possible sequence enthalpy and entropy or the minimum possible sequence enthalpy and entropy into equation (3-43). The differences between the maximum and the best estimated *Tm* values were always close to the differences between the best estimated and the minimum *Tm* values, therefore the averages of the two was reported (Table 3.6).

3.9.5 Results

Propagated uncertainty, SEM and χ^2-distribution of enthalpy, entropy and free energy for sequence are reported in Table 3.6, left.

Enthalpy, entropy and free energy, and melting temperature at 1 μM total concentration of each sequence are reported in Table 3.6, right.

Table 3.6: Propagated Uncertainty ($\overline{\sigma_A}$), Standard Error of the Mean ($\overline{\sigma_B}$),), and χ^2 Test Results; $\Delta H°$, $\Delta S°$, $\Delta G°_{37}$, and Tm at 1 µM C_T of Sequences

Sequence	$\Delta H°$ (kcal/mol)			$\Delta S°$ (e.u.)			$\Delta G°_{37}$ (kcal/mol)			$\Delta H°$ (kcal/mol)	$\Delta S°$ (e.u.)	$\Delta G°_{37}$ (kcal/mol)	*Tm* (°C)
	$\overline{\sigma_A}$	$\overline{\sigma_B}$	χ^2 distribution	$\overline{\sigma_A}$	$\overline{\sigma_B}$	χ^2 distribution	$\overline{\sigma_A}$	$\overline{\sigma_B}$	χ^2 distribution				
5PZ	2.26	1.76	0.22	7.1	5.4	0.49	0.058	0.096	3.82E-02[a]	-56.2±2.3	-151.6±7.1	-9.1±0.1	35.8±0.6
3PZ	2.41	0.64	0.71	7.5	2.1	0.70	0.070	0.009	0.85	-58.4±2.4	-158.2±7.5	-9.4±0.1	37.0±0.4
2PZ	2.92	2.04	0.19	9.1	6.5	0.43	0.082	0.032	0.73	-58.5±2.9	-158.9±9.1	-9.23±0.1	36.3±0.5
5GZ	2.41	1.06	0.60	7.8	3.6	0.75	0.005	0.070	3.04E-124[a]	-47.9±2.4	-129.2±7.8	-7.8±0.1	27.6±0.4
3GZ	3.36	NA	NA	10.9	NA	NA	0.015	NA	NA	-49.6±3.4	-137±10.9	-7.07±.02	23.5±1.0
2GZ	2.93	3.55	8.7E-02[a]	9.4	11.9	7.4E-02[a]	0.014	0.144	2.58E-46[a]	-52.8±3.6	-146±11.9	-7.6±0.1	26.2±0.2
5PC	2.61	1.28	0.49	8.7	4.5	0.46	0.076	0.127	1.81E-02[a]	-58.6±2.6	-171.0±8.7	-5.5±0.1	18.0±0.3
3PC	2.45	1.64	0.14	8.2	5.3	0.63	0.081	0.038	0.78	-52.6±2.4	-151.4±8.2	-5.6±0.1	16.4±0.7
2PC	2.82	1.78	0.33	9.3	5.7	0.60	0.075	0.026	0.77	-49.2±2.8	-139.8±9.3	-5.8±0.1	16.1±0.6
5GT	4.13	1.14	0.70	13.7	3.8	0.70	0.133	0.031	0.74	-49.0±4.1	-140.3±13.7	-5.4±0.1	14.2±1.0
3GT	4.53	0.14	0.96	15.1	0.6	0.95	0.138	0.055	0.57	-45.6±4.5	-129.7±15.1	-5.4±0.1	12.0±1.2
2GT	3.68	0.94	0.72	12.2	3.0	0.73	0.097	0.012	0.87	-47.5±3.7	-134.1±12.2	-5.9±0.1	15.9±1.0
5PT	2.14	3.80	2.2E-05[a]	7.2	12.7	5.8E-03[a]	0.088	0.151	0.10[a]	-61.3±3.8	-183.2±12.7	-4.5±0.2	14.2±0.7
3PT	2.64	0.85	0.65	8.9	2.9	0.65	0.111	0.116	0.14[a]	-39.8±2.6	-113.6±8.9	-4.6±0.1	3.6±1.0
2PT	3.38	1.60	0.50	11.2	5.0	0.53	0.104	0.056	0.44	-45.4±3.4	-128.1±11.2	-5.7±0.1	13.6±1.2
5AZ	3.47	0.87	0.72	11.7	3.1	0.71	0.156	0.081	0.47	-37.6±3.5	-105.9±11.7	-4.7±0.2	3.1±2.0
3AZ	2.06	NA	NA	7.0	NA	NA	0.112	NA	NA	-38.1±2.1	-108.5±7.0	-4.4±0.1	1.6±0.6
2AZ	3.39	0.42	0.86	11.3	1.4	0.86	0.102	0.006	0.93	-41.3±3.4	-115.7±11.3	-5.4±0.1	9.9±1.6
5AC†	1.21	6.51	2.4E-14[a]	4.1	22.2	3.1E-14[a]	0.078	0.385	3.52E-12[a]	-48.9±6.5	-146.4±22.2	-3.5±0.4	3.8±2.0
3AC†	0.45	1.09	7.3E-04[a]	1.2	2.2	1.1E-02[a]	0.076	0.400	1.25E-13[a]	-53.9±1.1	-164.0±2.2	-3.0±0.4	4.4±2.5
2AC	2.93	0.18	0.93	9.8	0.5	0.94	0.104	0.024	0.75	-34.9±2.9	-95.6±9.8	-5.2±0.1	4.3±1.4
23PZ	2.19	3.80	1.4E-02[a]	6.9	11.4	1.9E-02[a]	0.064	0.277	8.00E-10[a]	-58.2±3.8	-156.8±11.4	-9.6±0.3	38.4±1.4
23GZ	2.07	2.81	5.5E-02[a]	6.8	9.3	5.2E-02[a]	0.030	0.071	9.35E-04[a]	-46.3±2.8	-128.4±9.3	-6.5±0.1	19.0±0.5
23PC†	0.33	0.08	0.74	1.0	0.5	0.52	0.026	0.224	4.32E-33[a]	-54.0±0.3	-165.4±1.0	-2.8±0.2	3.1±0.3

23GT†	0.18	1.09	3.2E-18[a]	0.5	2.5	4.7E-11[a]	0.013	0.320	1.01E-282[a]	-52.9±1.1	-164.8±2.5	-1.8±0.3	-2.0±2.2
23PT†	0.03	0.18	1.2E-18[a]	0.1	0.7	8.4E-11[a]	0.018	0.034	8.39E-03[a]	-53.0±0.2	-166.8±0.7	-1.29±0.03	-4.2±0.1
23AZ†	0.09	NA	NA	0.3	NA	NA	0.003	NA	NA	-51.7±0.1	-160.6±0.3	-1.959±0.003	-1.8±0.1
23AC‡	0.15	NA	NA	0.3	NA	NA	0.053	NA	NA	-52.2±0.2	-163.2±0.3	-1.56±.05	-3.4±0.1
GC	2.68	2.03	2.80E-01	8.5	6.4	0.29	0.027	0.028	0.15	-58.5±2.7	-163.0±8.5	-8.40±.03	31.9±0.6
Short-tail Results													
5ACs	2.75	2.38	0.22	9.5	8.3	0.22	0.197	0.187	0.18	-38.7±2.8	-111.2±9.5	-4.2±0.2	0.3±0.7
3ACs	2.78	0.72	0.71	10.1	2.2	0.75	0.354	0.032	0.90	-36.8±2.8	-106.4±10.1	-3.8±0.4	-4.0±0.6
23PCs	3.60	1.44	0.57	13.0	5.1	0.58	0.437	0.137	0.66	-41.4±3.6	-122.4±13.0	-3.4±0.4	-1.8±0.5
23GTs	3.64	1.42	0.58	13.4	5.4	0.57	0.512	0.247	0.49	-39.0±3.6	-115.1±13.4	-3.4±0.5	-4.21±0.04
23PTs	8.65	2.36	0.70	31.8	8.5	0.71	1.216	0.265	0.76	-37.2±8.6	-109.2±31.8	-3.3±1.0	-6.2±0.9
23AZs	4.92	NA	NA	18.1	NA	NA	0.706	NA	NA	-37.9±4.9	-111.8±18.1	-3.3±0.7	-6.0±0.3

[a] SEM is larger than propagated uncertainty and χ^2 distribution is small;

† Alternative short-tail results available (more discussion see Chapter 3.3);

‡ Fitting failed.

3.10 Analysis to trials with melting temperature below 15 °C

3.10.1 Thoughts on reliability of fitted parameters

The very first motivation to investigate the reliability of fitting results derived from low-melting-temperature trials was the concern that the relatively short double-stranded baseline would lead to an inaccurate calculation of m_{ds}. As described in Chapter 3.1, the absorbance generated by meltable strands is expressed as a function of several parameters including m_{ds} and fraction in double-stranded form of meltable strands, the later one further expressed as a function of $\Delta H°$ and $\Delta S°$. During the standard baseline fitting, the number of data points used in calculation of m_{ds} is designated somewhat arbitrarily, and this $m_{ds_{stand}}$ is involved in obtaining the first set of best-fit values of enthalpy and entropy $\Delta H°_{stand}$ and $\Delta S°_{stand}$. In the optimized baseline fitting, $\Delta H°_{stand}$ and $\Delta S°_{stand}$ are used to calculate the prediction of fraction in double-stranded form, which decided the number of data points used in calculation of the optimized m_{ds}, i.e. $m_{ds_{opt}}$, and again this $m_{ds_{opt}}$ is further used in second round of fitting to obtain the best-fit of enthalpy and entropy $\Delta H°_{opt}$ and $\Delta S°_{opt}$. Eventually in the unrestrained baseline fitting, $\Delta H°_{opt}$ and $\Delta S°_{opt}$ are assigned to enthalpy and entropy as their starting values. The effect of a short or even missed double-stranded baseline on the fitting results, especially its impact on the intertwined m_{ds} and $\Delta H°$ and $\Delta S°$ is not clear.

The observation on fitting results of trials with Tm lower than 15 °C find that very short double-stranded baseline were present, and that trials with Tm lower than 10 °C almost did not show any baseline. In addition, many estimated uncertainties in $\Delta H°$ and $\Delta S°$ derived from these trials are unusually small (less than 1%, some are even less than

0.1%) compared to those derived from trials with higher Tm (normally around 5%). Further examination and treatment on these trials are discussed in Chapter 3.10.2 to 3.10.5.

3.10.2 Observation on goodness of fit and classification of trials with low Tm

The 9 trials with Tm between 15 °C and 10 °C are 3PT_1, 3PT_6, 5AZ_1, 5AZ_5, 5AC_1, 3AC_1, 2AC_1, 2AC_6, and 23PC_5; and the 10 trails with Tm lower than 10 °C are 3AZ_6, 5AC_5, 3AC_6, 23PC_1, 23GT_1, 23GT_5, 23PT_1, 23PT_5, 23AZ_1, and 23AC_1. Among the 9 trials with Tm between 15 °C and 10 °C, 6 showed satisfying goodness of fit in the result of unrestrained baseline fitting: the absorbance derivative residuals evenly separate above and below the zero line after two rounds of optimization (Figure 3-7 left); the predicted absorbances, absorbance derivatives, and fractions in double-stranded form all agree well with corresponding to experimental values (Figure 3-7 right). Trials with all above characters were classified as middle-low Tm trials and their best-fit values were considered reliable.

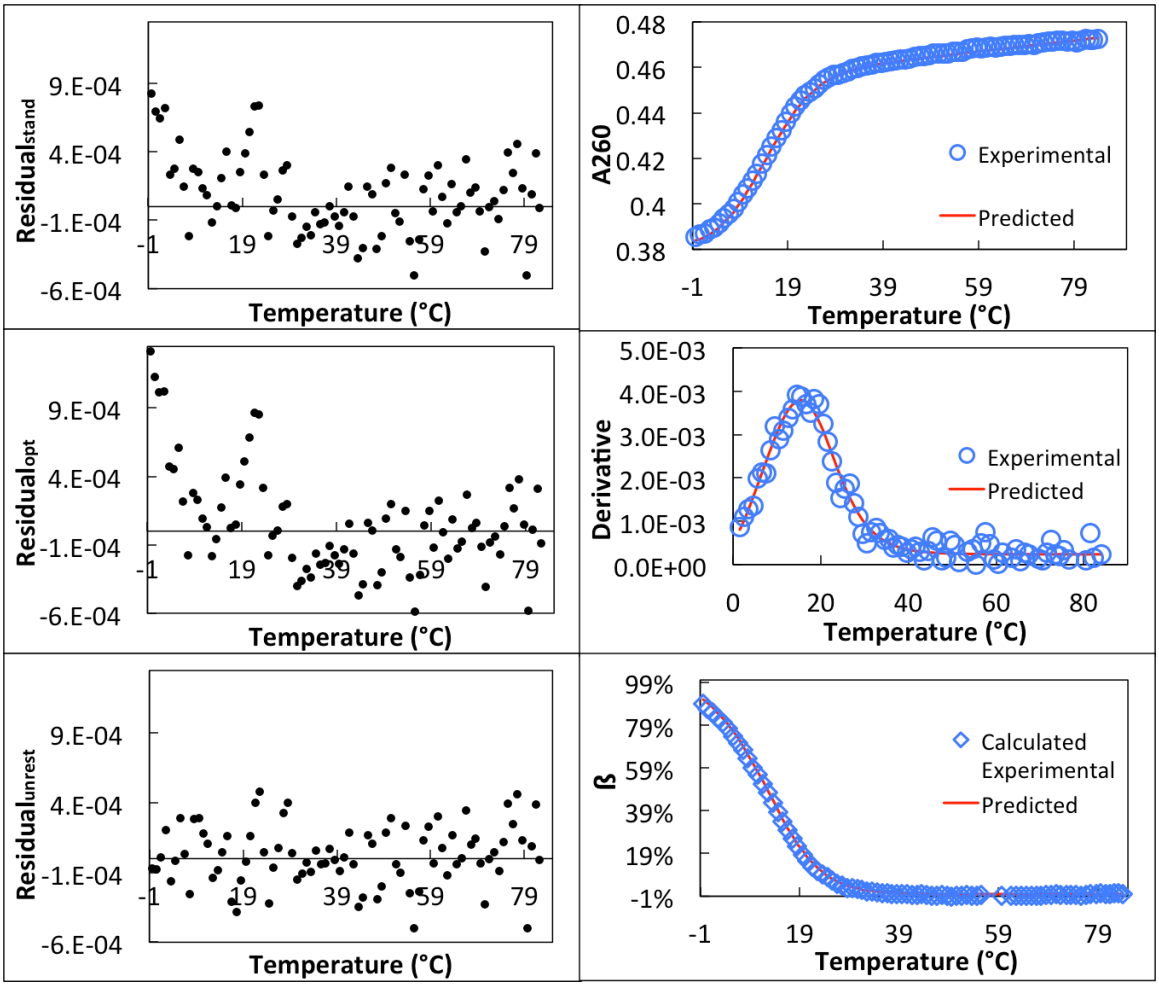


Figure 3-7. Absorbance derivative residuals in standard, optimized, and unrestrained baseline fitting (left top, middle and bottom), and predicted melting, derivative, and fraction in double-stranded curves presented with experimental data, 3PT_1. Increased randomness of residual distribution and good agreement between experimental data and simulation prediction were shown. This is a representation of fitting results with experimental Tm between 15°C to 10 °C.

3AC_1, 5AC_1 and 23PC_5 however showed poor goodness of fit. In the standard and optimized baseline fitting, although the absorbance derivative residual distributions were even in the middle and high temperature area, both remarkably bias towards the positive area in low temperature area; in the unrestrained baseline fitting, the distribution

was even in the low temperature area however bias towards the positive side in the middle and high temperature area (Figure 3-8, left). Obvious discrepancies exist between predicted absorbances and the experimental values (Figure 3-8, right top). The agreement between the predicted absorbance derivatives and the experimental values was acceptable (Figure 3-8, right middle). The calculated values of fraction in double-stranded form at several temperatures exceeded the theoretical range (0,1) dramatically, thus did not allow a complete overlap of the calculated experimental data over the predicted curve (Figure 3-8, right bottom). The examination to fitting results of the 10 trials with Tm lower than 10 °C proved that most of them - except for the two trials with the highest Tm, i.e. 3AZ_6 (Tm 9.57 °C) and 5AC_5 (Tm 8.21 °C), both showing middle-low Tm characters - shared the same characters with 3AC_1, 5AC_1 and 23PC_5. Trials with above characters were classified as low Tm trials and their best-fit values were considered not accurate.

Now the 19 trials with short double-stranded baseline or no baseline are assigned to two groups. The total 8 members in the middle-low Tm group are 3PT_1, 3PT_6, 5AZ_1, 5AZ_5, 2AC_1, 2AC_6, 3AZ_6 and 5AC_5, and the total 11 members in the low Tm group members are 3AC_1, 5AC_1, 23PC_5, 3AC_6, 23PC_1, 23GT_1, 23GT_5, 23PT_1, 23PT_5, 23AZ_1, and 23AC_1.

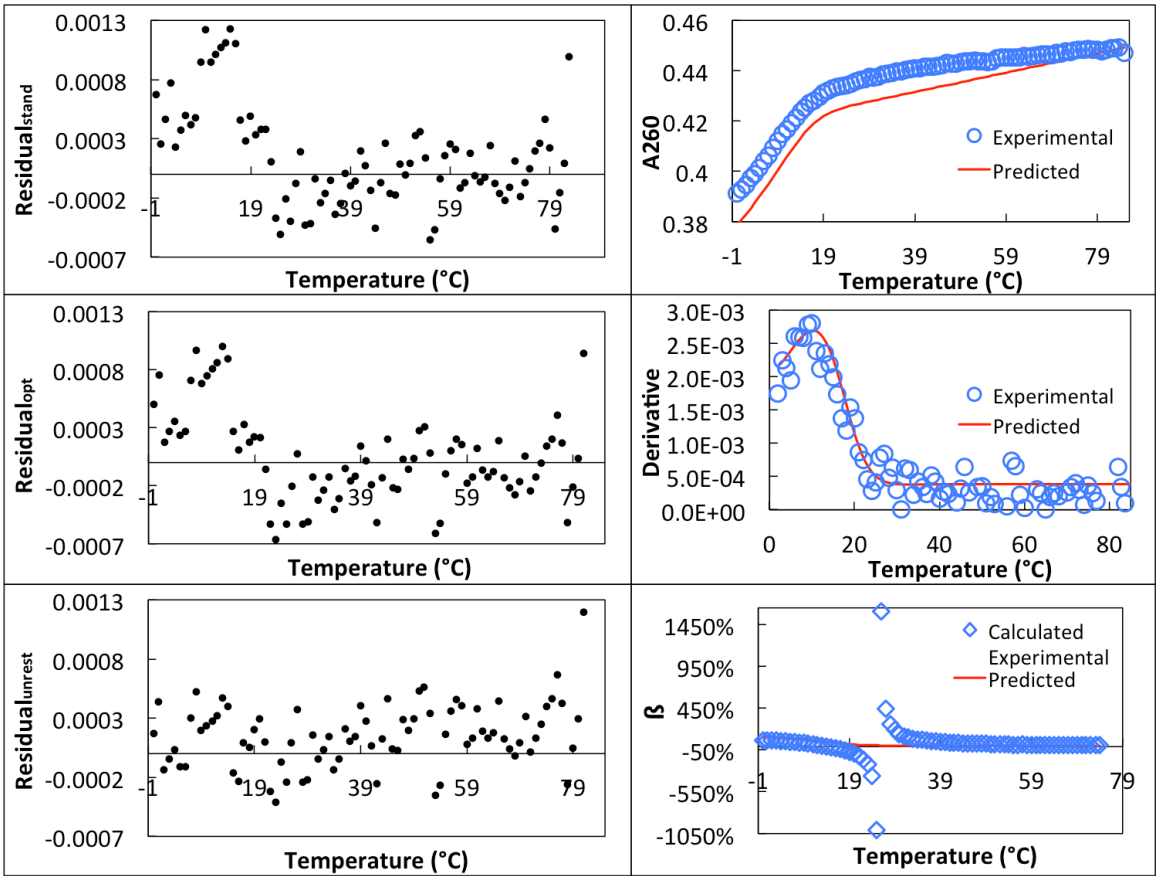


Figure 3-8. Absorbance derivative residuals (left, top to bottom: residual from standard baseline fitting, optimized baseline fitting, and unrestrained baseline fitting), and regenerated and experimental data: melting curve (right top), derivative curve (right middle), and fraction in double stranded form (right bottom) of 3AC_1. Optimal randomness of residual distribution in lower temperature area (below 19 °C) in the unrestrained fitting compared with that in the standard-baseline and the optimized-baseline fitting, with the residual distribution above 19 °C biased towards positive values; good agreement between experimental absorbance derivatives and predicted absorbance derivatives, and remarkable disagreement between experimental and predicted absorbances and fraction in double-stranded form.

3.10.3 Effect of removing low temperature points on analysis to high Tm trials

For the purpose of imitating trials with middle-low and low Tm, four randomly picked trials with higher Tm were treated by removing certain data points. The four trials are GC_5 (Tm 38.79 °C), 5PZ_1 (Tm 43.85 °C), 2GZ_2 (Tm 33.62 °C), and 3PC_4 (Tm 24.36°C). Removing some of the data points such that the temperatures of the remaining points were not smaller than (a) 15 °C, (b) 10 °C, (c) 5 °C, or (d) 0 °C below the experimental Tm. For example as for GC_5, data points with temperature lower than (a) 23.79 °C, (b) 28.79 °C, (c) 33.79 °C, and (d) 38.79 °C were removed as imitations to data sets that would generate Tm as (a) 15 °C, (b) 10 °C, (c) 5 °C or (d) 0 °C respectively.

First how Treatment (a) and (b) affect fitting results were examined as they generate the two extreme situations in middle-low Tm class. The goodness of fit of

Treatment (a) and (b) were not compromised: for all the four trials, the residual distribution was random, and the predicted physical quantities agreed well with experimental values (data not shown). The change in best-fit values of ΔH°, ΔS°, ΔG° and Tm were reported in Table 3.7. The best-fit values of ΔH°, ΔS°, ΔG° and Tm were generally smaller than best-fit values derived from complete data set, except for that treatment (b) causes an 0.02 kcal/mol increment in $\Delta G_{37}°$ to 3PC_4. The drop ranges of ΔH°, ΔS°, $\Delta G_{37}°$ and Tm are 2% to 21%, 2% to 21%, 2% to 6%, and 0.1°C to 3.0 °C. Therefore it is believed that fitting of the 8 middle-low Tm trials (3PT_1, 3PT_6, 5AZ_1, 5AZ_5, 2AC_1, 2AC_6, 3AZ_6 and 5AC_5) produced reliable $\Delta G_{37}°$ and Tm, and there may be some error in ΔH° and ΔS°.

Table 3.7: Change in Thermodynamic Parameters by Treatment of (a) to (d).

Trial	-ΔH° difference (%)				-ΔS° difference (%)				-ΔG°$_{37}$ difference (%)				Tm difference (°C)			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
GC_5	-14	-19	-7	-9	-15	-21	-7	-11	-4	-6	-4	3	-1.5	-3.0	-1.7	1.7
5PZ_1	-12	-14	2	5	-14	-15	2	4	-4	-6	-3	5	-1.6	-2.5	-1.4	2.3
2GZ_2	-11	-8	-1	-18	-12	-9	-1	-24	-3	-2	1	7	-1.9	-1.4	0.3	3.8
3PC_4	-2	-17	-3	23	-2	-19	-3	25	0	-2	-4	-3	-0.1	-2.6	-1.6	1.6

It was expected that the fitting to data treated by Treatment (c) and (d) would produce quite different predicted absorbances than the experimental data, just like what happened to the low Tm data (Figure 3-8, right top and right bottom), because these two treatments were meant to produce severely truncated data sets. Surprisingly none of the four trials showed compromised goodness of fit; instead great agreement between predicted absorbance and the experimental data remained (Figure 3-9). The agreement between predicted and calculated experimental fraction in double-stranded form remained as well (Data not shown). Treatment (c) only affected the best-fit values of ΔH°, ΔS°, $\Delta G_{37}°$ and Tm within a small range, treatment (d) however caused a larger difference in ΔH°, ΔS° (Table 3.7). It was noticed that the best-fit values of $\Delta G_{37}°$ and Tm were not severely affected by neither of the two treatments. More specifically, compared with those derived from complete data set, ΔH°, ΔS°, $\Delta G_{37}°$ and Tm derived from treatment (c) differ by -7% to -1%, -7% to 2%, -4% to 1%, and -1.7 °C to 0.3 °C respectively, and from treatment (d) differ by -18% to 23%, -24% to 25%, -3% to 7%, and 1.6 °C to 3.8 °C respectively.

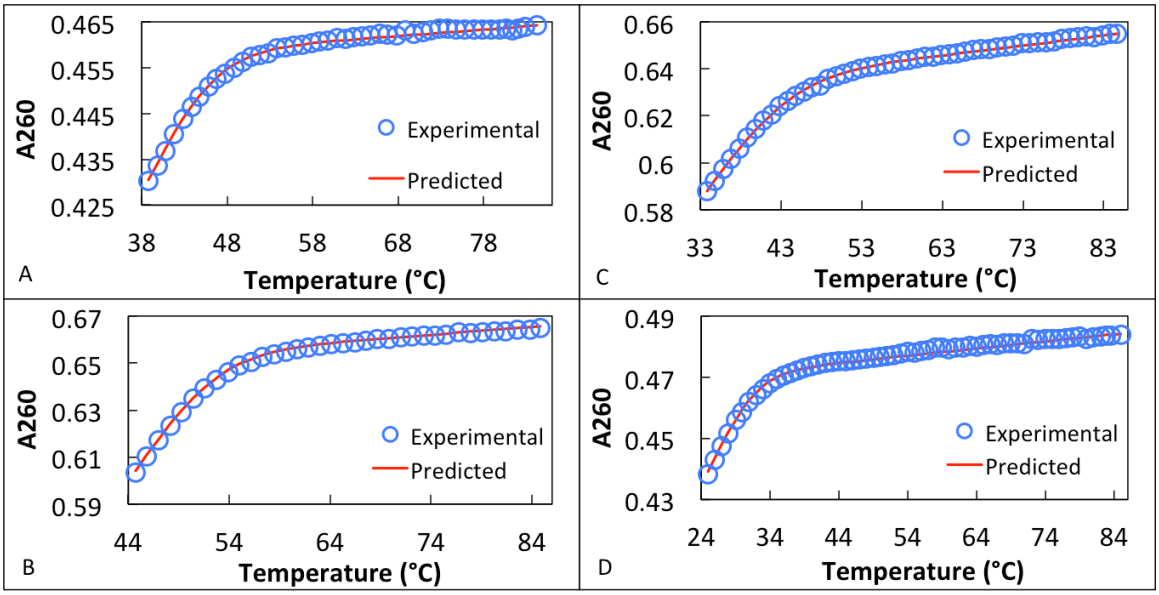


Figure 3-9. Agreement achieved between predicted absorbance curve and experimental data when data points before melting points were removed, Treatment (d). From A to D: GC_5, 5PZ_1, 2GZ_2, 3PC_4.

The associated estimated uncertainties of $\Delta G°_{37}$ derived from Treatment (a) to (d) ranged from 0.9% to 2.0%, 0.2% to 0.9%, 0.1% to 0.5%, and 0.7% to 2.0% respectively; and estimated uncertainties in Tm ranged from 0.6 to 0.9 °C, from 0.2 to 0.5 °C, 0.1 to 0.5 °C, and 0 to 1.6 °C respectively. Both parameters were fitted with high precision for individual trials treated by any of the four treatments. The estimated uncertainties in ΔH° and ΔS° from individual fittings of data treated by Treatment (a) to (c) were within the reasonable range (mostly around 5% with several exceptions of a little more than 10%), however Treatment (d) caused largely increased uncertainties in ΔH° and ΔS°: 39% uncertainty in ΔH° and 46% uncertainty in ΔS° for GC_5, 15% uncertainty in ΔH° and 17% uncertainty in ΔS° for 5PZ_1, 56% uncertainty in ΔH° and 72% uncertainty in ΔS° for 2GZ_2, and 1% uncertainty in ΔH° and 1% uncertainty in ΔS° for 3PC_4.

The conclusion was that when the data points before melting point were available over a temperature range smaller than 5 °C, the fitting process started to give wider range of estimated uncertainties in ΔH° and ΔS°, however the precision as well as accuracy in $\Delta G_{37}°$ and Tm was not compromised. Also, short double-stranded baseline was not the reason for poor goodness of fit as originally thought.

3.10.4 Effect of removing high temperature points on analysis to low Tm trials

Considering that absorbance versus temperature data were collected until 85 °C for all trials, the temperature range on which absorbance was monitored after melting point was small to high Tm trials compared with that to low Tm trials. That is the higher the Tm was, the shorter the single-stranded plateau was. Four example, among the four high Tm trials mentioned above, the melting temperature of 5PZ_1 was 43.8 °C, and absorbances were recorded over a temperature range of about 41 °C after the melting point until the last data point. Trials with Tm lower than 10 °C had absorbances recorded over a temperature range larger than 75 °C, which exceeded 41°C by 34 °C. As a result, the single-stranded plateau in trials with Tm lower than 10 °C was almost twice as long as in 5PZ_1. It was conceivable that the poor goodness of fit for low Tm trials is a consequence of long single-stranded plateau.

To clarify the effect of long single-stranded plateau in fitting process, the 11 trials in low Tm group were treated by removing part of the high temperature data points, and this data version was termed Short-tail. Any data points with temperature more than 40 °C than Tm were removed. For example, for 3AC_1 (Tm 13.9 °C), all the data points with temperature higher than 53.9 °C were removed. Except for 23AC_1 (fitting failed), all

the rest 10 trials showed improved goodness of fit by fitting Short-tail version. The result of short-tail 3AC_1 was shown as a demonstration (Figure 3-10).

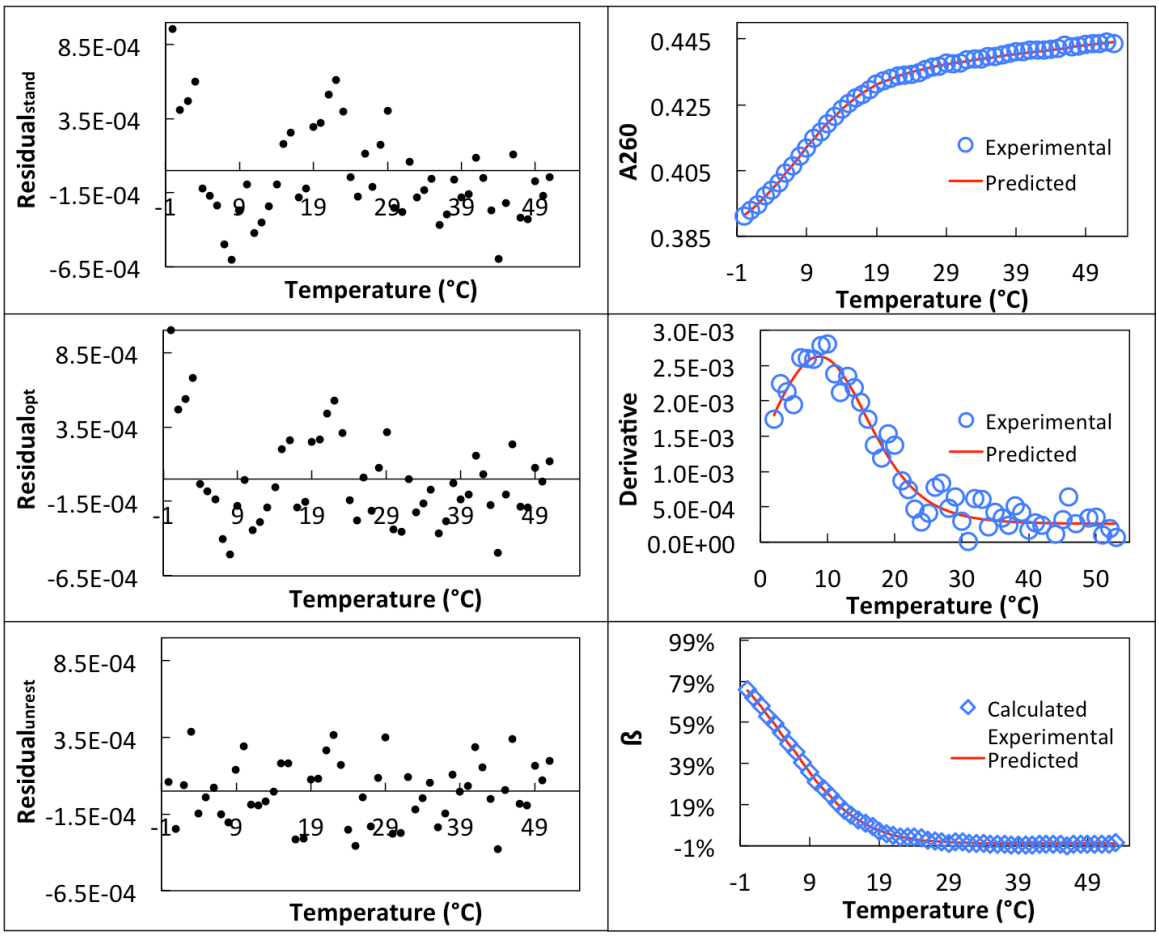


Figure 3-10 Fitting results of 3AC_1, short-tail version.

Several features were noticeable as the indication of improved goodness of fit. The absorbance derivative residuals in the unrestrained fitting process spread evenly above and below the zero line, and previous bias towards the positive area no longer existed (Figure 3-10 left). The agreement between the predicted absorbance derivatives and the experimental values remained (Figure 3-10 right middle). The agreement between predicted absorbances and the experimental values was satisfying (Figure 3-10, right top) and much better than the result from the complete data set. The experimental fractions in

double-stranded form, which was calculated based on values characterizing the shape of melting curve, including m_{ds} and m_{ss}, was within (0,1) over the whole fitting temperature range, and agreed very well with the predicted values, which was regenerated from thermodynamic parameters and physical amounts characterizing the chemical reaction, including $\Delta H°$ and $\Delta S°$ (Figure 3-10, right bottom). These features demonstrated significant improvement on goodness of fit.

3.10.5 Discussion

The fitting results from complete data set typically report both large enthalpy and large entropy change, and a combinatory small free energy change for low Tm trials. More specifically, enthalpy and entropy range from -50.90 kcal/mol to -55.30 kcal/mol and -152.9 eu to -169.5 eu, compared to relatively small enthalpy and entropy change (below -50 kcal/mol and -150 eu) of many trials with Tm around 20 °C. The short-tail data produces significantly smaller enthalpies and entropies and to various extents enlarged free energies. The drop range of enthalpy and entropy were 19% to 57% and 22% to 72%, and the free energy increased by 6% to 65%. Tm changed from -8.2 °C to 0.5 °C. The goodness of fit was significantly improved to 10/11 low Tm trials.

The estimated uncertainties in $\Delta H°$, $\Delta S°$ and $\Delta G°_{37}$ derived from complete data set are uncommonly small (mostly less than 1%, some even less than 0.1%). Short-tail data report significantly larger uncertainties: 4% to 10% for four trials (5AC_1, 3AC_1, 3AC_6, 23PC_1), and more than 10% uncertainty in at least one of $\Delta H°$, $\Delta S°$ and $\Delta G°_{37}$ for the rest of six trials. Compared with higher Tm trials, these uncertainty ranges are large; however considering these low Tm trials provide very limited information on absorbances before melting temperature, it is not surprising that the fitting process is not

able to provide more precise results. The larger range of uncertainties in enthalpy and entropy are also seen in Treatment (d) fitting results to the four higher Tm trials (Table 3.7). The uncommonly small uncertainties derived from complete data (with poor goodness of fit) might have been resultant from small uncertainties in m_{ss}, which is calculated from long however non-linear single-stranded plateau and is not accurate. The Tm uncertainties however remain small and are smaller than 1 °C for all trials except for 23PT_1 (Tm uncertainty is 1.6 °C)

The model considers a melting curve possesses a homogeneous complete single-stranded area with consistent slope, the value of slope being half of m_{ss}. In fact, when the single-stranded plateau becomes long, the relationship between absorbance and temperature is not overall linear, leading to inconsistent slopes. For the 10 trials processed by short-tail treatment, the m_{ss} values derived from complete data and short-tail data differ from -48% to 27%, with an average of -20%. The model needs certain modification in order to accurately describe the behavior of absorbance change in accordance with temperature change under the circumstance of none-linear single-stranded plateau. From the data aspect, the reason for poor goodness of fit to low Tm trials is not a too short double-stranded baseline, but a too long and none-linear single-stranded plateau. Conceivably, to improve the precision of ΔH°, ΔS°, and ΔG°, more data points before melting point is required.

The best-fit values and estimated uncertainties derived from short-tail data were reported in corresponding tables. ΔH°, ΔS°, ΔG°, and Tm were graphed according to categories of base pairs or mismatches (Figure 3-12). Parameters derived from original

data and short-tail data were both drawn, and error bars were only labeled for short-tail result if there was one.

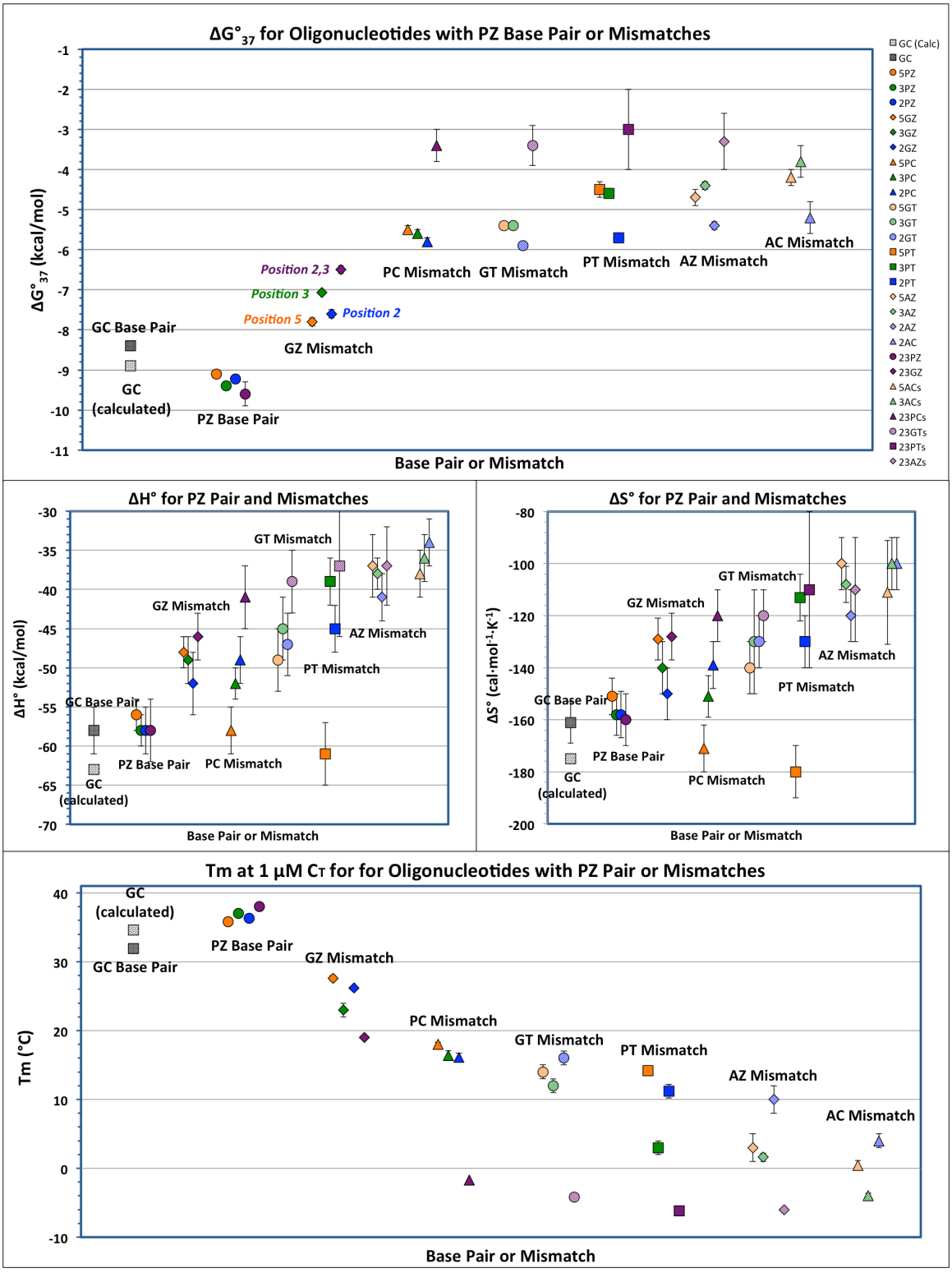


Figure 3-11. Thermodynamic parameters of base pairs and mismatches.

Chapter 4 Discussion, Application and Future Directions

4.1 Nearest Neighbor Parameters

Nearest-neighbor model considers the free energy change of duplex formation is the sum of the three terms: (1) an entropy penalty for the loss of translational freedom associated to formation of the first hydrogen bonded base pair, which is the initiation free energy; (2) the sum of enthalpy and entropy contributions for base stack formation between adjacent base pairs; and (3) an entropy penalty for self-complementary sequences pertaining to their C2 symmetry (SantaLucia, Allawi and Seneviratne, 1996). According to this model, the free energy change of the reference GC sequence duplex formation is

```
5'GCCAGTTAA= initiation + symmetry + GC + GG + CA + CT + GT + AA + TA + AA
3'CGGTCAATT                         CG   CC   GT   GA   CA   TT   AT   TT
```

Enthalpy and entropy changes are needed to predict the stability of the sequence. Nearest-neighbor thermodynamic parameters for all 10 Watson-Crick base pair have been obtained by Allawi and SantaLucia (1997) (Table 4.1).

Table 4.1 Nearest-Neighbor Thermodynamic Parameters for Watson-Crick Base Pair Formation in 1 M NaCl (Allawi and SantaLucia, 1997)

Propagation Sequence	ΔH° (kcal/mol)	ΔS° (eu)	$\Delta G°_{37}$ (kcal/mol)
AA/TT	-7.9 ± 0.2	-22.2 ± 0.8	-1.00 ± 0.01
AT/TA	-7.2 ± 0.7	-20.4 ± 2.4	-0.88 ± 0.04
TA/AT	-7.2 ± 0.9	-21.3 ± 2.4	-0.58 ± 0.06
CA/GT	-8.5 ± 0.6	-22.7 ± 2.0	-1.45 ± 0.06
GT/CA	-8.4 ± 0.5	-22.4 ± 2.0	-1.44 ± 0.04
CT/GA	-7.8 ± 0.6	-21.0 ± 2.0	-1.28 ± 0.03
GA/CT	-8.2 ± 0.6	-22.2 ± 1.7	-1.30 ± 0.03
CG/GC	-10.6 ± 0.6	-27.2 ± 2.6	-2.17 ± 0.05
GC/CG	-9.8 ± 0.4	-24.4 ± 2.0	-2.24 ± 0.03
GG/CC	-8.0 ± 0.9	-19.9 ± 1.8	-1.84 ± 0.04
init. w/term. G-C	0.1 ± 1.1	-2.8 ± 0.2	0.98 ± 0.05
init. w/term. A-T	2.3 ± 1.3	4.1 ± 0.2	1.03 ± 0.05
symmetry correction	0	-1.4	0.4

Calculation of predicted enthalpy change and uncertainty in enthalpy change of the reference GC sequence using values in Table 4.1 is illustrated below.

$\Delta H°$ = (0.1)+(2.3)+(0)+(-9.8)+(-8.0)+(-8.5)+(-7.8)+(-8.4)+(-7.9)+(-7.2)+(-7.9) kcal/mol

= -63.1 kcal/mol

Uncertainty

$= \sqrt{1.1^2 + 1.3^2 + 0 + 0.4^2 + 0.9^2 + 0.6^2 + 0.6^2 + 0.5^2 + 0.2^2 + 0.9^2 + 0.2^2}$ kcal/mol

= 2.394 kcal/mol

The predicted entropy and free energy change of the reference GC sequence were calculated using the same method quoting corresponding entropy and free energy change values from Table 4.1. The predicted entropy change is -174.8 eu, and the predicted free energy change -8.82 kcal/mol.

4.1.1 G·C, A·C, and G.T sequences: comparison of experimental vs predicted thermodynamics

In addition to the 10 Watson-Crick base pairs, thermodynamic parameters of nearest neighbors for 8 A·C mismatch nearest neighbors (Allawi and SantaLucia, 1998b) and 11 G·T mismatch nearest neighbors (SantaLucia, Allawi and Seneviratne, 1997) were obtained (Table 4.2).

Table 4.2 Nearest-neighbor Thermodynamics of A·C Mismatches (left) and G·T Mismatches (right) in 1 M NaCl (Allawi and SantaLucia, 1998b; SantaLucia, Allawi and Seneviratne, 1997)

Propagation Sequence	ΔH° (kcal/mol)	ΔS° (eu)	$\Delta G°_{37}$ (kcal/mol)	Propagation Sequence	ΔH° (kcal/mol)	ΔS° (eu)	$\Delta G°_{37}$ (kcal/mol)
AA/TC	2.3	4.6	0.88	AG/TT	1.0	0.9	0.71
AC/TA	5.3	14.6	0.77	AT/TG	-2.5	-8.3	0.07
CA/GC	1.9	3.7	0.75	CG/GT	-4.1	-11.7	-0.47
CC/GA	0.6	-0.6	0.79	CT/GG	-2.8	-8	-0.32
GA/CC	5.2	14.2	0.81	GG/CT	3.3	10.4	0.08
GC/CA	-0.7	-3.8	0.47	GG/TT	5.8	16.3	0.74
TA/AC	3.4	8	0.92	GT/CG	-4.4	-12.3	-0.59
TC/AA	7.6	20.2	1.33	GT/TG	4.1	9.5	1.15
				TG/AT	-0.1	-1.7	0.43
				TG/GT	-1.4	-6.2	0.52
				TT/AG	-1.3	-5.3	0.34

All these parameters are used to calculate the predicted enthalpy, entropy and free energy change of (1) the reference GC sequence, (2) the three position variants of A·C sequences, and (3) the all four position variations of G·T sequences. The predicted and experimental values are reported below (Table 4.3).

Table 4.3 Thermodynamic Parameters of G·C, A·C and G·T Containing Sequences

Sequence	Predicted			Experimental			
	ΔH° (kcal/mol)	ΔS° (eu)	$\Delta G°_{37}$ (kcal/mol)	ΔH° (kcal/mol)	ΔS° (eu)	$\Delta G°_{37}$ (kcal/mol)	Tm (°C)
GC	-63.1±2.4	-174.8±5.1	-8.82±0.13	-58±3	-161±8	-8.40±0.03	31.9±0.6
2AC	-40.8±2.2	-120.8±4.4	-3.46±0.12	-34±3	-100±10	-5.2±0.1	4±1
3AC	-42.6±2.1	-124.8±4.4	-3.82±0.11	-36±3	-100±10	-3.8±0.4	-4.0±0.6
5AC	-39.3±2.3	-112.2±4.3	-4.45±0.12	-38±3	-111±9	-4.2±0.2	0.4±0.7
2GT	-46.4±2.2	-132.4±4.4	-5.25±0.12	-47±4	-130±10	-5.9±0.1	16±1
3GT	-50.7±2.1	-145.5±4.4	-5.51±0.11	-45±4	-130±20	-5.4±.1	12±1
5GT	-48.4±2.3	-138.8±4.3	-5.32±0.12	-49±4	-140±10	-5.4±.1	14±1
23GT	-39.2±2.3	-115.0±4.3	-3.46±0.12	-39±4	-120±10	-3.4±.5	-4.18±0.04

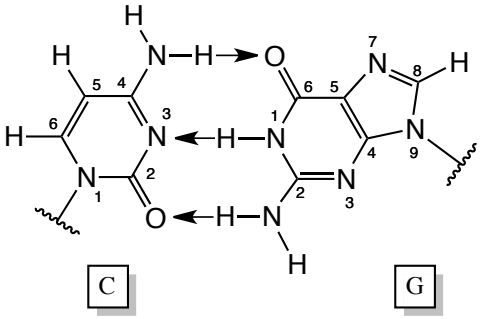
The percent differences of ΔH° and ΔS° range from -20% and 4%. The percent differences of $\Delta G°_{37}$ range from 5% to -50% with an average of 8%. The Tm differences range from -5.7 °C to 7 °C with an average of -0.4 °C. In general the results of this work are in good agreement with nearest neighbor prediction using SantaLucia's parameters. The best agreement is seen for 5AC among the three AC sequences, and is seen for 5GT

and 23GT among the four GT sequences. Compared to the prediction, significantly larger experimental $\Delta G°_{37}$ (-50% and -12% difference) and Tm (7 °C and 4 °C) for 2AC and 2GT are observed. This tendency is not surprising because Santa Lucia's parameters were obtained from internal A·C and G·T mismatches.

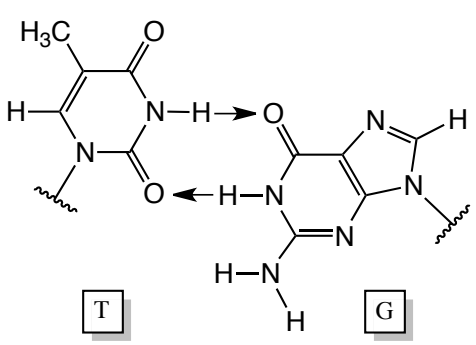
It was reported by SantaLucia that different thermodynamic characters were observed for terminal and internal G·T wobble pair, but only internal nearest neighbor thermodynamics were published (Allawi and SantaLucia, 1997). Our work shows that the Position 2 variant is the most stable position variant for all mismatches (P·C, G·T, P·T, A·Z and A·C). DNA end fraying is a reasonable explanation for this trend. The terminal base pairs in a double helix can be in non-hydrogen bonded status (Patel, 1974; Patel et al., 1982), which was termed fraying. Fraying can occur to the last three base pairs at the end of a duplex, with the opening extents decreasing from the most outside position to relatively internal positions (von Hippel, Johnson and Marcus, 2013). Base pair dissociation constants of the terminal pair during fraying have been measured (Kochoyan, Lancelot and Leroy, 1988; Nonin, Leroy and Gueron, 1995). It is conceivable that when a mismatch is located at an internal position, the destabilizing effects, i.e. their unfavorable base paring and stacking to adjacent base pairs, are fully dispersed into the whole sequence. On contrary, when a mismatch is located at a terminal position, fraying already reduces the base pairing and stacking contributions from these positions. The destabilizing effects produced by the mismatch are screened out to some extent and only partially influence the stability of the whole sequence.

4.2 Proposed Structures *for P·Z* Pair and M*ismatches*

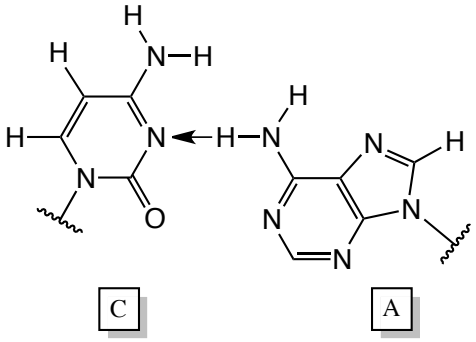
Structures of the G·C base pair, G·T mismatch and A·C mismatch are known from X-ray and NMR studies (Kalnik et al., 1988; Allawi, 1998a; Guo and Patel, 1987). G·T is an especially stable mismatch. Both G·T and A·C adopt wobble pair configuration: the two hydrogen bonds formed between G and T are guanine C6=O···HN3 thymine, and guanine N1H···O=C2 thymine; and the one hydrogen bond formed between A and C is adenine C5-NH_2···N3 cytosine.



G·C base pair



G·T wobble pair



A·C mismatch

Figure 4-1. Structures of G·C base pair, G·T wobble pair and A·C mismatch.

It was suggested that A·C is stabilized by a second hydrogen bond between protonated adenine N1 and cytosine C2 carbonyl group in acidic solution (Hunter et al., 1986). The G·T and A·C structures are drawn in Figure 4-1.

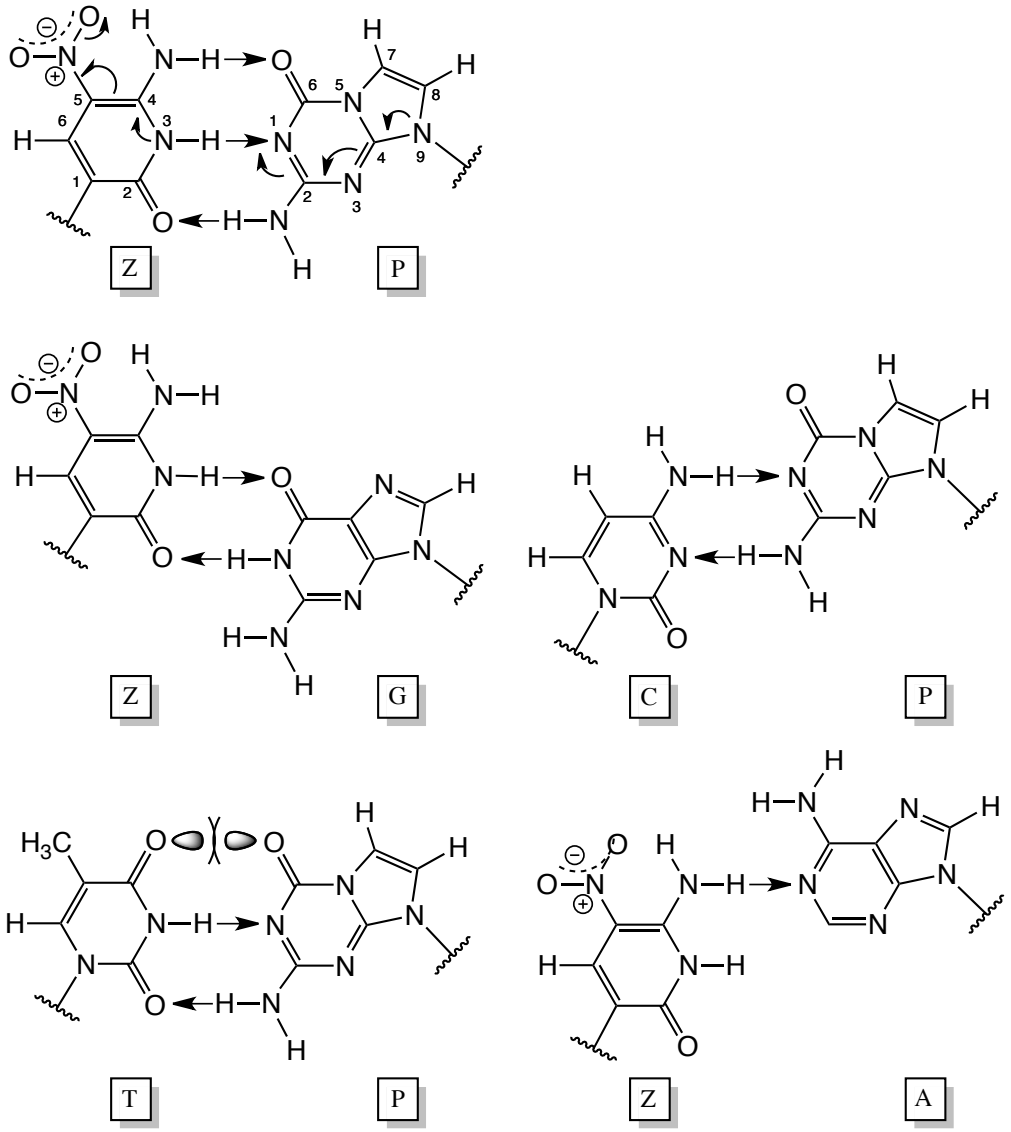


Figure 4-2. Proposed structures of the P·Z pair and mismatches containing P and Z; the stability trend: P·Z > G·C > G·Z > P·C ≈ G·T > P·T ≈ A·Z > A·C.

Averaged ΔG°_{37} values were calculated for P·Z base pair or for mismatches by averaging ΔG°_{37} of Position 2, Position 3 and Position 5 variants. The Position 2,3 variants are left out because they are doublets of base pairs or mismatches. The averaged ΔH°, ΔS° and ΔG°_{37} values for sequences containing P·Z, G·Z, P·C, G·T, P·T, A·Z, A·C and G·C are reported in Table 4.4. The averaged ΔG°_{37} is shown in Figure 4-3.

Table 4.4 Averaged Thermodynamic Parameters for P·Z Base Pair and Mismatches

	ΔH° (kcal/mol)	ΔS° (eu)	$\Delta G_{37}°$ (kcal/mol)	ΔΔH° (kcal/mol)	ΔΔS° (eu)	$\Delta\Delta G_{37}°$ (kcal/mol)
PZ	-57.7±4.4	-156.2±13.8	-9.2±0.2	0.8±4.4	6.8±13.8	-0.8±0.2
GZ	-50.1±5.5	-137.4±17.9	-7.5±0.1	8.4±5.5	25.6±17.9	0.9±0.1
PC	-53.5±4.5	-154.1±15.1	-5.6±0.2	5.0±4.5	8.9±15.1	2.8±0.2
GT	-47.4±7.1	-134.7±23.8	-5.6±0.2	11.1±7.1	28.3±23.8	2.8±0.2
PT	-48.8±5.7	-141.6±19.1	-4.9±0.2	9.7±5.7	21.4±19.1	3.5±0.2
AZ	-39.0±5.3	-110.0±17.7	-4.8±0.2	19.5±5.3	53.0±17.7	3.6±0.2
AC	-36.8±4.9	-104.4±17.0	-4.4±0.4	21.7±4.9	58.6±17.0	4.0±0.4
GC	-58.5±2.7	-163.0±8.0	-8.40±0.03	0±2.7	0±8.0	0±0.03

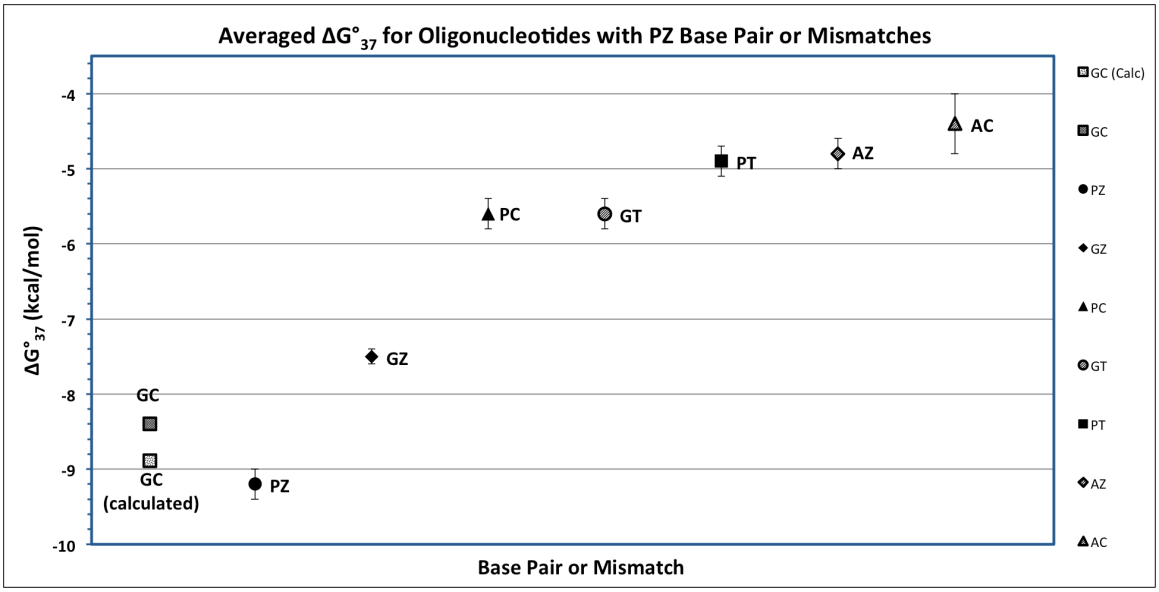


Figure 4-3. Averaged $\Delta G°_{37}$ of the three position variants for P·Z base pair or mismatches.

The general trend is that P·Z is more stable than G·C and both are the most stable base pairs; G·Z follows next and is more stable than P·C and G·T; P·C and G·T show almost the same stability and are more stable than P·T and A·Z; P·T and A·Z show almost the same stability and are slightly more stable than A·C. Compared to the G·C base pair, the increased stability caused by P·Z is entropic, and the decreased stabilities

caused by mismatches are largely entropic as well. Stacking might have played a role in stabilizing and destabilizing the whole sequence. The melting experiment alone however cannot separate the contributions from base paring and stacking. Obtaining UV profiles of sequences with P or Z dangling ends next to specific nucleotides should give more insight into stacking energy. We propose structures of the P·Z base pair and mismatches (Figure 4-2) according to known G·C, G·T and A·C structures. The structure of each one is discussed below.

P·Z base pair: the P·Z pair is more stable than G·C pair by -0.8 kcal/mol. P·Z pair has an electron-withdrawing nitro group attached to C5. This should cause the six-member ring's electron density move towards C4 and lead to a larger positive charge on the hydrogen attached to N3, thus a better hydrogen bond donor compared to the N1 hydrogen in Guanine. The lone pair of N9 in P can delocalize into the purine ring and confer a partial negative charge to N1, thus making N1 a better hydrogen bond acceptor compared to N3 in Cytosine, which should not be able to benefit in the same way from N1 electron delocalizaiton because the C2 carbonyl group is highly electron-withdrawing and the other way around in the ring there is no additional nitrogen between N1 and N3 like N3 is between N9 and N1 in P. The N3 hydrogen in Z may be deprotonated at higher pH, and melting experiments at high pH will probably find a lower Tm for oligonucleotides with P·Z pairs.

G·Z and P·C wobble pairs: the G·Z and the P·C mismatches are less stable than the G·C pair by +0.9 kcal/mol and +2.8 kcal/mol, respectively. The likely structures for G·Z and P·C are wobble pairs. The spatial configuration of G·Z would be very close to G·T, with Z intruding into the major groove. As discussed before, the N3 hydrogen in Z is a

better hydrogen donor, thus N3-H···O=C6 in Z·G is a stronger hydrogen bond than that in T·G, making G·Z a more stable wobble pair than G·T. The P·C wobble pair requires rotating the N9 glycosidic bond in P into the major groove, and presumably needs more energy than rotating N1 glycosidic bonds in pyrimidine because purine is more bulky than pyrimidine. It is possible that C4-NH···N1 in C·P is a stronger hydrogen bond than usual C4-NH···N1 bond therefore makes some compensation to the less favorable spatial configuration. Both structures would be expected to show two imino protons in an NMR study.

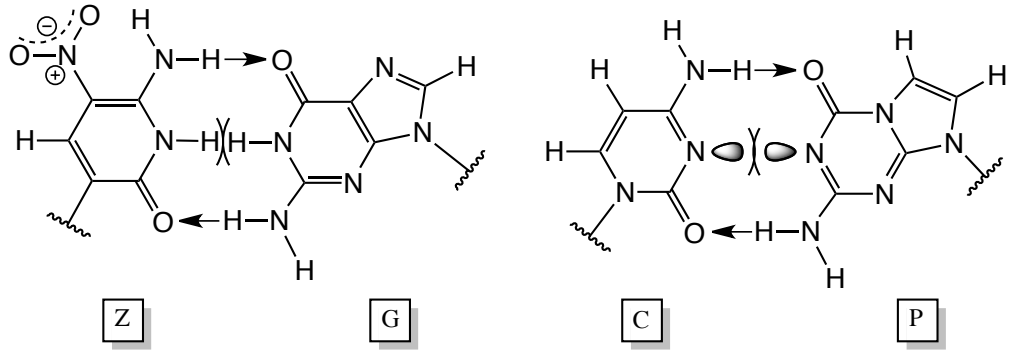


Figure 4-4. Possible structures of G.Z and P.C in basic or acidic environment.

Alternative structures for G·Z and P·C are shown below (Figure 4-4). These two structures were proposed by Yang (2011). The N3-H…N1-H clash in Z·G and N3 lone pair…N1 lone pair clash in C·P should be destabilizing in neutral solution. The left structure might be preferred in basic environment, as one of the two clashing protons could be stripped away; the right structure on the other hand might be preferred in acidic environment as one proton could be dropped in between the two opposing nitrogen atoms and form a stable N3··H··N1 structure. Increasing pH for G·Z containing sequences and decreasing pH for P·C containing sequences in melting experiments will observe increased Tm if these two structures are taken. These two configurations, however, are

hard to believe to be the structure for G·Z or P·C in neutral solution considering their relatively high stability compared to P·T mismatch, which present a carbonyl-carbonyl repulsive component in its configuration with almost no doubt.

P·T mismatch: the P·T mismatch is less stable than the G·C pair by +3.5 kcal/mol. The proposed structure presents a Watson-Crick configuration with thymine C4 carbonyl group and P's C6 carbonyl group facing towards each other. Its relatively low stability indicates that the repulsion between the two oxygen atoms' lone pair electrons may have pushed the two bases away, or even might have forced the base pair plane to propeller twist from the carbonyl-carbonyl side. The enthalpy and entropy change of P·T mismatches exhibit strong context-dependent character. Decreasing the pH for P·T containing sequences in melting experiments should observe increased Tm.

A·Z mismatch: the A·Z mismatch is less stable than the G·C pair by +3.6 kcal/mol, and more stable than the A·C pair by -0.4 kcal/mocl. The structure of A·Z mismatch should resemble the that of A·C mismatch, and requires more rotation of Adenine N9 glycosidic bond toward the major groove. Both form one hydrogen bond and A·C seems more sterically favored. It is not clear that why A·Z is more stable than A·C.

4.3 Reverse Selection of Desired Secondary Structure in Probe Design

A solution to the structure design problem at the end of Chapter 1 is proposed here using P·Z as reverse selection building blocks. For quantitative illustration, using parameters in Table 4.5, changing the 3rd base G into P, and changing the 23rd base C into Z (Figure 4-5).

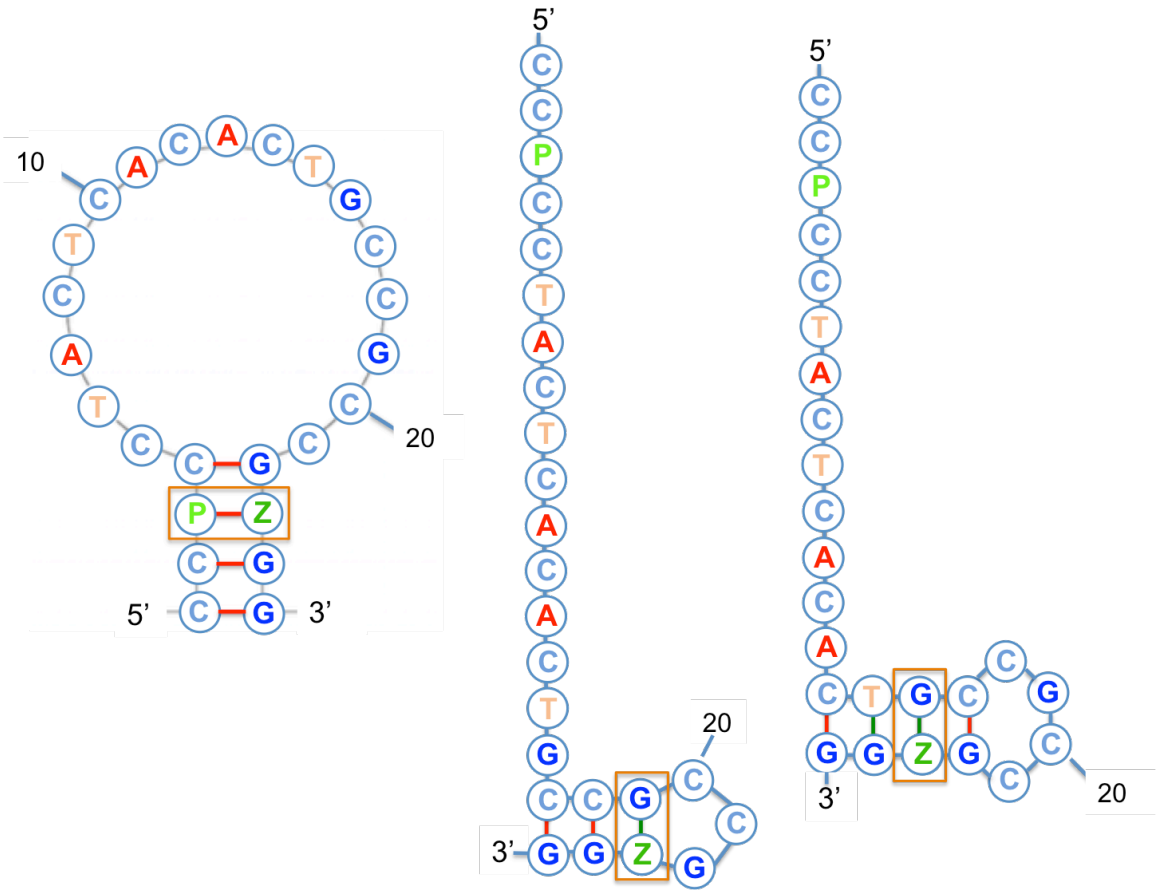


Figure 4-5. Reverse selection using P and Z. By replacing the 3rd base G with P, and replacing the 23rd base C with Z, the stability of the first structure is increased by $\Delta\Delta G°_{37}$ of 0.8 kcal/mole upon formation of a P·Z pair, and the stability of the secondary and the third structure is decrease by $\Delta\Delta G°_{37}$ of 0.9 kcal/mol upon formation of a G·Z pair either between position 19 and position 23 (the middle structure), or between position 16 and position 23 (the left structrue). The total $\Delta G_{37}°$ for the left, the middle, and the right structures become -2.65 kcal/mole, -0.33 kcal/mol, and -0.09 kcal/mol respectively.

In fact, a better solution is to change the 3rd base G into Z, and changing the 23rd base C into P, therefore turns the relatively stable mismatch G·Z into a conceivably much more unstable mismatch G·P. The reverse selectivity of P·Z pair comes from two aspects: similar to backbone and sugar modified nucleic acid analogues, the higher stability of the

pair enhances local structure; more importantly, their inability to paring with A, T, G or C excludes the undesired structures.

4.4 P·Z as the Third Pair of Expanded Genetic System

The trend of stabilizing contribution of P·Z and P, Z containing mismatches to a duplex is P·Z > G·C > G·Z > P·C ≈ G·T > P·T ≈ A·Z > A·C, P and Z involved mismatches are underlined. It is noticeable that the most stable mismatch is G·Z, and the second stable mismatch is P·C. In polymerase-catalyzed reactions, considering reverse mutation (mutation from P·Z to G·C or A·T only), G·Z leads to mutation from Z to C, P·C leads to mutation from P to G, P·T leads to mutation from P to A, and A·Z leads to mutation from Z to T. If only consider the stability trend of mismatches, mutation frequency is Z→C > P→G > P→A > Z→C. In PCR experiments which only provided natural nucleobase triphosphates as substrates for Taq polymerase to amplify P, Z containing templates, the mutation frequency was observed as Z→C >> Z→T, and P→A was slightly more frequently than P→G > (Yang et al., 2011), indicating that the Taq polymerase preferred P·C a lot more than P·T, and preferred Z·A slightly over Z·G. The pH of Taq polymerase reaction buffer is 8.3, which might explain the preference of P·C over P·T. The preference of Z·A over Z·G might have due to the same reason.

4.5 Future Directions

Melting experiments under variant pH conditions should give more information on structures of G·Z, P·C, and P·T as discussed in 4.2. More specifically, G·Z containing sequences are expected to exhibit increased stability under low pH condition, and P·C or P·T containing sequences expect to see increased stability under high pH condition.

DNA and RNA folding programs have been widely used in nucleic acid secondary structure prediction. Mfold uses nearest-neighbor parameters measured by SantaLucia's Lab in DNA secondary prediction (Zuker, 2003). Turner Group compiled thermodynamic experimental results from 1972 (Gralla and Crothers, 1973) to (Schroeder et al., 2003) for RNA secondary prediction (Turner and Mathews, 2009). Obtaining the full set of nearest-neighbor thermodynamic parameters of P·Z pair and integrating them into existing algorithms will enable accurate prediction to secondary structures folded by P·Z containing DNA and RNA sequences.

Appendix A: Improving Ligation

A.1 List of Abbreviations

TBE	Tris/Borate/EDTA
EDTA	Ethylenediaminetetraacetic acid
BSA	Bovine Serum Albumin
TEMED:	tetramethylethylenediamine
Phusion HF	Phusion® High-Fidelity DNA Polymerase
NaAc	sodium acetate

A.2 Discussion on Why Many Previous Cyclization Reactions Have Failed

In cyclization experiments the two restriction ends of DNA are joined together through ligation. In the past thirty years phenol/chloroform extraction followed by ethanol precipitation has been used as the standard purification method to prepare DNA for subsequent cyclization work. During the same period of time, a constant difficulty encountered by people working in this area has been to obtain fully reactive DNA at high purity. Typical consequences are that large fraction of non-ligatable linear DNA is observed, and that bimolecular ligation continues slowly after unimolecular cyclization is completed due to the presence of one-end reactive DNA.

Daniel Gowetski noticed that using QIAquick PCR Purification Kit instead of phenol/chloroform extraction in DNA preparation improved cyclization results significantly. Later an alternative commercial product, the GeneJET PCR Purification Kit® was found to improve ligation yield as well. The purpose of this work is to

investigate how this purification kit helped, and why the previous cyclization reactions failed often.

A.3 Materials and Methods

All enzymes and buffers were purchased from New England BioLabs® if not specified.

A.4 Radioactive Labeling of DNA

9C14(+4) DNA molecules were labeled by adding α-^{32}P labeled dATP in PCR amplification. PCR was performed 65 days after the reference date of α-^{32}P labeled dATP (3000 Ci/mmol, 10 μCi/μL on reference date). On the day PCR was performed, the stock concentration of α-^{32}P labeled dATP was 1.09 μM, and the stock concentration of unlabeled dATP was 2.24 μM.

A single PCR mixture (50 μL) contained 2 ng of pRM9C14 plasmid, 0.4 μM of BsaHI(+4) 2.1 top primer and 0.4 μM M13/pUC bottom primer, 100 μM of four cold dNTPs, 2 unit of Phusion HF; 1X Phusion HF buffer (diluted from 5X Phusion HF buffer); additional 0.04 mM $MgCl_2$, 1 μL of purchased α-^{32}P labeled dATP (1.09 μM of labeled dATP and 2.24 μM of unlabeled dATP), and purified water. Two drops of pure mineral oil were added on top of the 50 μL PCR mixtures by a p-200 pipette to prevent water evaporation. Detailed preparation process using Master Mix method is described in the next paragraph.

A 200 μL of Master Mix was prepared by mixing 4 μL of 2 ng/μL pRM9C14 plasmid (water dilution of mini-prep), 8 μL of 10 μM BsaHI(+4) 2.1 top primer (in 1X TE buffer), 8 μL of 10 μM M13/pUC bottom primer (in 1X TE buffer), 8 μL of 2.5 mM dNTP (in 1X TE buffer), 4 μL of 2 unit/μL of Phusion HF, 40 μL of 5X Phusion HF buffer, 8 μL of 1 mM $MgCl_2$, 4 μL of 3.33 μM α-P-32 labeled dATP (3000 Ci/mmol, 10 μCi/μL on reference date), and 116 μL of Barnstead Nanopure Purification System purified water ("the purified water"). A single PCR mixture was made by adding 50 μL of Master Mix to a PCR tube. Three single PCR mixtures were made.

PCR program PHU-52 was used: one cycle of 95 °C for 5 minute; followed by 30 cycles of (95 °C for 1 minute, 52 °C for 30 seconds, 72 °C for 30 seconds); followed by 72 °C for 5 min; and finally 4 °C forever.

After amplification, 150 μL of raw products were split and treated in different ways to test the pre-restriction influence of ethanol precipitation, column purification, phenol/chloroform extraction and residual Phusion HF polymerase.

A.5 Pre-restriction Treatment

Ethanol Precipitation (E). 90 μL (3 volumes) of 100% ethanol and 3 μL (1/10 volume) of 3M NaAc (pH 5.2) were added to 30 μL of raw PCR product. The mixture was gently mixed, placed in -80 °C freezer for 15 min, followed by 15 min centrifugation at 13,200 rpm under 4 °C. The supernatant was decanted, and the residual liquid was removed by careful pipetting. Additional 1 mL of 70% ethanol was added to rinse DNA pellet. The 70 % ethanol and DNA pellet mixture was place on a vortex mixer briefly, followed by 15 min centrifugation at 13,200 rpm at 4 °C. Again the supernatant was

decanted, and the residual liquid was removed by careful pipetting. DNA pellet was allowed for 15 min air dry, and finally dissolved in 90 μL of the purified water.

Phenol-chloroform Extraction Followed by Ethanol Extraction (ΦE). 15 μL (1 volume) of phenol/chloroform/isoamyl alcohol (25:24:1, purchased from Sigma-Aldrich) was added to 15 μL of raw PCR product. The mixture was vigorously mixed, followed by 3 min centrifugation at 13,200 rpm at room temperature. The mixture separated into two layers: a top aqueous layer containing DNA and salts, and a bottom organic layer containing proteins. The bottom organic layer was removed by pipetting. 15 μL (1 volume) of chloroform was added to the remaining top aqueous layer for reverse extraction of the residual organic component in aqueous layer. This reverse extraction mixture was vigorously mixed, followed by 3 min centrifugation at 13,200 rpm at room temperature. Again the bottom organic layer was removed by careful pipetting. The remaining 15 μL aqueous layer underwent an ethanol precipitation.

Proteinase K Treatment Followed by Phenol/Chloroform Extraction and Ethanol Precipitation (KΦE). 2 μL of 0.8 unit/μL Proteinase K, 2 μL of 10 X NEBuffer 4, and 1 μL of the purified water were added to 15 μL of raw PCR product. This 20 μL Proteinase K digestion mixture was incubated at 37 °C for 60 min, followed by phenol-chloroform extraction and ethanol precipitation.

GeneJET PCR Purification Kit Method (GeneJET Kit, G). 45 μL of raw PCR product was purified using the purification kit following the product instruction, and was eluted in 45 EB provided by the kit. The purified PCR product was split into a 15 μL part (for addition of Phusion HF polymerase) and a 30 μL part.

GeneJET Kit Method Followed by Addition of Phusion HF (GP). 1 µL of 2 unit/µL Phusion HF was added to the 15 µL part of the purification kit purified PCR product.

A.6 BsaHI Restriction and Gel Purification

E, ΦE, KΦE, G and GP treated PCR products were subjected to an overnight BsaHI restriction (0.2 unit/µL BsaHI, 1X NEBuffer 4 and 0.5 µg/µL BSA at 37 °C); 2 µL of sample G dissolved in 10 µL of 1X NEBuffer 4 with 1 µg/µL BSA overnight at 37 °C was used as restriction control (c). Restriction mixtures and control were loaded to a 0.8-mm 6% 75:1 acrylamide:bis-acrylamide gel (50 mL gel with 1XTBE buffer, 1% ammonium persulfate and 0.04% TEMED), electrophoresed for 1.5 hour at 400 volts. After electrophoresis the gel was laid on the glass plate, wrapped by a plastic film, put under a phosphor storage plate for several hours, and the storage plate was imaged by a Storm PhorphorImager® (Figure A-1). GP sample was degraded and was not recycled.

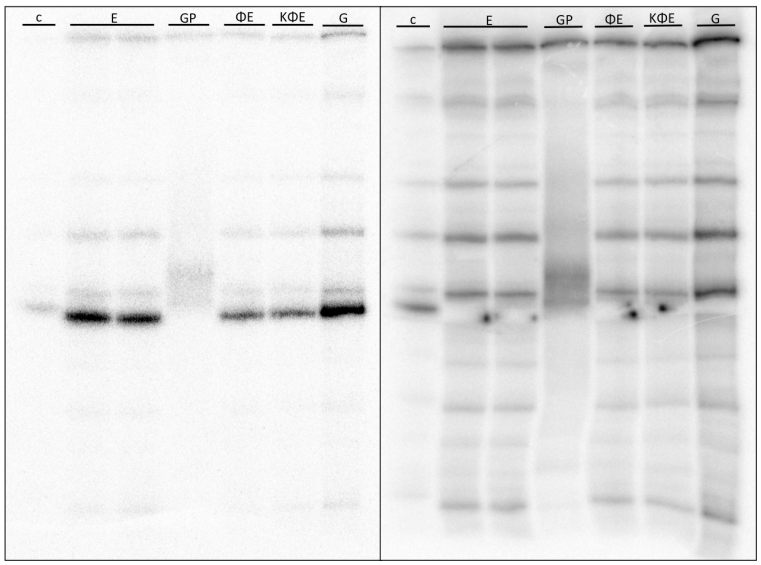


Figure A-1: BsaHI restriction results of ^{32}P labeled 9C14(+4) processed by variant pre-restriction treatments: before (left) and after (right) cutting out gel slice.

The full-sized picture of the radioactive gel was print out and laid behind the glass on which the gel was placed, as guidance for cutting out gel slices. Gel slices were cut out at corresponding positions where restriction product bands were imaged with a razor rinsed by pure ethanol, and diced into small cubes with side length about 0.5 cm. The gel was re-imaged after the desired slices were cut out to make sure that cutting was performed at the correct position (Figure A-1, right). The gel cubes from each gel slices were soaked in 200 μL gel elution buffer (0.5 M ammonium acetate, 1 mM EDTA, pH 8.0), frozen at -80 °C for 10 min, and incubated in 37 °C over night.

This overnight gel slice elution (1^{st} elution) was separated from the gel cubes by pipetting, and then placed into speed vacuum in order to reduce the solution volume. At the same time an additional 200 μL gel elution buffer (2^{nd} elution) was added to the gel cubes, and incubated at 37 °C for on hour. The 1^{st} and 2^{nd} elution were combined and mixed. This combined gel elution was subject to ethanol precipitation or GeneJET Kit method as the post-restriction treatment.

A.7 Post-restriction Treatment

The combined gel elution of samples treated by ethanol precipitation as the re-restriction treatment was split evenly into two parts. One part was ethanol precipitated and dissolved into 30 μL of purified water, the other part was treated by GeneJET Kit and eluted in 30 μL of GeneJET EB buffer. The resultant two samples were E-E and E-G. The combined gel elution of samples treated by GeneJET Kit as the pre-restriction treatment was subjected to the same post-restriction treatment, and the two resultant samples were G-E and G-G. The combined gel elution of samples treated by the other

two pre-restriction treatments, i.e. ΦE and KΦE, were ethanol precipitated as the pre-restriction treatment, and the two resultant samples were ΦE-E and KΦE-E.

A.8 Ligation Experiments, BsaHI Re-restriction and BAL-31 Digestion of Ligation Products

Ligation was performed 4 days after PCR labeling due to an overnight restriction, an overnight elution, and some preparation time as well. The stock T4 DNA ligase concentration was 400 unit/μL, and was diluted to 40 unit/μL in 1X T4 ligase storage buffer (10 mM Tris-HCl, 50 mM KCl, 1 mM DTT, 0.1 mM EDTA, 50 % glycerol, pH 7.4). T4 DNA ligase was kept on ice during ligation sample preparation. A 20 μL ligation mixture contained 2 nM DNA, supplementary 1 μL of 2mM ATP, 2 μL of 40 unit/μL T4 DNA ligase, 2μL of 10X T4 DNA ligase buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM ATP, 10 mM DTT, pH 7.5), and the purified water. The negative control contained G-G sample as the DNA and all the other reaction components except for T4 ligase. For topoisomer identification purpose, G-G sample was ligated in the presence of 0.6 ng/μL ethidium bromide. Ligation was performed at room temperature for 30 min, followed by 30 min of quenching, or re-restriction, or BAL-31 digestion. A 40 μL quenching reaction contained 20 μL of ligation mixture and 20 μL of 2X quenching mixture (4 μg/μL Proteinase K and 75 mM EDTA), and was incubated at room temperature. A 25 μL of BsaHI re-restriction contained 20 μL of ligation mixture, 1 μL of 10 unit/μL BsaHI, 1.5 μL of 10 μg/μL BSA and 2.5 μL of 10X NEBuffer 4, and was incubated at 37 °C. A 42 μL BAL-31 digestion reaction contained 20 μL ligation mixture, 1 μL of 1 unitl/μL BAL-31 and 21 μL of 2X BAL-31 reaction buffer (20 mM Tris-HCl, 600 mM NaCl, 12

mM $MgCl_2$, 12 mM $CaCl_2$, 1mM EDTA, pH 8), and was incubated at room temperature. The quenched, re-restricted, and BAL-31 treated samples were ethanol-precipitated and dissolved in 16 μL of the purified water and 4 μL of 6X loading buffer (30% glycerol, 0.25% bromophenol blue, and 0.25% xylene cyanol), and were analyzed on a 6 % acrylamide gel (75:1) containing 7.5 μg/mL chloroquine in 1X TBE buffer (50 mM Tris base, 50 mM boric acid, and 1 mM EDTA, filtered by 0.2 micron Whatman® filter paper) at 120 volts for 12 hours. The regular addition of 1/10 volume of in ethanol precipitation was skipped to 3 M NaAc to BAL-31 treated samples, which had a high salt concentration (600 mM NaCl).

A.9 Results

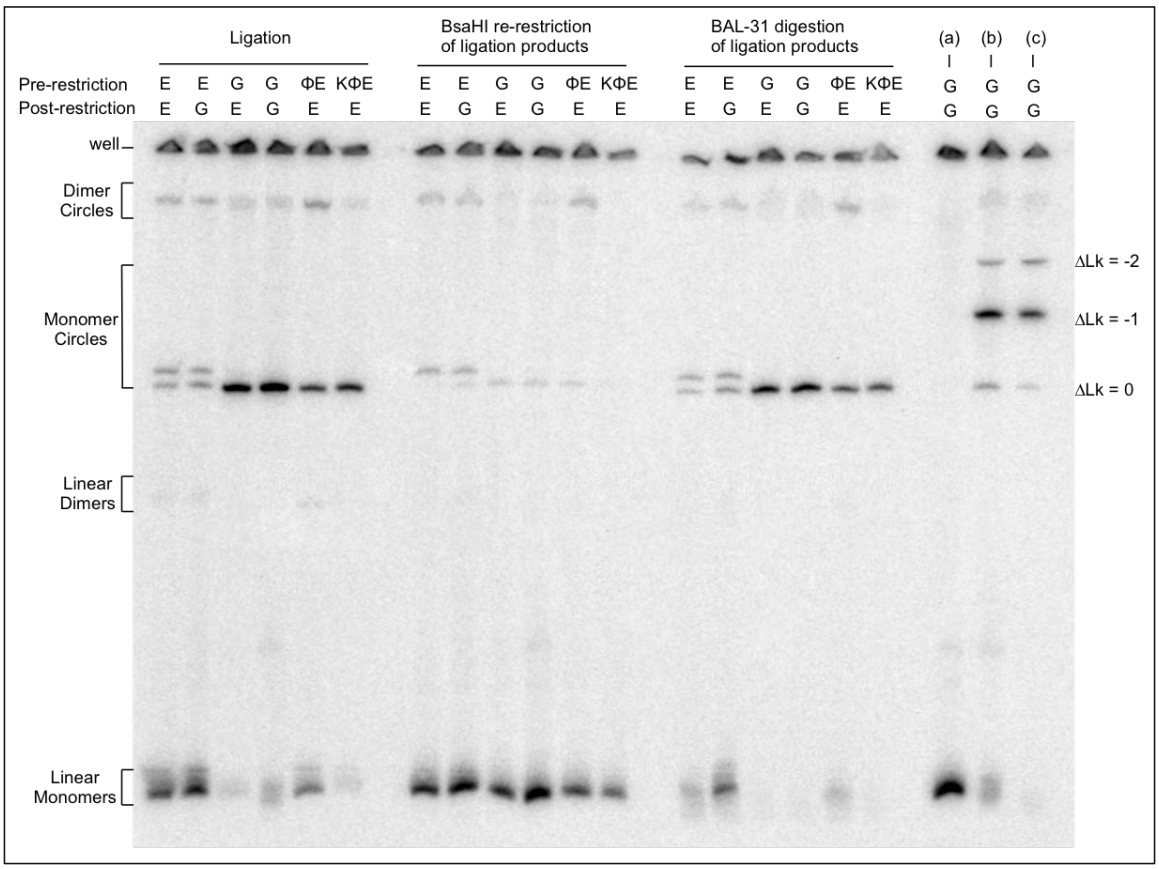


Figure A-2: Chloroquine gel electrophoresis of ligation products, BsaHI re-restricted ligation products, and BAL-31 digested ligation products, 9C14(+4). (a) no T4 ligase control; (b) ligation performed in the presence of EB; (c) BAL-31 digestion of ligation products from (b).

A.10 Discussion

A.10.1 Design of 9C14 molecule, and structure of 9C14(+4) molecule

9C14 molecule was originally designed for study of LacI mediated DNA looping. The molecule was characterized by a phased A-tract bend bracketed by two O_{sym} Lac

operators (Mehta and Kahn, 1999). Runs of 4-6 adenine base pairs (A-tract) repeated with the helical periodicity gives global curvature of the DNA double helix (Haran and Mohanty, 2009). A-tracts are bend towards the minor groove at the center. 9C14 molecule contains 8 A_6 tracts separated by CGGGC or CGGC sequences. It was determined by Koo et al (1990) that A_6 tracts separated by CGGGC or CGGC sequences bends the DNA helix by 17-21° by a single A_6 tract. The estimated total bending angle 9C14 is estimated to be 136-168°. The two ends of this highly bent molecule are therefore brought close to each other compared to the regular linear B-DNA.

9C14(+4) PCR product is a 427 bp long DNA molecule characterized by the same A_6 tracts, which locate from the 219th position to the 298th position on the strand containing polyA. BsaHI restricts double stranded DNA at site 5'…G R▼C G Y C…3' / 3'…C Y G C▲R G…5', generating 5' CG overhangs on both ends. 9C14(+4) PCR product contains two BsaHI sites: one locates from the 9th to the 14th position (5'-GG|CGCC-3'), and the other locates from the 415th to the 420th position (5'-GG|CGTC-3') (Figure A-3).

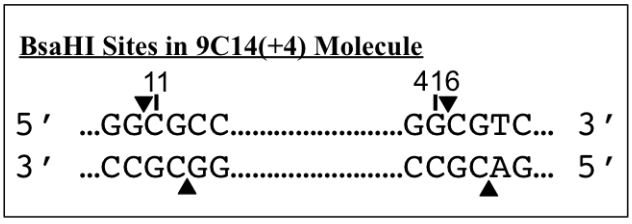


Figure A-3: Sketch of two BsaHI sites in 9C14(+4) PCR products.

BsaHI cleaves before the 11th position and after the 416th position at the A_6 tract strand, and before the 13th position and after the 418th position at the complementary strand, resulting a duplex that is 406 bp long in each single strand, with 404 base pairs and the 5'

CG overhangs on both ends (Figure A-4). The perfectly restricted 9C14(+4) would have the two 5' CG overhangs well aligned to each other for cyclization: the two arms bracketing the 80 bp A-tracts (helical repeat 10.33 bp/turn) are 326 bp long regular B-DNA (helical repeat 10.45/turn), therefore the most relaxed linking number (Lk) of the monomer circle formed by joining the two 5' CG overhangs is (80/10.33) + (326/10.45) = 38.94, which requires only slightly over twist of the two ends to get ligated. (Alternatively, according to Mehta and Kahn, the length of A_6 tracts is 84 bp, then the rest B-DNA length is 322 bp, and the most relaxed Lk is 38.95.) The decreased deformation energy required for bringing the two DNA ends together because of the intrinsic bending, as well as the well-aligned two restriction ends facilitates the monomer cyclization by perfectly restricted 9C14(+4) molecule, i.e. the Double 5' CG.

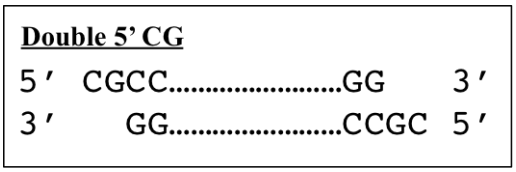


Figure A-4: Sketch of the PCR product with perfectly restricted ends, the Double 5' CG.

A.10.2 Assays applied for identification of restriction products

BsaHI re-restriction. T4 ligase catalyzes the formation of a phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini in duplex DNA or RNA. Ligation of two perfectly restricted BsaHI sites always produces a new BsaHI site. A second BsaHI restriction after ligation should restore any circular product to the linear

form with such new site. BsaHI re-restriction assay therefore can be used to test the existence of perfectly restricted sites.

BAL-31 digestion. BAL-31 is an exonuclease that degrades both 3' and 5' termini of duplex DNA, and is a highly specific single-stranded endonuclease that cleaves at nicks, gaps and single-stranded regions of DNA and RNA. BAL-31 digestion assay therefore can be used to test whether or not a species is in the circular form.

Topoisomer differentiation by chloroquine gel eletrophoresis. In a cyclization reaction, the 5' phosphate and the 3' hydroxyl termini of a linear DNA were covalently joined by T4 ligase. Ethidium bormide (EB) is a DNA intercalator that binds DNA by inserting in between stacked base pairs. When a linear DNA is intercalated by EB molecules, the double helix is unwound, resulting in decreased number of twist. Cyclization of unwound DNA leads to circular products with smaller Lk. The number of ΔLk depends on the extent of EB resulted unwinding. During electrophoresis, the intercalated EB molecules dissociates from the double helix, the circular product regains some twists because of the torsional force possessed by the double helix, and leads to a negative writhe change (ΔWr) that compensates to the positive twist change (ΔTw). Similar to EB, chloroquine is another DNA intercalator. In the chloroquine gel electrophoresis, chloroquine intercalating into circular products, introducing positive writhe to the circles, and differentiating circles with variant ΔLks. With high enough chloroquine concentration, the more negatively writhed species exhibits slower gel mobility.

A.10.3 Pre-restriction treatment versus Post-restriction treatment

Applying ethanol precipitation or GeneJET PCR purification kit (the "GeneJET kit") method as the post-restriction purification method did not make difference to the ligation results, as evidenced by E-E, E-G samples and G-E, G-G samples. The smearing of linear monomer product that was observed in the G-G ligation result but not in the G-E ligation result should be due to salt effect, as G-E sample was dissolved in the purified water while G-G sample was eluted in the GeneJET elution buffer. It was the pre-restriction treatment that made a difference to restriction products, which further generated different ligation products and are discussed below.

A.10.4 Pre-restriction Treatments and Corresponding Restriction Products

A.10.4.1 The GeneJET kit:

The GeneJET kit removes primers, triphosphates, enzymes and salts from PCR reaction. The cleanly purified G-G sample was then subjected to BsaHI restriction, and produced perfectly restricted ends, which is the Double 5' CG. The Double 5' CG linear monomer should appear in the no T4 ligase control, and should form monomer and dimer circles in the ligation.

The Double 5' CG was observed in the G-G sample no T4 ligase control. As expected it is susceptible to BAL-31 digestion.

Monomer circle formed by Double 5' CG was observed in G-G and G-E ligation. The identity of this monomer circle was confirmed by its topoisomer distribution in cyclization reaction with and without EB, by its susceptibility to BsaHI re-restriction, and by its resistance to BAL-31 digestion. The trace amount of remaining Double 5' CG

monomer circle after re-restriction should be due to an insufficient restriction considering the relatively short reaction time (30 min). The majority of the cyclized Double 5' CG monomer circle should be Lk^0 = 39 (ΔLk = 0) species, and indeed a single band was observed in cyclization in the absence of EB. In the presence of EB, the other two circular species with slower gel mobility appeared, and these were circular products with negative writhes, presumably Lk = 38 (ΔLk = -1) and Lk = 37 (ΔLk = -2) topoisomers respectively. The Double 5' CG monomer circle was observed in E-E, E-G, ΦE-E, and KΦE-E ligation as well.

Dimer circle formed by the Double 5' CG species was observed in G-G and G-E ligation. The length of the Double 5' CG dimer is 812 bp thus moved much slower in the chloroquine gel. The fact that this species was susceptible to BsaHI re-restriction and was resistant to BAL-31 digestion further confirmed its identity. Again the trace amount of remaining Double 5' CG monomer circle after re-restriction should be due to the incomplete restriction. This dimer circle was observed in G-E and KΦE-E ligation results as well. The somewhat blurry bands observed in BAL-31 digestion may be because of the low radiation signal detected as a result of sample loss during ethanol precipitation, which was evidenced by the decreased band darkness of all ΔLk = 0 species in BAL-31 digestion compared to those in ligation. The Double 5' CG dimer circle was observed in KΦE-E ligation as well. The species appearing in E-E, E-G and ΦE-E ligation and having very similar gel mobility as Double 5' CG dimer circle may however not be the exactly same species, and is discussed below.

A.10.4.2 Ethanol Precipitation:

Ethanol precipitation is a common method used for concentrating and de-salting DNA. The process however is not designed for removing enzymes. Phusion HF as well as residual triphosphates entered into the following BsaHI restriction reaction; presumably resulting in various imperfectly restricted 9C14(+4) products with G or C filled in the 5' CG overhangs (Figure A-5).

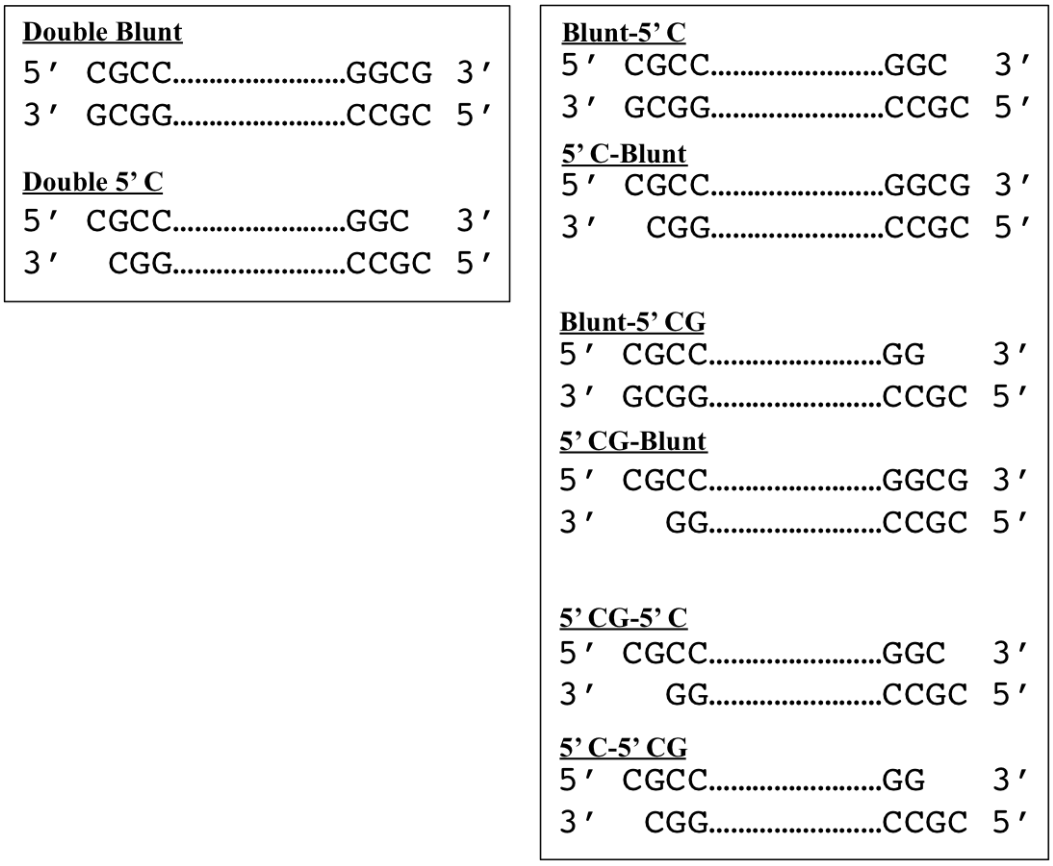


Figure A-5: Sketch of possible species with fill-in ends resulted from Phusion HF and triphosphate presence in BsaHI restriction: products with symmetric fill-in ends (left panel), and with asymmetric ends (right panel).

The two species with symmetric fill-in ends, the Double Blunt and Double 5' C can be ligated to form monomer circle. The circular products formed by these two species are

resistant to the re-restriction because of the loss of BsaHI site: for the Double Blunt, the ligated site is 5'-GGCGCGCC-3' (the underlined are the two fill-in bases, C and G), and for the Double 5' C, the ligated site is 5'-GGCCGCC-3' (the underlined is the fill-in base, C). All of the fill-in species may form dimer circles. The Double Blunt and Double 5' C may form head-to-head or head-to-tail self-dimer circles, and the species with asymmetric fill-in ends may form head-to-head self-dimer circles. Three heterogeneous dimer circles may form by head-to-head ligation of the blunt-5 C' and the 5' C-blunt, the blunt-5' CG and the 5' CG-blunt, and the 5' CG-5' C and 5' C-5' CG.

At least three unligated linear monomers were observed in E-G and E-E ligation. It is hard to give the definite identity to each one because of the variety of fill-in ends. It can be concluded however that that these three monomers have very close yet different lengths, and they are unlikely to form ligation products. The three species, Double 5' CG, Double 5' C and Double Blunt can form monomer circles and therefore not likely to remain unligated. The six species with asymmetric ends cannot form monomer circle in ligation and therefore are more likely to be in linear form. The lengths of the double-stranded parts of them are 405 bp (5' CG-5' C and 5' C-5' CG), 406 bp (Blunt-5' CG and 5' CG-Blunt), and 407 bp (Blunt-5' C and 5' C-Blunt), which could explain the three species observed. The same unligated linear monomers were observed in ΦE-E ligation results with smaller amount, and in KΦE-E ligation results with extremely small amount.

A very faint band or maybe several very faint bands appeared between linear monomers and monomer circles were identified as the linear dimer(s). Mehta and Kahn studied cyclization kinetics of 9C14 DNA (1999), and observed the linear dimer with

similar gel mobility. The susceptibility to BAL-31 digestion confirmed the linear shape. Moreover, the fact that the linear dimer(s) did not appear in G-G or G-E ligation sample suggested that it (they) should be dimer(s) of fill-in species.

Two monomer circles were observed in ligation results of E-E sample and E-G sample. One was the monomer circle produced by the Double 5' CG, and the other with slightly slower mobility should be monomer circle of either the Double Blunt or the Double 5' C. The Lk^0 of these two circles are 38.98 (80/10.33+328/10.5) for the Double Blunt, and 38.89 (80/10.33+327/10.5) for the Double 5' C. It is more likely that Double 5' C would form monomer circle because ligation of blunt ends is less efficient than ligation of sticky ends.

At least one dimer circle species was observed in ligation results of E-E sample and E-G sample, which appeared at almost the same position as the Double 5' CG dimer circle. The fact that this species was somewhat if not totally resistant to BsaHI re-restriction showed that it must had contained dimer circle(s) generated from the fill-in species without any 5' CG overhang. It was likely that multiple dimer circles formed from various combinations of fill-in species may present, and the gel was not able to differentiate them. This dimer circle species was observed in ΦE-E ligation as well.

A.10.4.3 Phenol-Chloroform Extraction

Phenol-chloroform extraction followed by ethanol precipitation is widely used for extracting nucleic acid from mixture containing proteins such as cell lysates and PCR mixtures. Phenol is a protein denaturant. Chloroform facilitates the separation between aqueous phase and organic phase, and eliminates the slightly water dissolving effect of

phenol. It was however unknown that to what extent the denaturation of protein was achieved by this method.

The following species were observed in ΦE-E ligation: linear monomers of fill-in species, linear dimer(s) of fill-in species, monomer circle of Double 5' CG, and dimer circle formed by fill-in species (and maybe dimer circles formed by the Double 5' CG as well, see discussion on dimer circles in ethanol precipitation section).

A.10.4.4 Proteinase K Digestion Followed by Phenol-chloroform Extraction

Proteinase K is a serine protease that hydrolyzes a variety of peptide bonds. Adding Proteinase K to the raw PCR mixture should inactivate Phusion HF. The following phenol-chloroform extraction was meant to inactivate Proteinase K to prevent hydrolysis of T4 ligase.

The following species were observed in KΦE-E ligation: extremely small amount of linear monomers of fill-in species, monomer circle of the Double 5' CG, and dimer circle of the Double 5' CG.

A.11 Conclusion

As a standard ligation/cyclization procedure, DNA was extracted from PCR mixtures by phenol-chloroform extraction, restricted, and then ligated. The unligatable species were noticed and was attributed to incorrect restriction or phosphatase activity (Kahn and Crothers, 1992). We provide evidence for polymerase modified DNA restriction ends, and the fill-in species characterized by those ends either changed the ligation product distribution, or were not able to be ligated. Cyclization kinetic assay measures the

monomer cyclization rate constant and bimolecular association rate constant, and derives *J* factor, the effective concentration of one properly aligned DNA end about the other, as the ratio of the two (Kahn and Crothers, 1992). *J* factor has been used to calculate parameters modeling DNA flexibility. A parameter *f*u was introduced to account for unligatable species, yet the bimolecular products were considered homogeneous and being ligated at a unique rate. Our results show that the amount of bimolecular products formed in ligation was changed by fill-in species, suggesting that the ligation mechanism of fill-in ends may be different to the perfectly restricted ends. The fact that the gel mobility of dimer circles formed by fill-in ends were almost the same as those formed by perfectly restricted ends made it very hard to differentiate these two dimer circles. Cyclization kinetic experiments performed on DNA with fill-in ends versus perfectly restricted ends would give more information on how the mechanisms differ. Nevertheless, now that the QIAquick PCR Purification Kit®/ GeneJET PCR Purification Kit® are available for complete removal of polymerase and producing perfectly restricted DNA, more reliable experimental data can be obtained to calculate *J* factor, which should improve the accuracy of DNA modeling parameters. From method perspective, the GeneJET Kit and Proteinase K followed by phenol-chloroform extraction were able to eliminate the polymerase activity, and ethanol precipitation failed doing so. While it was generally believed that phenol-chloroform extraction denatures proteins, we showed that a carefully performed phenol-chloroform extraction was not able to inactivate Phusion HF completely.

Appendix B: Cyclization Analysis of Lac Repressor-mediated DNA Loops

B.1 Definition and Significance of Research Problem

DNA looping is an important mechanism widely involved in transcription, replication and recombination. In the classic gene regulation system of Lac operon, looping formed by Lac repressor (LacI) binding at the primary operator O_1 and neighboring secondary operator O_2/O_3 enhances repression via increasing local concentration of LacI to nearby operators. An understanding of DNA loop geometry and stability is essential to quantitatively understanding DNA-looping involved biological processes. The Kahn lab has been focused on designing protein-mediated DNA loops to test whether existing theories and models accurately describe DNA looping geometry and stability. The FRET studies (Haeusler et al., 2012) proposed six loop topologies and mapped out globally the loop geometry distribution for a systematically constructed DNA sequence landscape. To further understand looping topology and stability, in particular what are the twist and writhe of a looped LacI-DNA complex depending on the inner-loop sequence, this work focuses on the comprehensive cyclization analysis of the designed DNA sequence landscape. The results will add knowledge to quantitative understanding of biological process involving DNA looping.

B.2 Research Plans

The cyclization DNA family consists of 75 members. 25 basic DNA constructs are built by varying the adaptor sequences between intrinsically bent A-tracts and operators by 5-13 bp (left adaptor), and 10-18 bp (right adaptor), such that the total DNA length ranges from 396 bp to 412 bp. Two length variants of each basic DNA are built by changing the tail length by -3 bp or +4 bp (Figure A-6).

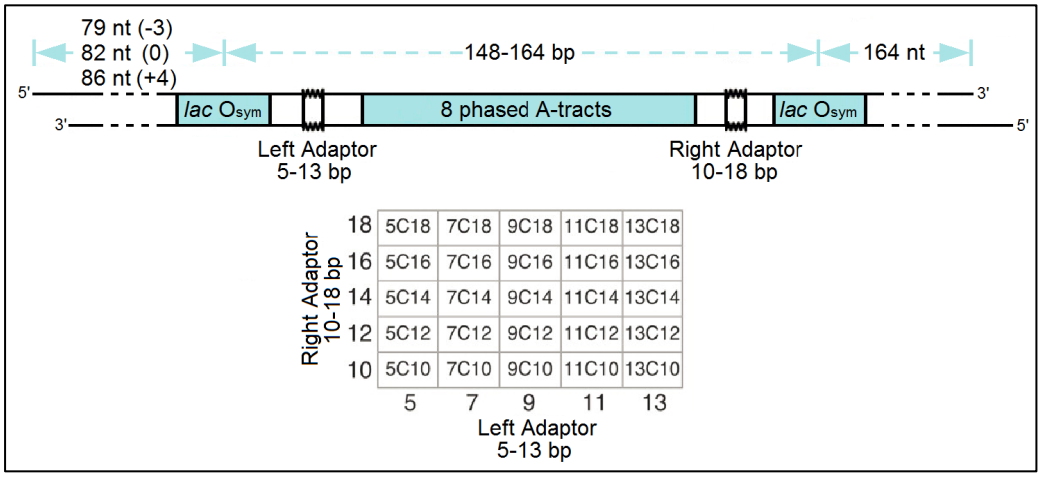


Figure A-6. Structures and nomenclatures of 25 basic DNA constructs.

Cyclization of DNA alone, and DNA with LacI is performed to all 75 DNAs. Data of each basic DNA and the two variants is analyzed as a set by three parameters: torsional modulus C, helical repeat hr, and ΔLkloop. Figure A-7 is a postulated cyclization result of one set DNAs of n bp, (n-3) bp, and (n+4) bp, showing relative population of -1 and 0 topoisomers. Data in Figure A-7 (A) is fitted by Gaussian distribution N(μ, δ2) in Figure A-8. The standard deviation is related to C and hr, and the mean is the value of Lk_0-Lk_m. Gaussian curve shifts by ΔLk_{loop} when LacI binds free

DNA and forms LacI-DNA looping complex (In this case $\Delta Lk_{loop}<0$ as the distribution curve shifts to the left). It can be speculated that Gaussian curves of -3 and +4 length variants shift by -3/hr and +4/hr relative to corresponding curves of basic DNA, with the same standard deviation and ΔLk_{loop}.

As described above, ΔLk_{loop} can be calculated for all 75 DNA constructs. To examine the influence of sequence between two operators on ΔLk_{loop}, ΔLk_{loop} is plotted versus inter-operator sequence length. $\Delta Lk_{loop} = \Delta Tw_{loop} + \Delta Wr_{loop}$, where ΔWr_{loop} depends on the distance between the two operator binding sites of LacI, and ΔTw_{loop} reflects twist strain of inter-operator sequences.

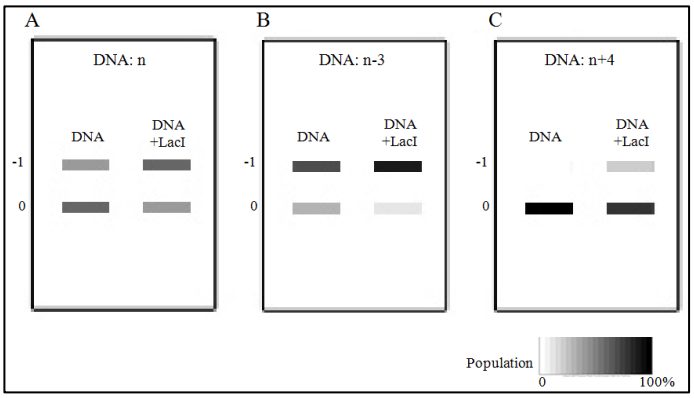


Figure A-7. Postulated cyclization products analyzed by polyacrylamide gel. DNA of n, n-3 and n+4 bp is shown in A, B and C. Relative population of -1 and 0 topoisomers is shown in degree of blackness.

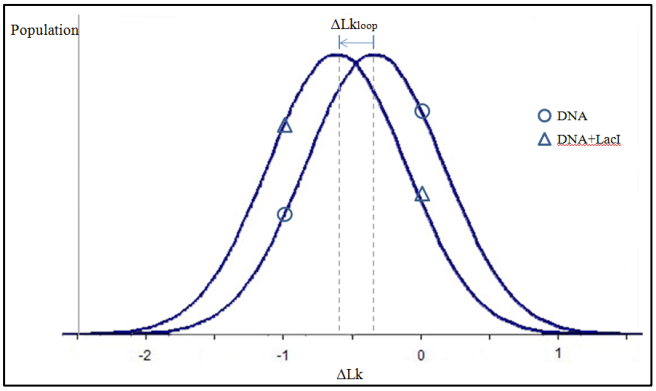


Figure A-8. Data in Figure A-7(A) fitted by Gaussian distribution.

B.3 Results

B.3.1 Cyclization of 5C(+4) DNA with and without LacI (Figure A-9).

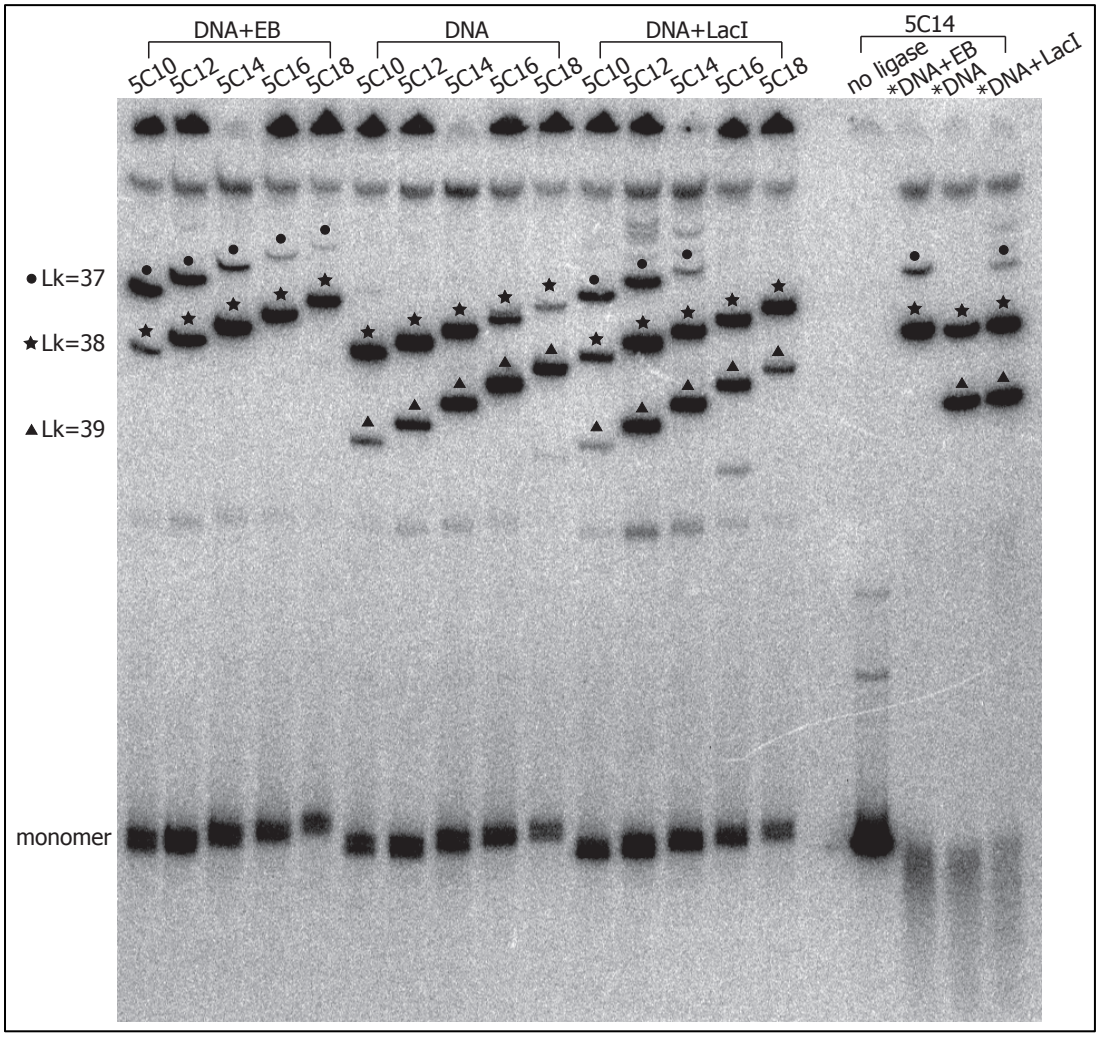


Figure A-9. Cyclization product of 5C (+4) DNA on a native 6% polyacrylamide gel (75:1) containing 7.5ug/ml choloroquine. 0.5 nM DNA is cyclized alone (the second group), with 1.5 nM Lac repressor (the third group), or with 0.15ug/mL ethidium bromide (the first group). * BAL-31 digested.

Negatively supercoiled topoisomers appear in cyclization product when Lac repressor binds 5C DNA. 5C molecules form negatively writhed loop upon Lac repressor binding. This result is in agreement with previous FRET study.

B.3.2 Titration of LacI to DNA (Figure A-10).

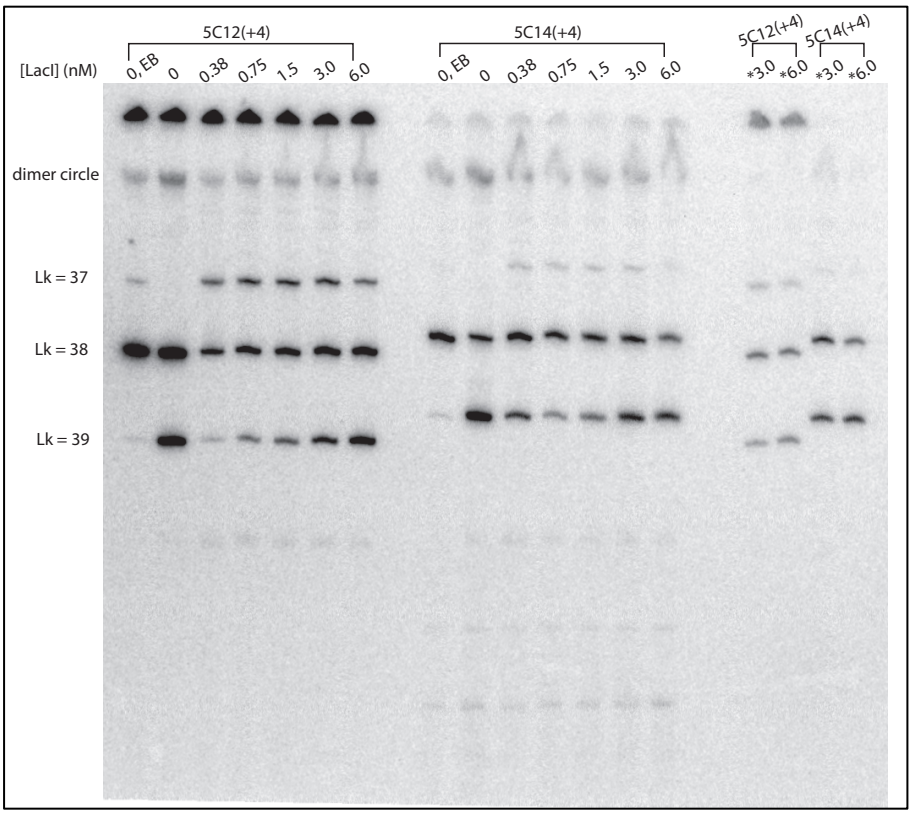


Figure A-10. Cyclization product of 5C12(+4) and 5C14(+4) with gradient concentration of Lac repressor on a native 6% polyacrylamide gel (75:1) containing 7.5ug/ml choloroquine. DNA concentration is 0.5 nM. * BAL-31 digested.

When LacI:DNA molar ratio approaches to 1:1, topoisomer distribution shifts to the maximum extent.

B.3.3 Cyclization of +4 DNA without LacI (Figure A-11).

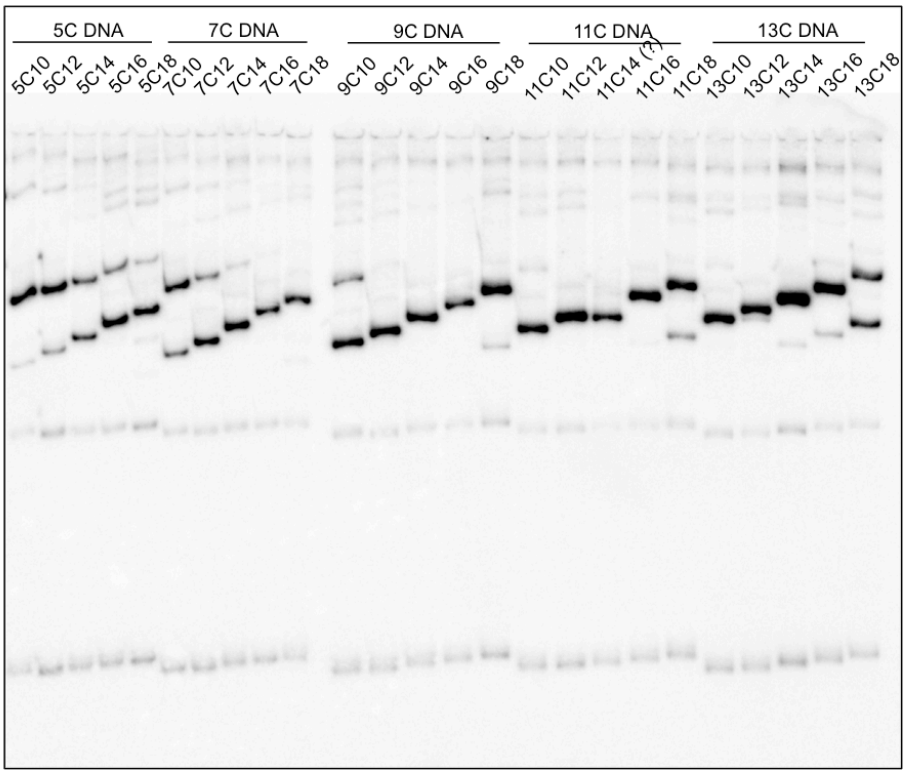


Figure A-11. Cyclization product of all the 25 DNA with +4 tail on a native 6% polyacrylamide gel (75:1) containing 7.5ug/ml choloroquine.

The 11C14 (+4) DNA showed the same cyclization result as that of 11C12 (+4). The identity of 11C14 plasmid needs to be verified.

Supplementary Materials

Sequences

9C14 (+4) PCR product amplified by BsaHI (+4) 2.1 primer and M13 pUC NcoI primer, A_6 tract strand.

GCTGCCATGGCGCCAGGGTTTTCCCAGTCACGACgttgtaaaacgacggccagtgagcg
cgcgtaatacgactcactatagggcgaattggagctccaccgcggtggcggccccccct
cgaggtcgacggtatcgataagcttgatatcaaagctttaccacaacgAATTGTGAGCG
CTCACAATTAGCTAGCTTCGATTCTAGGACGCGTTCAGCGCAAAAAACGGGCAAAAAAC
GGCAAAAAACGGGCAAAAAACGGCAAAAAACGGGCAAAAAACGGCAAAAAACGGGCAAA
AAACCGGATCCGTACGAATTGAGATCTAATTGTGAGCGCTCACAATTacctagactgac
ctgcagatatcggtacccagcttttgttccctttagtgagggttaattggtaccgcgct
tggcgtcgtaatca

BsaHI (+4) 2.1 Primer

5’ tgattacgacgccaagcgcggtaccaattaaccctcac 3’

M13 pUC NcoI Primer

5’ gctgccatggcgccagggttttcccagtcacgac 3’

Bibliography

Allawi, H. and SantaLucia, J. (1997). Thermodynamics and NMR of internal G·T mismatches in DNA. *Biochemistry*, 36(34), pp.10581-10594.

Allawi, H. and SantaLucia, J. (1998a). NMR solution structure of a DNA dodecamer containing single G·T mismatches. *Nucleic Acids Research*, 26(21), pp.4925-4934.

Allawi, H. and SantaLucia, J. (1998b). Nearest-neighbor thermodynamics of internal A·C mismatches in DNA: Sequence dependence and pH effects. *Biochemistry*, 37(26), pp.9435-9444.

Bain, J., Switzer, C., Chamberlin, R. and Benner, S. (1992). Ribosome-mediated incorporation of a non-standard amino acid into a peptide through expansion of the genetic code. *Nature*, 356(6369), pp.537-539.

Eritja, R., Kaplan, B., Mhaskar, D., Sowers, L., Petruska, J. and Goodman, M. (1986). Synthesis and properties of defined DNA oligomers containing base mispairs involving 2-aminopurine. *Nucleic Acids Research*, 14(14), pp.5869-5884.

Gao, F., Lei, J. and Ju, H. (2013). Label-Free surface-enhanced Raman spectroscopy for sensitive DNA detection by DNA-mediated silver nanoparticle growth. *Analytical Chemistry*, 85(24), pp.11788-11793.

Goel, G., Kumar, A., Puniya, A., Chen, W. and Singh, K. (2005). Molecular beacon: a multitask probe. *Journal of Applied Microbiology*, 99(3), pp.435-442.

Gralla, J. and Crothers, D. (1973). Free energy of imperfect nucleic acid helices. *Journal of Molecular Biology*, 73(4), pp.497-511.

Guo, X. and Patel, D. (1987). NMR studies of A·C mismatches in DNA dodecanucleotides at acidic pH. Wobble A(anti)·C(anti) pair formation. *The Journal of Biological Chemistry*, 262(35), pp.16973-16984.

Haeusler, A., Goodson, K., Lillian, T., Wang, X., Goyal, S., Perkins, N. and Kahn, J. (2012). FRET studies of a landscape of Lac repressor-mediated DNA loops. *Nucleic Acids Research*, 40(10), pp.4432-4445.

Haran, T. and Mohanty, U. (2009). The unique structure of A-tracts and intrinsic DNA bending. *Quarterly Reviews of Biophysics*, 42(01), p.41.

Herdewijn, P. (1999). Conformationally restricted carbohydrate-modified nucleic acids and antisense technology. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1489(1), pp.167-179.

Horlacher, J., Hottiger, M., Podust, V., Hubscher, U. and Benner, S. (1995). Recognition by viral and cellular DNA polymerases of nucleosides bearing bases with nonstandard hydrogen bonding patterns. *Proceedings of the National Academy of Sciences*, 92(14), pp.6329-6333.

Hu, Q., Hu, W., Kong, J. and Zhang, X. (2015). Ultrasensitive electrochemical DNA biosensor by exploiting hematin as efficient biomimetic catalyst toward in situ metallization. *Biosensors and Bioelectronics*, 63, pp.269-275.

Hunter, W., Brown, T., Anand, N. and Kennard, O. (1986). Structure of an adenine·cytosine base pair in DNA and its implications for mismatch repair. *Nature*, 320(6062), pp.552-555.

Kalnik, M., Kouchakdjian, M., Li, B., Swann, P. and Patel, D. (1988). Base pair mismatches and carcinogen-modified bases in DNA: an NMR study of A·C and A·O^4meT pairing in dodecanucleotide duplexes. *Biochemistry*, 27(1), pp.100-108.

Kahn, J. and Crothers, D. (1992). Protein-induced bending and DNA cyclization. *Proceedings of the National Academy of Sciences*, 89(14), pp.6343-6347.

Kessler, C. (2000). *Nonradioactive analysis of biomolecules*. Berlin: Springer.

Kim, J., Cheong, C. and Moore, P. (1991). Tetramerization of an RNA oligonucleotide containing a GGGG sequence. *Nature*, 351(6324), pp.331-332.

Kochoyan, M., Lancelot, G. and Leroy, J. (1988). Study of structure, base-pair opening kinetics and proton exchange mechanism of the d-(AATTGCAATT) self-complementary oligodeoxynucleotide in solution. *Nucleic Acids Research*, 16(15), pp.7685-7702.

Koo, H., Drak, J., Rice, J. and Crothers, D. (1990). Determination of the extent of DNA bending by an adenine-thymine tract. *Biochemistry*, 29(17), pp.4227-4234.

Kuehn, B. (2012). 1000 Genomes project finds substantial genetic variation among populations. *The Journal of the American Medical Association*, 308(22), p.2322.

Laos, R., Shaw, R., Leal, N., Gaucher, E. and Benner, S. (2013). Directed evolution of polymerases to accept nucleotides with nonstandard hydrogen bond patterns. *Biochemistry*, 52(31), pp.5288-5294.

Leal, N., Kim, H., Hoshika, S., Kim, M., Carrigan, M. and Benner, S. (2015). Transcription, reverse transcription, and analysis of RNA containing artificial genetic components. *ACS Synthetic. Biology*, 4(4), pp.407-413.

McTigue, P., Peterson, R. and Kahn, J. (2004). Sequence-dependent thermodynamic parameters for Locked Nucleic Acid (LNA)−DNA duplex formation. *Biochemistry*, 43(18), pp.5388-5405.

Mehta, R. and Kahn, J. (1999). Designed hyperstable lac Repressor·DNA loop topologies suggest alternative loop geometries. *Journal of Molecular Biology*, 294(1), pp.67-77.

Nielsen, P., Egholm, M., Berg, R. and Buchardt, O. (1991). Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science*, 254(5037), pp.1497-1500. *Journal of Biotechnology*, 14(3), pp.303-308.

Nonin, S., Leroy, J. and Gueron, M. (1995). Terminal base pairs of oligodeoxynucleotides: imino proton exchange and fraying. *Biochemistry*, 34(33), pp.10652-10659.

Nguyen, C., Grimes, J., Gerasimova, Y. and Kolpashchikov, D. (2011). Molecular-beacon-based tricomponent probe for SNP analysis in folded nucleic acids. *Chemistry - A European Journal*, 17(46), pp.13052-13058.

Patel, D. (1974). Peptide antibiotic-oligonucleotide interactions. Nuclear magnetic resonance investigations of complex formation between actinomycin D and d-ApTpGpCpApT in aqueous solution. *Biochemistry*, 13(11), pp.2396-2402.

Patel, D., Kozlowski, S., Marky, L., Rice, J., Broka, C., Dallas, J., Itakura, K. and Breslauer, K. (1982). Structure, dynamics, and energetics of deoxyguanosine-thymidine wobble base pair formation in the self-complementary d(CGTGAATTCGCG) duplex in solution. *Biochemistry*, **21**(3), pp.437-444.

Petersen, M. and Wengel, J. (2003). LNA: a versatile tool for therapeutics and genomics. *Trends in Biotechnology*, 21(2), pp.74-81.

Piccirilli, J., Benner, S., Krauch, T., Moroney, S. and Benner, S. (1990). Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature*, 343(6253), pp.33-37.

Rappaport, H. (1988). The 6-thioguanine/5-methyl-2-pyrimidinone base pair. *Nucleic Acids Research*, 16(15), pp.7253-7267.

SantaLucia, J., Allawi, H. and Seneviratne, P. (1996). Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, 35(11), pp.3555-3562.

Schroeder, S., Fountain, M., Kennedy, S., Lukavsky, P., Puglisi, J., Krugh, T. and Turner, D. (2003). Thermodynamic stability and structural features of the J4/5 Loop in a pneumocystis carinii Group I intron. *Biochemistry*, 42(48), pp.14184-14196.

Sudmant, P., Rausch, T., Gardner, E., Handsaker, R., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M., Malhotra, A., Stütz, A., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J., Kong, Y., Lameijer, E., McCarthy, S., Flicek, P., Gibbs, R., Marth, G., Mason, C., Menelaou, A., Muzny, D., Nelson, B., Noor, A., Parrish, N., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A., Untergasser, A., Walker, J., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M., McCarroll, S., Mills, R., Gerstein, M., Bashir, A., Stegle, O., Devine, S., Lee, C., Eichler, E. and Korbel, J. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571), pp.75-81.

Switzer, C., Moroney, S. and Benner, S. (1989). Enzymatic incorporation of a new base pair into DNA and RNA. *J. Am. Chem. Soc.*, 111(21), pp.8322-8323.

Switzer, C., Moroney, S. and Benner, S. (1993). Enzymatic recognition of the base pair between isocytidine and isoguanosine. *Biochemistry*, 32(39), pp.10489-10496.

Tor, Y. and Dervan, P. (1993). Site-specific enzymatic incorporation of an unnatural base, N6-(6-aminohexyl)isoguanosine, into RNA. *Journal of the American Chemistry Society*, 115(11), pp.4461-4467.

Turner, D. and Mathews, D. (2009). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38(Database), pp.D280-D282.

von Hippel, P., Johnson, N. and Marcus, A. (2013). 50 years of DNA 'Breathing': Reflections on old and new approaches. *Biopolymers*, *99*(12), 923–954. http://doi.org/10.1002/bip.22347.

Wang, J., Nielsen, P., Jiang, M., Cai, X., Fernandes, J., Grant, D., Ozsoz, M., Beglieter, A. and Mowat, M. (1997). Mismatch-sensitive hybridization detection by peptide nucleic acids immobilized on a quartz crystal microbalance. *Analytical Chemistry*, 69(24), pp.5200-5202.

Yang, Z., Hutter, D., Sheng, P., Sismour, A. and Benner, S. (2006). Artificially expanded genetic information system: a new base pair with an alternative hydrogen bonding pattern. *Nucleic Acids Research*, 34(21), pp.6095-6101.

Yang, Z., Chen, F., Chamberlin, S. and Benner, S. (2009). Expanded Genetic Alphabets in the Polymerase Chain Reaction. *Angewandte Chemie*, 122(1), pp.181-184.

Yang, Z., Chen, F., Alvarado, J. and Benner, S. (2011). Amplification, mutation, and sequencing of a six-letter synthetic genetic system. *Journal of Amemican Chemistry Society*, 133(38), pp.15105-15112.

Yang, Z., Durante, M., Glushakova, L., Sharma, N., Leal, N., Bradley, K., Chen, F. and Benner, S. (2013). Conversion strategy using an expanded genetic alphabet to assay nucleic acids. *Analytical Chemistry*, 85(9), pp.4705-4712.

Zhang, L., Yang, Z., Sefah, K., Bradley, K., Hoshika, S., Kim, M., Kim, H., Zhu, G., Jiménez, E., Cansiz, S., Teng, I., Champanhac, C., McLendon, C., Liu, C., Zhang, W., Gerloff, D., Huang, Z., Tan, W. and Benner, S. (2015). Evolution of functional six-nucleotide DNA. *Journal of American Chemistry Society*, 137(21), pp.6734-6737.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13), pp.3406-3415.