Thermodynamics and NMR of Internal G·T Mismatches in DNA

Hatim T. Allawi and John SantaLucia, Jr.*

Department of Chemistry, Wayne State University, Detroit, Michigan 48202

Received October 16, 1996; Revised Manuscript Received June 18, 1997[⊗]

ABSTRACT: Thermodynamics of 39 oligonucleotides with internal G·T mismatches dissolved in 1 M NaCl were determined from UV absorbance versus temperature profiles. These data were combined with literature values of six sequences to derive parameters for 10 linearly independent trimer and tetramer sequences with G·T mismatches and Watson−Crick base pairs. The G·T mismatch parameters predict ΔG°_{37}, ΔH°, ΔS°, and T_M with average deviations of 5.1%, 7.5%, 8.0%, and 1.4 °C, respectively. These predictions are within the limits of what can be expected for a nearest-neighbor model. The data show that the contribution of a single G·T mismatch to helix stability is context dependent and ranges from +1.05 kcal/mol for AGA/TTT to −1.05 kcal/mol for CGC/GTG. Several tests of the applicability of the nearest-neighbor model to G·T mismatches are described. Analysis of imino proton chemical shifts show that structural perturbations from the G·T mismatches are highly localized. One-dimensional NOE difference spectra demonstrate that G·T mismatches form stable hydrogen-bonded wobble pairs in diverse contexts. Refined nearest-neighbor parameters for Watson−Crick base pairs are also presented.

Mismatches occur naturally in DNA as a result of errors from misincorporation of bases during replication (Goodman et al., 1993), due to heteroduplex formation during homologous recombination (Bhattacharyya et al., 1989), and from mutagenic chemicals (Leonard et al., 1990a; Plum et al., 1995), ionizing radiation (Brown, 1995), and spontaneous deamination. In addition to Watson−Crick base pairing, there are eight possible mispairs, namely A·A, A·C, A·G, C·C, C·T, G·G, G·T, and T·T. Repair of these mismatches requires the recognition and excision of mismatched bases by proofreading enzymes or by postreplication mismatch repair systems (Modrich & Lahue, 1996). Understanding the thermodynamics of mismatches in DNA duplexes will improve our understanding of these processes (Aboul-ela et al., 1985; Werntges et al., 1986; Petruska et al., 1988; Mendelman et al., 1989; Johnson, 1993).

Several molecular biological techniques require accurate prediction of hybridization thermodynamics to "matched" versus "mismatched" sites (Wallace et al., 1979; Aboul-ela et al., 1985; Kawase et al., 1986; Ikuta et al., 1987) including PCR[1] (Saiki et al., 1988), Kunkel mutagenesis (Kunkel et al., 1987), sequencing by hybridization (Fodor et al., 1993), and gene diagnostics (Freier, 1993). In each of these techniques, the choice of a nonoptimal sequence or temperature can lead to amplification or detection of wrong sequences (Steger, 1994; SantaLucia et al., 1996). In addition, knowledge of mismatch stability is an important step toward acquiring a parameter database for DNA secondary-structure prediction algorithms (Allawi, Peyret, and SantaLucia, unpublished experiments).

Previously, we and others (SantaLucia et al., 1996; Sugimoto et al., 1994; Doktycz et al., 1995) showed that, despite the structural variability observed in DNA structures (Callidine & Drew, 1984; Hunter, 1993), a nearest-neighbor model is sufficient to reliably predict the stability of DNA duplexes with Watson−Crick pairs. We hypothesized that a nearest-neighbor model could also apply for DNA duplexes with internal G·T mismatches. To test this hypothesis, thermodynamic measurements of 39 G·T mismatch-containing DNA oligonucleotides were combined with six literature values to derive G·T mismatch nearest-neighbor parameters in 1 M NaCl buffer. The availability of nearest-neighbor parameters for G·T mismatches along with refined parameters for Watson−Crick pairs allows the reliable prediction of duplex stability from sequence. Exchangeable proton one-dimensional NMR spectra show that G·T mismatches form a wobble hydrogen-bonded structure in diverse contexts.

MATERIALS AND METHODS

DNA Synthesis and Purification. Oligonucleotides were supplied by Hitachi Chemical Research and were synthesized on solid support using standard phosphoramidite chemistry (Brown & Brown, 1991). DNA oligomers were removed from the solid support and deblocked by treatment with concentrated ammonia at 50 °C overnight. Each sample was evaporated to dryness, and the crude mixture was dissolved in 250 mL of water and purified on a Si500F thin-layer chromatography plate (Baker) by eluting for 5 h with *n*-propanol/ammonia/water (55:35:10 by volume) (Chou et al., 1989). Bands were visualized with a UV lamp, and the least mobile band was cut out and eluted three times with 3 mL of distilled deionized water. The sample was then evaporated to dryness. Oligonucleotides were desalted and further purified with a Sep-pak C-18 cartridge (Waters). The DNA was eluted with 30% acetonitrile buffered with 10 mM ammonium bicarbonate, pH 7.0. Purities were checked by analytical C-8 HPLC (Perceptive Biosystems) and were greater than 95%.

* To whom correspondence should be sent. Phone: (313) 577-0101. FAX: (313) 577-8822. E-mail: jsl@chem.wayne.edu.

[⊗] Abstract published in *Advance ACS Abstracts,* August 1, 1997.

[1] Abbreviations: Na_2EDTA, disodium ethylenediaminetetraacetate; eu, entropy units (cal/K mol); HPLC, high-performance liquid chromatography; NOE, nuclear Overhauser enhancement; PCR, polymerase chain reaction; SVD, singular value decomposition; T_M, melting temperature; UV, ultraviolet.

S0006-2960(96)02590-1 CCC: $14.00

Melting Curves. Absorbance versus temperature profiles (melting curves) were measured at 280 or 260 nm with a heating rate of 0.8 °C min^{-1} on an AVIV 14DS UV−vis spectrophotometer as described previously (SantaLucia et al., 1996). The buffer used for thermodynamic studies was 1.0 M NaCl, 10 mM sodium cacodylate, and 0.5 mM Na_2EDTA, pH 7.0. Oligonucleotide samples were "annealed" and degassed by raising the temperature to 85 °C for 5 min. While at 85 °C, the absorbance of each sample was measured at 260 nm for determination of oligonucleotide concentrations (C_T) using extinction coefficients calculated from dinucleoside monophosphates and nucleotides, as described previously (Richards, 1975).

Data Analysis. Thermodynamic parameters for duplex formation were obtained from absorbance versus temperature melting curves using the program MELTWIN v2.1 (McDowell & Turner, 1996) by two methods: (1) enthalpies and entropies from fits of individual melting curves with sloping base lines were averaged (Petersheim & Turner, 1983), and (2) plots of reciprocal melting temperature (T_M^{-1}) vs ln C_T according to eq 1 (Borer et al., 1974) were made.

$$T_M^{-1} = R/\Delta H^\circ \ln C_T + \Delta S^\circ/\Delta H^\circ \quad (1)$$

For non-self-complementary molecules, C_T in eq 1 was replaced by $C_T/4$. Both methods assume that the transition equilibrium involves only two states (i.e., duplex and random coil) and that the difference in heat capacities (ΔC_p°) of these states is zero (Petersheim & Turner, 1983; Freier et al., 1986). Agreement of parameters derived by the two methods is a necessary, but not sufficient, criterion to establish the validity of the two-state approximation (see below; SantaLucia et al., 1990; Marky & Breslauer, 1987).

Design of Sequences. Sequences were designed to have a melting temperature (T_M) between 30 and 60 °C and to minimize the possibility of forming stable alternative secondary structures such as slipped duplexes or hairpins; this maximizes the likelihood of observing two-state thermodynamics. In addition, sequences were chosen to provide uniform representation of the 11 different G·T mismatch containing nearest-neighbors. Throughout this paper nearest-neighbor base pairs are represented with a slash separating the strands in antiparallel orientation and with mismatches underlined (e.g., A<u>T</u>/T<u>G</u> means $^{5'}AT^{3'}$ paired with $^{3'}TG^{5'}$). The 11 G·T nearest-neighbors occur in this study with the following frequencies: A<u>G</u>/T<u>T</u> = 12, A<u>T</u>/T<u>G</u> = 12, C<u>G</u>/G<u>T</u> = 17, C<u>T</u>/G<u>G</u> = 12, G<u>G</u>/C<u>T</u> = 9, G<u>T</u>/C<u>G</u> = 20, T<u>G</u>/A<u>T</u> = 23, T<u>T</u>/A<u>G</u> = 17, <u>GT</u>/<u>TG</u> = 3, <u>TT</u>/<u>GG</u> = 3, <u>TG</u>/<u>GT</u> = 3. 18 of the 45 sequences used to derive nearest-neighbor parameters form self-complementary duplexes with two nonadjacent G·T mismatches.

Three-State Equilibrium Calculations. Some of the sequences in this study are better described by a three-state model rather than a two-state model. We follow the method described by Longfellow et al. (1990) to carry out three-state equilibrium calculations. Self-complementary sequences have the potential to form both duplex and hairpin species as described by the following coupled equilibria:

$$2A \rightleftharpoons A_2 \qquad K_1 = [A_2]/[A]^2 \quad (2a)$$

$$A \rightleftharpoons A_H \qquad K_2 = [A_H]/[A] \quad (2b)$$

where A, A_2, and A_H represent A in the random coil, duplex, and hairpin states, respectively. The total strand concentration, C_T, is given by

$$C_T = [A] + 2[A_2] + [A_H] \quad (3)$$

and the equilibrium constants are given by

$$K_i = \exp(-\Delta H^\circ_i/RT + \Delta S^\circ_i/R) \quad (4)$$

where $i = 1$ or 2 and ΔH°_i and ΔS°_i are measured or predicted thermodynamic parameters for the individual equilibria. Substituting eqs 3 and 2b into eq 2a gives the quadratic equation

$$0 = C_T - (1 + K_2)[A] - 2K_1[A]^2 \quad (5)$$

The concentration of A is given by the analytical solution of the quadratic equation (only one root is physical). The concentrations of A_2 and A_H are calculated from eqs 2a and 2b. Similar analytical solutions can be derived for non-self-complementary sequences that have hairpin intermediates (Allawi and SantaLucia, unpublished results).

Determination of the G·T Mismatch Contribution to Helix Formation. The thermodynamic increments associated with the folding of the mismatch portion of an oligonucleotide cannot be directly measured. Instead, each thermodynamic measurement provides the total energy change for strands going from the random coil state to duplex state. According to the nearest-neighbor model, the total energy change is the sum of energy increments for helix initiation (see below), helix symmetry, and nearest-neighbor interactions between base pairs (Freier et al., 1986). The nearest-neighbor model can be extended to include parameters for interactions between mismatches and neighboring base pairs (He et al., 1991). For example (underlined residues are mismatched):

```
5'-GGACTGACG-3' = initiation + symmetry +
3'-CCTGGCTGC-5'

    GG + GA + AC  + CT + CG + GA + AC + CG
    CC   CT   TG    GG   GT   CT   TG   GC      (6)
```

Thus, to derive the mismatch contribution to duplex formation, ΔG°_{37}(mismatch), the contributions from the Watson−Crick pairs, helix initiation, and helical symmetry (for self-complementary sequences) are subtracted from the total free energy. In the example above, this amounts to simply rearranging eq 6:

```
CT + CG = GGACTGACG - initiation - symmetry -
GG   GT   CCTGGCTGC

              GG - GA - AC - GA - AC - CG
              CC   CT   TG   CT   TG   GC      (7)
```

Inserting the experimental free energy (Table 3) and the Watson−Crick nearest-neighbor numbers (Table 1) into eq 7, we obtain

```
CT + CG = -8.37 - 2 (0.98) - 0 - (-1.84) - (-1.30) - (-1.44) -
GG   GT

                (-1.30) - (-1.44) - (-2.17)

= -0.84 kcal/mol = ΔG°37(mismatch)                    (8)
```

The nearest-neighbors C<u>T</u>/G<u>G</u> and C<u>G</u>/G<u>T</u> in eq 8 are

Table 1: Nearest-Neighbor Thermodynamic Parameters for Watson−Crick Base Pair Formation in 1 M NaCl[a]

propagation sequence	$\Delta H°$ (kcal/mol)	$\Delta S°$ (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[b]	0.1 ± 1.1	−2.8 ± 0.2	0.98 ± 0.05
init. w/term. A−T[b]	2.3 ± 1.3	4.1 ± 0.2	1.03 ± 0.05
symmetry correciton	0	−1.4	0.4

[a] Errors are resampling standard deviations (see text). [b] See text for how to apply the initiation parameters.

unknowns. Similar calculations for $\Delta H°$ and $\Delta S°$ are carried out to calculate $\Delta H°$(mismatch) and $\Delta S°$(mismatch).

Alternatively, the mismatch contribution can be determined by measuring the thermodynamics of a "core sequence" and adding back the nearest neighbor that is interrupted by the mismatch (Wu et al., 1995), as shown below:

```
CTG = GGACTGACG - GGACGACG + CG
GGC   CCTGGCTGC   CCTGCTGC   GC          (9)
```

where CTG/GGC = CT/GA+ CG/GT. The two methods in eqs 7 and 9 are equally reliable due to uncertainties in measurements and in the nearest neighbor parameters (Table 1). We chose the former method because the measurement of the thermodynamics of all the core sequences is not required.

Analysis of the G·T Mismatch Contribution in Terms of Linearly Independent Sequences. Previous work by Gray and Tinoco (1970), Vologodskii et al. (1984), and Goldstein and Benight (1992) have shown that nucleic acid thermodynamics should be analyzed in terms of linearly independent sequences. For duplexes with Watson−Crick base pairs, there are 10 different nearest neighbors and two initiation parameters that can be determined from a carefully designed set of oligonucleotides (D. M. Gray, personal communication). If constraints are imposed on all the sequences in the data set, however, such as fixing the oligonucleotide length, fixing the ends of the duplex, or making measurements in polymers that formally do not contain ends, then each sequence constraint results in one fewer parameter that can be derived from the set of sequences (Gray & Tinoco, 1970). A fundamental assumption of the nearest-neighbor model is that terminal nearest neighbors make the same contribution as internal neighbors. For Watson−Crick pairs in RNA and DNA (see below), it appears that this assumption is reasonable as evidenced by the fact that the nearest-neighbor parameters make accurate predictions for sequences with diverse nucleotide content and termini. However, for G·T mismatches in DNA this assumption is invalid. Terminal G·T mismatches always make favorable contributions to helix stability (Jenkins and SantaLucia, unpublished results), while internal G·T mismatches can make favorable or unfavorable contributions, depending on the context. Hence, our data set for internal G·T mismatches does not contain sequences that have terminal G·T mismatches. As a result, the maximum number of linearly independent parameters that can be derived from our data set is 10. This is verified from the column rank of the stacking matrix, which is 10 for our data set. These 10 uniquely determined parameters are linear combinations of the 11 G·T nearest-neighbor dimers. A simple way to construct a set of sequences that are linearly independent is to start with the 11 G·T mismatch containing dimers and add a Watson−Crick base pair to each end of the base pair doublet that has a G·T mismatch. For example, the neighbor AT/TG becomes ATC/TGG (again, the "/" indicates pairing of strands in the antiparallel orientation) by the addition of a $^{5'}$G−C$^{3'}$ base pair to the left side. The trimer sequence ATC/TGG is simply the sum of the nearest neighbors AT/TG and GG/CT. For neighbors with two G·T mismatches such as GT/TG, the linearly independent sequence is GGTC/CTGG which is the sum of GG/CT + GT/TG and GG/CT. When this is done for all 11 nearest neighbors, two of the trimers are identical (GTC/CGG occurs twice) so the total number of unique parameters is reduced to 10 and these are linearly independent (see Tables 4 and 5). The choice of which base pair to place at the left is arbitrary, but placing the same pair at the end of each sequence simplifies the analysis of sequences in terms of linearly independent sequences. For example, eq 6 can be rewritten in terms of the following linearly independent sequences:

```
5'-GGACTGACG-3' = initiation + symmetry + GG +
3'-CCTGGCTGC-5'                           CC

 GA + AC  + CTC + CGC - GTC + GA + AC + CG
 CT   TG    GGG   GTG   CGG   CT   TG   GC   (10)
```

Note the term GTC/GGC is subtracted to account for the extra terms for TC/GG and GC/TG that are found in the sequences CTC/GGG and CGC/GTG but are not found in the actual sequence. This equation can be rearranged to give the mismatch contribution to helix stability.

```
CTG = CTC + CGC - GTC =
GGC   GGG   GTG   CGG

  5'-GGACTGACG-3' - initiation - symmetry -
  3'-CCTGGCTGC-5'

        GG - GA - AC - GA - AC - CG
        CC   CT   TG   CT   TG   GC        (11)
```

It is important to note that the 10 linearly independent parameters obtained still conform to a nearest-neighbor model and do not account for next-nearest-neighbor interactions.

Duplex Initiation Parameter. Recently, it has been shown that the duplex initiation parameter contains contributions from the terminal base pairs (D. M. Gray, personal communication). Gray's work shows that the difference between DNA sequences with terminal G−C and T−A base pairs can be accounted for by introducing two initiation parameters. This model assumes that the contribution to duplex stability for terminal G−C equals terminal C−G and terminal A−T equals terminal T−A. To accomodate this, two parameters are introduced: "initiation with terminal G−C" and "initiation with terminal A−T" (Table 1). A duplex with two terminal G−C pairs would use total initiation = 2 × (initiation with terminal G−C). A duplex with one terminal G−C and one terminal A−T would use total initiation =

Table 2: Thermodynamics of Duplex Formation of Oligonucleotides with G·T Mismatches[a]

DNA duplex	1/T_M vs ln C_T parameters				curve fit parameters		
	-$\Delta G°_{37}$ (kcal/mol)	-$\Delta H°$ (kcal/mol)	-$\Delta S°$ (eu)	T_M (°C)[b]	-$\Delta G°_{37}$ (kcal/mol)	-$\Delta H°$ (kcal/mol)	-$\Delta S°$ (eu)
Molecules with Two-State Transitions							
CGTGTCTCC GCACGGAGG	7.75 ± 0.60	51.5 ± 2.1	141.0 ± 4.7	50.0	7.79 ± 0.10	52.6 ± 3.1	144.3 ± 9.7
GGACGCTCG CCTGTGAGC	8.49 ± 0.51	63.9 ± 2.0	178.5 ± 5.1	51.3	8.44 ± 0.10	61.1 ± 2.9	169.8 ± 9.1
GGACTCTCG CCTGGGAGC	7.47 ± 0.21	57.7 ± 0.8	162.1 ± 2.0	47.0	7.47 ± 0.04	58.1 ± 1.6	163.3 ± 5.3
GGACTGACG CCTGGCTGC	8.30 ± 0.63	59.5 ± 2.4	165.2 ± 5.8	51.3	8.44 ± 0.15	65.7 ± 2.9	184.5 ± 9.1
GGAGTCACG CCTCGGTGC	8.84 ± 0.50	66.3 ± 2.1	185.3 ± 5.1	52.5	8.79 ± 0.08	64.5 ± 1.8	179.6 ± 5.6
GACCGTGCAC CTGGTGCGTG	7.17 ± 0.41	55.8 ± 1.6	156.8 ± 3.8	45.5	7.18 ± 0.08	50.9 ± 3.0	140.9 ± 9.9
GACGTGACCG CTGCGTTGGC	8.06 ± 0.52	63.4 ± 2.0	178.3 ± 5.3	49.1	8.09 ± 0.08	64.5 ± 2.0	181.9 ± 6.4
GACGTTGGAC CTGCGGCCTG	7.91 ± 0.40	59.1 ± 1.6	164.9 ± 3.8	49.2	7.97 ± 0.08	61.8 ± 2.4	173.6 ± 7.5
GACGTTAGGC CTGCGGTCCG	7.66 ± 0.25	46.4 ± 0.8	124.9 ± 1.8	50.9	7.67 ± 0.07	45.6 ± 3.6	122.2 ± 11.5
GCAGGTCTGC	6.38 ± 0.74	36.8 ± 2.1	98.1 ± 4.5	43.1	6.56 ± 0.07	39.7 ± 9.9	106.9 ± 31.8
GGAGTGCTCC	7.66 ± 0.21	66.9 ± 1.0	191.0 ± 2.5	46.5	7.61 ± 0.05	64.4 ± 1.6	183.2 ± 5.0
GGCAGTTCGC CCGTTGAGCG	6.87 ± 0.51	54.3 ± 1.9	152.8 ± 4.6	43.9	6.86 ± 0.08	54.9 ± 1.7	154.9 ± 5.5
CATGAGGCTAC GTACTTCGATG	8.57 ± 0.56	68.1 ± 2.4	192.0 ± 6.0	50.8	8.65 ± 0.14	71.7 ± 3.0	203.3 ± 9.1
CATGTGACTAC GTACATTGATG	7.71 ± 0.35	64.6 ± 1.5	183.5 ± 3.7	47.1	7.72 ± 0.05	63.7 ± 3.0	180.6 ± 9.5
GATCATTGTAC CTAGTGACATG	7.78 ± 0.29	66.1 ± 1.3	188.1 ± 3.2	47.2	7.84 ± 0.08	72.4 ± 2.1	208.1 ± 6.7
GATCTGTGTAC CTAGATACATG	7.78 ± 0.38	68.6 ± 1.7	196.0 ± 4.3	46.7	7.83 ± 0.07	72.7 ± 2.8	209.0 ± 8.8
GTAGCGTCATG CATCGTAGTAC	9.76 ± 0.46	75.6 ± 2.1	212.3 ± 5.1	54.7	9.84 ± 0.11	77.7 ± 1.7	218.7 ± 5.1
GTAGTGACATG CATCATTGTAC	7.88 ± 0.98	69.4 ± 4.5	198.3 ± 11.3	47.2	7.87 ± 0.12	67.1 ± 2.8	191.1 ± 8.6
CCATGCGTAACG GGTATGCGTTGC	8.94 ± 0.37	69.5 ± 1.6	195.4 ± 3.8	52.2	9.02 ± 0.08	72.9 ± 2.7	205.9 ± 8.5
CGAGACGTTTCG	6.96 ± 0.46	62.8 ± 2.1	179.9 ± 5.3	43.5	6.99 ± 0.06	59.2 ± 4.2	168.3 ± 13.6
CGAGCATGTTCG	7.20 ± 0.72	60.6 ± 3.2	172.1 ± 7.9	45.0	7.25 ± 0.07	59.1 ± 5.3	167.1 ± 16.9
CGTGACGTTACG	8.19 ± 0.42	76.4 ± 2.1	219.9 ± 5.5	47.6	8.07 ± 0.08	70.1 ± 3.3	200.1 ± 10.6
CGTGTCGATACG	8.42 ± 1.29	75.5 ± 6.6	216.1 ± 17.1	48.7	8.36 ± 0.16	72.2 ± 3.7	205.7 ± 11.7
CGTTACGTGACG	7.86 ± 0.33	65.7 ± 1.5	186.6 ± 3.8	47.7	7.87 ± 0.08	64.1 ± 3.9	181.4 ± 12.4
CTCGGATCTGAG	8.49 ± 0.28	77.6 ± 1.4	222.8 ± 3.7	48.7	8.36 ± 0.04	72.3 ± 1.7	206.2 ± 5.6
CTCTCATGGGAG	6.46 ± 0.29	51.7 ± 1.2	145.8 ± 3.0	41.8	6.47 ± 0.03	49.0 ± 4.4	137.2 ± 14.1
CTCTGATCGGAG	7.54 ± 0.40	63.3 ± 1.8	179.7 ± 4.4	46.4	7.50 ± 0.05	57.4 ± 5.1	160.8 ± 16.5
CTGTCATGGCAG	8.30 ± 0.50	60.8 ± 2.0	169.2 ± 5.0	51.0	8.30 ± 0.21	57.7 ± 8.4	159.1 ± 26.5
CTGTGATCGCAG	8.74 ± 0.32	70.6 ± 1.5	199.5 ± 3.8	51.1	8.56 ± 0.09	63.4 ± 5.1	176.9 ± 16.3
CTTGGATCTAAG	5.89 ± 0.25	66.4 ± 1.4	195.0 ± 3.6	38.0	5.92 ± 0.04	61.6 ± 1.5	179.4 ± 4.9

(initiation with terminal G−C) + (initiation with terminal A−T). A duplex with two terminal A−T pairs would use total initiation = 2 × (initiation with terminal A−T).

Regression Analysis. The free energy and enthalpy contributions of the 10 linearly independent G·T mismatch containing trimer and tetramer sequences were determined by multiple linear regression using MATHEMATICA (Wolfram 1992). Thermodynamic parameters for 45 G·T mismatch containing duplexes (Table 2) were used to construct a list of 45 equations (analogous to eq 11 above) with 10 unknowns. These equations were then cast in the form of matrices for input into linear regression. The $\Delta G°_{37}$-(mismatch) for all 45 sequences formed the column matrix $\mathbf{G}_{Mis}$ with elements ΔG_i, where the subscript i denotes different oligonucleotides. The number of occurrences of each G·T mismatch containing trimer or tetramer sequence formed the "stacking matrix," **S**, with dimensions of 45 × 10. The unknown values of the 10 G·T mismatch trimers and tetramers form the column matrix $\mathbf{G}_{NN}$ with elements ΔG_j, where the subscript j denotes the 10 linearly independent sequences. Therefore, the data for all the sequences are written as

$$\mathbf{G}_{Mis} = \mathbf{S}\cdot\mathbf{G}_{NN} \quad (12)$$

The solution of eq 12 for the unknowns, $\mathbf{G}_{NN}$, was obtained using singular value decomposition (SVD) (Press et al., 1989) which effectively inverts the stacking matrix and minimizes the error weighted squares of the residuals (Bevington, 1969):

$$\chi^2 = \Sigma_{ij}|(\Delta G_i - S_{ij}\Delta G_j)/\sigma_i|^2 \quad (13)$$

where σ_i are the propagated errors in ΔG_i, and S_{ij} are the matrix elements of **S**. The σ_i were calculated as the square

Table 2 (Continued)

DNA duplex	$1/T_M$ vs ln C_T parameters				curve fit parameters		
	$-\Delta G^\circ_{37}$ (kcal/mol)	$-\Delta H^\circ$ (kcal/mol)	$-\Delta S^\circ$ (eu)	T_M (°C)[b]	$-\Delta G^\circ_{37}$ (kcal/mol)	$-\Delta H^\circ$ (kcal/mol)	$-\Delta S^\circ$ (eu)
Molecules with Marginally Non-Two-State Transitions							
CGTCTGTCC GCAGGCAGG	8.01 ± 0.80	53.1 ± 2.6	145.2 ± 5.9	51.3	8.06 ± 0.10	60.1 ± 1.2	167.7 ± 4.1
GACTGGAGAG CTGATTTCTC	4.61 ± 0.19	42.2 ± 0.7	121.2 ± 1.6	29.3	4.16 ± 0.17	49.9 ± 2.3	147.4 ± 7.7
GCGATGTCGC	7.98 ± 0.54	59.7 ± 2.3	166.6 ± 5.5	49.5	7.85 ± 0.21	50.3 ± 11.1	136.7 ± 35.0
GATCTTTGTAC CTAGAGACATG	7.29 ± 0.25	63.8 ± 1.1	182.2 ± 2.7	45.1	7.32 ± 0.08	71.4 ± 2.7	206.7 ± 8.5
CGAGTCGATTCG	7.71 ± 0.57	74.2 ± 2.9	214.3 ± 3.2	45.7	7.59 ± 0.12	65.0 ± 3.5	185.1 ± 11.2
CTTGCATGTAAG	6.10 ± 0.42	60.9 ± 2.1	176.8 ± 5.3	39.2	6.17 ± 0.12	52.4 ± 3.9	148.9 ± 12.9
CGTGTCTAGATACG	9.40 ± 0.41	84.0 ± 2.1	240.7 ± 5.6	51.4	9.09 ± 0.10	73.6 ± 5.4	208.1 ± 17.2
Molecules with Non-Two-State Transitions							
GACGTGAGGC CTGCGTTCCG	6.49 ± 0.48	31.4 ± 1.1	80.2 ± 2.0	45.23	6.39 ± 0.10	39.8 ± 3.8	107.8 ± 12.1
CGTTGCGTAACG	7.91 ± 0.44	60.3 ± 1.8	168.9 ± 4.5	48.9	7.87 ± 0.08	57.1 ± 4.5	158.7 ± 14.2
CTCGCATGTGAG	8.48 ± 0.21	75.9 ± 1.1	217.3 ± 2.8	48.9	7.90 ± 0.03	53.4 ± 6.2	146.0 ± 19.9
CTGGCATGTCAG	8.73 ± 0.31	75.4 ± 1.5	215.0 ± 3.8	50.1	8.34 ± 0.10	60.2 ± 1.9	167.3 ± 6.1
CTGGGATCTCAG	7.81 ± 0.36	81.3 ± 2.1	236.9 ± 5.6	45.4	7.50 ± 0.19	66.4 ± 2.1	189.8 ± 6.0
GCGTACGCATGCG CGCATGTGTACGC	12.96 ± 0.22	85.7 ± 1.0	234.38 ± 2.3	65.8	12.33 ± 0.07	76.1 ± 1.2	205.7 ± 3.9

[a] Listed in alphabetical order and by oligomer length. For self-complementary sequences only the top strand is given. For non-self-complementary duplexes, both strands are given in antiparallel orientation. Underlined residues are mismatched. Molecules listed as two-state had ΔH° agreement within 10% by two different methods. Molecules listed as marginally non-two-state had ΔH° agreement between 10 and 20% by two different methods. Molecules listed as non-two-state had ΔH° disagreement greater than 20% by two different methods. Solutions are 1 M NaCl, 10 mM sodium cacodylate, 0.5 mM Na_2EDTA, pH 7. Errors are standard deviations from the regression analysis of the melting data. Extra significant figures are given to allow accurate calculation of ΔG°_{37} and T_M. [b] Calculated for 10^{-4} M oligomer concentration for self-complementary sequences and 4×10^{-4} M for non-self-complementary sequences.

root of the sum of the squares of the errors for ΔG_i(Total) and the standard errors for the nearest neighbors (Table 1). Analogous calculations were performed to obtain G·T mismatch nearest-neighbor ΔH° parameters. Entropic parameters for G·T mismatch contributions were calculated from ΔG°_{37} and ΔH° using the equation

$$\Delta S^\circ = (\Delta H^\circ - \Delta G^\circ_{37})/310.15 \quad (14)$$

To verify our calculation methodology, we derived the G·T mismatch nearest-neighbor entropic contributions as it was done for ΔG°_{37} and ΔH°, using SVD, and the results agreed with those obtained using eq 14.

Error Analysis. The sampling errors reported in Table 2 for $1/T_M$ vs log C_T plots and for the fits of the shapes of melting curves were obtained by standard methods (SantaLucia et al., 1991; McDowell & Turner, 1996) and reflect the precision or reproducibility in the experimental measurement (Bevington, 1969). The accuracy of the ΔG°_{37}, ΔH°, and ΔS° parameters derived from the van't Hoff analysis of the UV melting curves are estimated as standard deviations of 4%, 5%, and 6%, respectively. These estimates are based on the typical agreement between model independent calorimetry and UV melting measurements of ΔH° and ΔS° (Albergo et al., 1981) and by measurements made by different laboratories on the same sequences (J. SantaLucia, unpublished results). The small errors observed for ΔG°_{37} and T_M are the result of the fact that ΔH° and ΔS° determined from a van't Hoff analysis of UV melting data are greater than 99% correlated (Petersheim & Turner, 1983; SantaLucia et al., 1991). Plots of experimental ΔH° vs ΔS° are provided in the Supporting Information. The error propagation from ΔH° and ΔS° to ΔG°_{37} and T_M using standard methods [eq 4.8 of Bevington (1969)] are given by the following equations (SantaLucia et al., 1991):

$$(\sigma_{\Delta G^\circ_{37}})^2 = (\sigma_{\Delta H^\circ})^2 + T^2(\sigma_{\Delta S^\circ})^2 - 2T(R_{\Delta H^\circ \Delta S^\circ})\sigma_{\Delta H^\circ}\sigma_{\Delta S^\circ} \quad (15)$$

$$(\sigma_{T_m})^2 = (\sigma_{\Delta H^\circ}T_M/\Delta H^\circ)^2 + (\sigma_{\Delta S^\circ}T_M{}^2/\Delta H^\circ)^2 - 2T_M{}^3R_{\Delta H^\circ \Delta S^\circ}\sigma_{\Delta H^\circ}\sigma_{\Delta S^\circ}/(\Delta H^\circ)^2 \quad (16)$$

where $R_{\Delta H^\circ \Delta S^\circ}$ is the correlation coefficient between ΔH° and ΔS° and the σ terms are the standard deviations in the measurements. The errors in experimental measurements are rigorously propagated to the nearest-neighbor parameters in the variance-covariance matrix given by the SVD analysis (Press et al., 1989; SantaLucia et al., 1996). The propagated errors in the nearest-neighbor parameters have been independently confirmed by resampling analysis of the data.

Resampling Analysis of the Data. Since our data set contains 45 equations with 10 unknowns, the problem is overdetermined. We took advantage of this and used a resampling analysis (Efron & Tibshirani, 1993) of our data to determine the uncertainties of the 10 linearly independent sequences. We performed 30 resampling trials. For each trial, a different set of 35 randomly selected sequences was used in the SVD analysis to calculate the 10 unknowns. For each trial, the rank of the matrix was confirmed to be 10. Then for each of the 10 unknowns the 30 trial values were averaged and the standard deviations determined. This

resampling analysis was performed for ΔG°_{37}, ΔH°, and ΔS°. The errors obtained from the resampling analysis have the advantage that no assumption about the magnitudes of the experimental errors or knowledge of the correct method of error propagation are required, and yet highly reliable error estimates are obtained (Efron & Tibshirani, 1993).

^{1}H-NMR Spectroscopy. Oligomers were dissolved in 90% H_2O and 10% D_2O with 1 M NaCl, 10 mM disodium phosphate, and 0.1 mM Na_2EDTA at pH 7. Sample concentrations were between 0.2 and 1.0 mM. ^{1}H-NMR spectra at 10 °C were recorded using a Varian Unity 500 MHz NMR spectrometer. One dimensional exchangeable proton NMR spectra were recorded using the WATERGATE pulse sequence with "flip-back" pulse to suppress the water peak (Piotto et al., 1992; Lippens et al., 1995). Spectra were recorded with the carrier placed at the solvent frequency and with high-power and low-power pulse widths of 8.8 and 1700 ms, sweep width of 12 kHz, gradient field strength of 10.0 G/cm, and duration of 1 ms. 512 transients were collected for each spectrum. Data were multiplied by a 4.0 Hz line-broadening exponential function and Fourier transformed by a Silicon Graphics Indigo^{2}Extreme computer with Varian VNMR software. No base line correction or solvent subtraction was applied. 3-(Trimethylsilyl)propionic-2,2,3,3-d_4 acid (TSP) was used as the internal standard for chemical shift reference. 1D-NOE difference spectra were acquired as described above, but with selective decoupling of individual resonances during the 1 s recycle delay. Each resonance was decoupled with a power sufficient to saturate <80% of the signal intensity so that spillover artifacts would be minimized. The spectra were acquired in an interleaved fashion in blocks of 16 scans to minimize subtraction errors due to long-term instrument drift, and 3200−6400 scans were collected for each FID.

RESULTS

Watson−Crick Nearest-Neighbor Parameters. Recently, two groups independently published improved nearest-neighbor parameters for predicting DNA duplex stability (SantaLucia et al., 1996; Sugimoto et al., 1996). The nearest-neighbor parameters derived by the two groups are similar in many respects; but, both the initiation parameters and the CG/GC neighbors are different [SantaLucia et al. reported ΔG°_{37}(initiation) = +1.82 kcal/mol and ΔG°_{37}(CG/GC) = −2.09 kcal/mol, whereas Sugimoto et al. reported +3.4 and −2.8 kcal/mol, respectively]. We have determined that these discrepancies are primarily due to two factors: (1) Sugimoto's regression analysis method did not produce the linear least squares fit, and (2) Sugimoto included data for the sequence CGCGTACGCGTACGCG (Raap et al., 1985) which is a clear outlier in the fit. We also reported this sequence in our paper (SantaLucia et al., 1996) but did not include it in the regression analysis since it has a T_M of 91 °C (at a strand concentration of 1×10^{-4} M) which leads to a large uncertainty in the derived thermodynamic parameters and also makes it unlikely that the two-state approximation is valid. When we removed the sequence CGCGTACGCGTACGCG and performed a least squares fit on the remainder of Sugimoto's data set (64 sequences), the initiation parameter obtained is +2.34 kcal/mol and the CG/GC neighbor is −2.37 kcal/mol, which agree with what we reported (SantaLucia et al., 1996). Two important results of Sugimoto et al. (1996) are that (1) sequences with terminal T−A base pairs behave similarly as sequences with terminal G−C pairs, and (2) the helix initiation parameter for sequences with only A−T pairs appears to be the same as that for sequences with mixed G−C and A−T pairs. Our earlier conclusions for these parameters (SantaLucia et al., 1996) were incorrect due to insufficient/incorrect literature data for sequences with terminal A−T pairs.

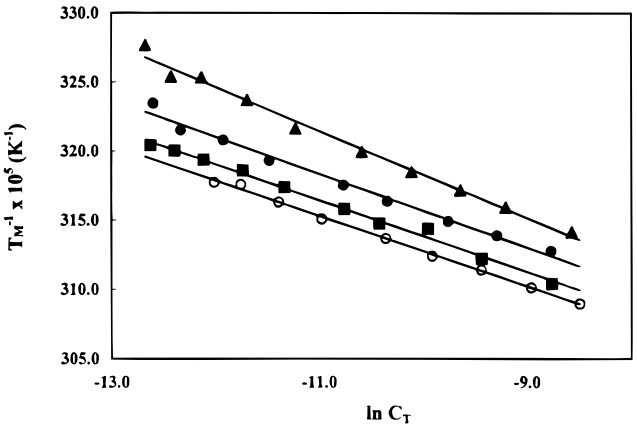

FIGURE 1: Reciprocal melting temperature vs ln C_T plots for CGAGTCGATTCG (▲), CGAGACGTTTCG (●), CGTGACGTTACG (■), and CTCGGATCTGAG (○).

In order to provide a unified set of thermodynamic parameters, we have combined data from Sugimoto's laboratory (58 sequences), data from our laboratory (38 sequences), and data from the literature from other laboratories (37 sequences) for a total of 131 sequences (see Supporting Information). 108 of these sequences that melted with two-state thermodynamics were used to derive new nearest-neighbor parameters (Table 1). The errors reported in Table 1 are from a resampling analysis of the data in which 30 trials with 78 sequences randomly selected from the 108 sequences were used to calculate the nearest-neighbor parameters (see Materials and Methods). For each trial the rank of the stacking matrix was verified to be 12. All of the parameters given in Table 1, including the initiation parameter and the CG/GC neighbor, agree within the reported error of our previously published parameters (SantaLucia et al., 1996). The parameters given in Table 1 predict the ΔG°_{37}, ΔH°, ΔS°, and T_M of the 108 sequences with average deviations of 3.9%, 6.4%, 7.5%, and 1.8 °C, respectively. A complete table with the experimental versus predicted thermodynamics of the unified data set of 131 sequences is provided in the Supporting Information.

Thermodynamic Data. Plots of T_M^{-1} versus ln C_T were linear (correlation coefficient ≥0.98) over the entire 80−100-fold range in concentrations and are shown in Figure 1 and Supporting Information. Thermodynamic parameters derived from fits of individual melting curves and from T_M^{-1} versus ln C_T plots are listed in Table 2. Sequences in which the ΔH° from the two methods agree within 10% are listed as two-state transitions. Seven of 45 of the sequences show agreement of ΔH° within 10−20% are listed as "marginally non-two-state" transitions. These sequences were also included in the regression analysis since they did not significantly affect the derived nearest-neighbor parameters. Sequences with ΔH° differences greater than 20% are listed as non-two-state transitions and were not included in the regression analysis. For molecules with two-state transitions, thermodynamic data derived from averages of the fits and

from T_M^{-1} vs ln C_T plots are equally reliable (SantaLucia et al., 1996; Sugimoto et al., 1996); therefore, the averages of these parameters were used for the linear regression analysis (Table 3).

Linear Regression Analysis of G·T Mismatches in Terms of 10 Linearly Independent Sequences. Table 4 lists parameters for 10 uniquely determined trimer and tetramer sequences with G·T mismatches obtained by multiple linear regression of the data listed in Table 3 (see Materials and Methods). The errors listed in Table 4 are from a resampling analysis of the data (see Materials and Methods). The resampling errors are within round-off of the errors obtained by propagating experimental errors in the SVD analysis (not shown). The parameters in Table 4 along with the Watson−Crick nearest-neighbor parameters (Table 1) predict the thermodynamics of the 45 sequences with two-state transitions in Table 3 with average deviations of 5.1%, 7.5%, 8.0%, and 1.4 °C for $\Delta G°_{37}$, $\Delta H°$, $\Delta S°$, and T_M, respectively. This level of agreement between experiment and prediction (SantaLucia et al., 1996; Freier et al., 1986) indicates that the nearest-neighbor model is valid for DNA sequences containing G·T mismatches. The 18 self-complementary sequences in Table 3 are predicted as well as the 27 non-self-complementary sequences by the parameters in Table 4, demonstrating the consistency of our approach.

Linear Regression Analysis of G·T Mismatches in Terms of 11 Non-Unique Dimers. An alternative presentation of the 10 linearly-independent parameters for G·T mismatches is to use SVD to calculate the 11 G·T nearest-neighbor dimer duplexes. The stacking matrix still has a rank of 10 and is thus singular, but SVD analysis can still provide a solution that is a least squares fit, but the solution is not unique (Press et al., 1989). The 11 nearest-neighbor dimer parameters (Table 5) make predictions that are equal within roundoff error of those made from the 10 linearly independent trimer and tetramer sequences (Table 4). In fact, the parameters in Table 4 are simple linear combinations of the parameters in Table 5. Conversely, the 10 parameters in Table 4 can not be used to derive the 11 parameters in Table 5 unless an 11th parameter is provided. The SVD analysis in terms of 11 parameters essentially assumes that this additional parameter is zero. It is important to note, however, that trends in the 11 parameters should not be considered physically relevant. The non-uniqueness of the 11 dimers is readily verified by noting that by adding a constant, C, to one of the 11 dimers and then subtracting or adding C or zero to the remaining 10 dimers subject to the constraints imposed by the 10 uniquely determined and linearly independent sequences (Table 4), an alternative 11 parameter solution is obtained that makes equal predictions for oligonucleotide duplexes, but the thermodynamic trends of the dimers themselves are different than those shown in Table 5. It is worth noting that the GG/TT neighbor is uniquely determined since GG/TT = GGGC/CTTG − GGC/CTG. The 11 parameters in Table 5 are useful because they are easier to apply than 10 linearly independent sequences since it is simpler to determine which dimers need to be added (analogous to eq 6) than it is to determine which trimers and tetramers need to be added and subtracted (analogous to eq 10) to predict the thermodynamics of an oligonucleotide duplex.

Molecules with Non-Two-State Thermodynamics. Six sequences are listed in Tables 2 and 3 that melt with non-two-state thermodynamics. We presume that these molecules are able to form structures other than the desired duplex or random coil states, including hairpins and "slipped duplexes". For four of these sequences, non-two-state behavior is manifested in differences greater than 20% for the van't Hoff enthalpies derived from $1/T_M$ vs ln C_T plots and from the fits of individual melting curves (Table 2). Below, we show two sequences that melt with non-two-state (NTS) behavior: NTS-1, $(CGTTGCGTAACG)_2$, and NTS-2, GCGTACGCATGCG/CGCATGTGTACGC (Plum et al., 1995), despite good agreement between $\Delta H°$ values derived by the two different van't Hoff methods (Table 2).

To test the hypothesis that NTS-1 is able to form a stable hairpin we synthesized and melted the mutant hairpin sequence HP-1, CGTTGCATAACG (underlined residues are in the loop) which is unlikely to form a duplex. The mutant sequence melted with a concentration independent T_M of 58 °C (see Supporting Information) and with $\Delta H°$, $\Delta S°$, and $\Delta G°_{37}$ of −31.2 kcal/mol, −94.1 eu, and −1.97 kcal/mol, respectively. We assume that the hairpin form of NTS-1, CGTTGCGTAACG, would have the same thermodynamic properties as HP-1. Figure 2 shows a simulation of the melting of NTS-1 using eqs 2−5. The simulation used parameters measured for HP-1 and the predicted thermodynamics for the duplex to random coil equilibrium (Table 3). The simulation clearly shows significant population of hairpin at temperatures near the duplex T_M. The results of the above simulation can be used to calculate a simulated melting curve using the equation

$$A(T) = b(\epsilon_{RC}[A] + \epsilon_{DH}2[A_2] + \epsilon_{Hairpin}[A_H]) \quad (17)$$

where $A(T)$ is the total temperature dependent absorbance, b is the optical pathlength, and ϵ_{RC}, ϵ_{DH}, and $\epsilon_{Hairpin}$ are extinction coefficient of random coil, double helix, and hairpin, respectively. The calculated melting curve so obtained is in good agreement with the observed melting curve for NTS-1 (not shown).

An alternative explanation for why NTS-1 is not well predicted is that the nearest-neighbor model does not apply in this case. The origin of such an effect would be structural perturbations that propagate more than 1 base pair away from the mismatch. To test this possibility, we synthesized the non-self-complementary duplex CCATGCGTAACG/GGTATGCGTTGC which has the same base pairs next to the mismatches and the same base pair composition further away from the G·T pair as NTS-1, but does not have the potential to form hairpin structures. This duplex is well predicted (Table 3) by the parameters in Table 4. Thus, we reject non-nearest-neighbor effects as an explanation of the thermodynamics observed for NTS-1.

We also investigated NTS-2 to determine why this sequence is not well predicted by the parameters in Table 4. Plum et al. (1995) observed the following thermodynamics for NTS-2: $\Delta H°$(av of fits) = −81.3 kcal/mol, $\Delta H°(1/T_M$ vs ln $C_T)$ = −86.6 kcal/mol, $\Delta H°$(calorimetry) = −89.6, $T_M(5 \times 10^{-6}$ M) = 55.3 °C. We melted this duplex and obtained the parameters shown in Table 2 that are in excellent agreement with those measured by Plum et al. (1995). This agreement by our lab and the Breslauer lab eliminates the possibilities that this sequence is not well predicted due to differences in instrumental calibration or sample preparation. Despite the agreement between the UV melting data and

Table 3: Experimental and Predicted Thermodynamic Parameters of Oligonucleotides with G·T Mismatches[a]

		$-\Delta G^\circ_{37}$ (kcal/mol)[c]		$-\Delta H^\circ$ (kcal/mol)[c]		$-\Delta S^\circ$ (eu)[c]		T_M (°C)[d]	
	ref[b]	expt	pred	expt	pred	expt	pred	expt	pred
Molecules with Two-State Thermodynamics									
CAAAGAAAG GTTTTTTTC	*e*	4.21	3.72	46.7	46.1	137.0	136.6	27.6	24.5
CAAATAAAG GTTTGTTTC	*e*	4.43	4.27	55.6	48.4	165.0	142.2	30.2	28.5
CGTGTCTCC GCACGGAGG		7.77	8.03	52.0	52.4	142.7	142.8	50.0	52.1
CGAGTGTCC GCTCGCAGG		8.47	8.43	62.5	59.5	174.1	164.8	51.6	51.8
GGACTCTCG CCTGGGAGC		7.47	7.61	57.9	50.5	162.7	138.4	46.9	49.1
GGACTGACG CCTGGCTGC		8.37	8.32	62.6	58.5	174.9	162.2	51.0	50.9
GGAGTCACG CCTCGGTGC		8.82	8.03	65.4	52.4	182.4	142.8	52.6	52.1
GACCGTGCAC CTGGTGCGTG		7.18	7.54	53.4	55.2	148.9	153.5	46.0	48.2
GACGTGACCG CTGCGTTGGC		8.08	8.13	63.9	61.1	180.1	170.6	49.1	50.3
GACGTTGGAC CTGCGGCCTG		7.94	7.85	60.4	54.3	169.3	150.0	49.1	49.5
GACGTTAGGC CTGCGGTCCG		7.67	7.73	46.0	51.3	123.5	140.2	51.1	50.5
GCAGGTCTGC		6.47	6.27	38.3	41.3	102.5	111.9	43.6	44.1
GGAGTGCTCC		7.64	7.14	65.7	58.0	187.1	163.6	46.5	45.7
GGCAGTTCGC CCGTTGAGCG		6.87	6.71	54.6	48.6	153.8	134.3	43.9	45.3
CATGAGGCTAC GTACTTCGATG		8.61	8.27	69.9	67.2	197.6	189.6	50.6	50.1
CATGTGACTAC GTACATTGATG		7.72	7.23	64.2	65.6	182.1	188.0	47.2	44.8
CCATCGCTACC GGTAGTGATGG	*f*	10.42	9.71	79.8	71.6	223.8	199.4	56.6	55.7
CCATTGCTACC GGTAATGATGG	*f*	9.06	8.51	75.7	67.3	214.8	189.3	51.5	51.0
GATCATTGTAC GTAGTGACATG		7:81	7.10	69.3	65.9	198.1	189.6	46.9	43.8
GATCTTTGTAC CTAGAGACATG		7.31	6.66	67.6	64.0	194.5	184.9	44.6	41.8
GTAGCGTCATG CATCGTAGTAG		9.80	9.06	76.6	72.0	215.5	202.6	54.7	52.8
GTAGTGACATG CATCATTCTAG		7.88	7.23	68.3	65.6	194.7	188.0	47.3	44.8
CCATGCGTAACG GGTATGCGTTGC		8.98	9.31	71.2	70.0	200.7	195.6	52.1	54.1
CGAGACGTTTCG		6.98	7.53	61.0	65.4	174.1	186.8	43.7	45.7
CGAGCATGTTCG		7.23	8.12	59.8	68.4	169.6	194.2	45.3	48.7
CGCGAATTTGCG	*g*	9.78	9.60	79.3	74.4	224.1	208.8	53.9	54.5
CGTGACGTTACG		8.13	8.37	73.3	68.0	210.0	192.6	47.7	49.3
CGTGTCGATACG		8.39	8.63	73.8	70.0	210.9	198.4	48.8	49.9
CGTTACGTGACG		7.87	8.37	64.9	68.0	184.0	192.6	47.8	49.3
CTCGGATCTGAG		8.43	7.86	75.0	69.2	214.5	198.4	48.8	46.2
CTCTCATGGGAG		6.47	7.06	50.4	55.0	141.5	154.4	41.9	45.3
CTCTGATCGGAG		7.52	7.86	60.3	69.2	170.3	198.4	46.7	46.2
CTGTCATGGCAG		8.30	7.90	59.2	58.8	164.2	163.2	51.4	50.8
CTGTGATCGCAG		8.65	8.70	67.0	73.0	188.2	207.2	51.4	50.6
CTTGGATCTAAG		5.91	5.46	64.0	60.6	187.3	178.2	38.1	35.2
CAACTTGATATTAATA GTTGAATTATAATTAT	*h*	9.40	9.77	91.3	98.5	264.0	286.4	47.1	50.1
CAACTTGATATTAATA GTTGAGCTATAATTAT	*h*	10.10	10.97	92.6	102.8	266.0	296.5	49.4	53.4
CAACTTGATATTAATA GTTGAACTATAGTTAT	*h*	10.50	10.94	95.5	100.1	274.0	287.7	50.5	54.0

Table 3 (Continued)

		$-\Delta G°_{37}$ (kcal/mol)[c]		$-\Delta H°$ (kcal/mol)[c]		$-\Delta S°$ (eu)[c]		T_M (°C)[d]	
	ref[b]	expt	pred	expt	pred	expt	pred	expt	pred
Molecules with Marginally Non-Two-State Thermodynamics									
CGTCTGTCC		8.04	8.32	56.5	58.5	156.5	162.2	50.1	50.9
GCAGGCAGG									
GACTGGAGAG		4.39	4.41	46.0	43.6	134.3	126.2	28.5	28.6
CTGATTTCTC									
GCGATGTCGC		7.92	8.40	55.0	63.4	151.7	177.0	50.2	51.5
GATCTGTGTAC		7.81	7.21	70.6	66.3	202.5	190.7	46.7	44.1
CTAGATACATG									
CGAGTCGATTCG		7.65	7.79	69.6	67.4	199.7	192.6	46.1	46.4
CTTGCATGTAAG		6.14	6.30	56.7	64.4	162.9	187.0	39.5	40.5
CGTGTCTAGATACG		9.22	9.60	78.3	82.2	222.6	234.4	51.7	52.1
Molecules with Non-Two-State Thermodynamics									
GACGTGAGGC		6.44	8.04	35.6	59.5	94.0	166.0	43.8	50.8
CTGCGTTCCG									
CGTTGCGTAACG		7.89	10.33	58.7	73.3	163.1	203.4	49.2	57.5
CTCGCATGTGAG		8.19	8.70	64.6	73.0	181.9	207.2	49.7	50.6
CTGGCATGTCAG		8.54	7.90	67.8	58.8	191.2	163.2	50.7	50.8
CTGGGATCTCAG		7.66	7.06	73.8	55.0	213.4	154.4	45.6	45.3
GCGTACGCATGCG		12.65	15.16	80.9	97.3	220.0	264.4	66.3	71.0
CGCATGTGTACGC									

[a] Listed in alphabetical order and by oligomer length. Experimental values are the averages of the T_M^{-1} versus ln C_T and the curve fit parameters given in Table 2. [b] Sequences without a literature reference are from Table 2 of this work. [c] Standard errors for experimental $\Delta G°_{37}$, $\Delta H°$, and $\Delta S°$ are assumed to be 4%, 8%, and 8%, respectively. [d] Calculated for 10^{-4} M oligomer concentration for self-complementary sequences and 4×10^{-4} M for non-self-complementary sequences. [e] Aboul-ela et al. (1985). [f] M. Arghavani, J. SantaLucia, Jr., and L. Romano, unpublished. [g] Leonard et al. (1990b). [h] Tibanyenda et al. (1984).

Table 4: Thermodynamic Parameters for Ten Linearly Independent Sequences with Internal G·T Mismatches in 1 M NaCl[a]

propagation sequence	$\Delta H°$ (kcal/mol)	$\Delta S°$ (eu)	$\Delta G°_{37}$ (kcal/mol)
AGC/TTG	-3.4 ± 1.6	-11.4 ± 5.0	0.12 ± 0.15
A$\overline{T}$C/T$\overline{G}$G	0.6 ± 1.8	1.6 ± 2.9	0.15 ± 0.14
C$\overline{G}$C/G$\overline{T}$G	-8.6 ± 1.4	-24.3 ± 4.5	-1.05 ± 0.10
C$\overline{T}$C/G$\overline{G}$G	0.4 ± 1.1	2.2 ± 3.3	-0.24 ± 0.13
G$\overline{T}$C/C$\overline{G}$G	-1.2 ± 1.3	-2.4 ± 3.9	-0.51 ± 0.08
T$\overline{G}$C/A$\overline{T}$G	-4.5 ± 1.0	-14.1 ± 3.4	-0.16 ± 0.12
T$\overline{T}$C/A$\overline{G}$G	1.9 ± 1.5	4.9 ± 3.5	0.42 ± 0.15
GG$\overline{G}$C/C$\overline{T}$TG	4.6 ± 2.8	14.1 ± 9.3	0.23 ± 0.05
G$\overline{GT}$C/C$\overline{TG}$G	10.6 ± 2.2	30.1 ± 5.1	1.31 ± 0.10
G$\overline{TG}$C/C$\overline{GT}$G	-9.7 ± 3.8	-29.2 ± 10.7	-0.66 ± 0.23

[a] Errors are resampling standard deviations (see text).

Table 5: Non-Unique Nearerst-Neighbor Thermodynamic Parameters of G·T Mismatches in 1 M NaCl[a]

propagation sequence	$\Delta H°$ (kcal/mol)	$\Delta S°$ (eu)	$\Delta G°_{37}$ (kcal/mol)
A$\underline{G}$/T$\underline{T}$	1.0	0.9	0.71
A$\overline{T}$/T$\overline{G}$	−2.5	−8.3	0.07
C$\overline{G}$/G$\overline{T}$	−4.1	−11.7	−0.47
C$\overline{T}$/G$\overline{G}$	−2.8	−8.0	−0.32
G$\overline{G}$/C$\overline{T}$	3.3	10.4	0.08
G$\overline{G}$/T$\overline{T}$[b]	5.8	16.3	0.74
$\overline{GT}$/$\overline{CG}$	−4.4	−12.3	−0.59
G$\overline{T}$/T$\overline{G}$	4.1	9.5	1.15
$\overline{TG}$/$\overline{AT}$	−0.1	−1.7	0.43
T$\overline{G}$/G$\overline{T}$	−1.4	−6.2	0.52
$\underline{TT}$/$\overline{AG}$	−1.3	−5.3	0.34

[a] These parameters are a linear least-squares fit of the data for a singular matrix with a rank of 10. These parameters make predictions that are within roundoff error of the parameters from Table 4. Linear combinations of the parameters in this table give the parameters in Table 4. Trends in these parameters should not be considered physically relevant (see text). [b] The $\underline{GG}$/$\underline{TT}$ nearest neighbor is uniquely determined (see text).

calorimetric data, Plum et al. were careful not to conclude that this sequence melts with two state thermodynamics. We melted the individual single strands and found both strands were able to form stable concentration dependent structures (Figure 3), probably consisting of partially self-complementary slipped duplexes. Therefore, discrepancy between experiment and prediction in this case is most likely due to the presence of the alternative structures formed by the single strands. The results for NTS-1 as well as those for NTS-2 suggest that caution is in order whenever the two-state approximation is applied.

NMR Spectroscopy of Molecules with Non-Two-State Thermodynamics. The NMR of three non-two-state RNA sequences with G·U mismatches were either broadened (suggesting intermediate conformational exchange) or showed extra resonances suggestive of slow exchange with hairpin or other species (He et al., 1991). The work of He et al. (1991) suggests that 1D-NMR is a good way to assess the validity of the two-state approximation. Figure 4 shows the imino region of the 1D-NMR spectra of NTS-1 and NTS-2. We do not find evidence for extra resonances or that peaks are excessively broad at 10 °C. This suggests that at low temperatures hairpin or slipped duplex species are present in low concentrations or that their resonances are either at the same chemical shifts as the desired duplex or they are too broad to observe due to chemical exchange. However, as the temperature is raised, different resonances begin to broaden at different temperatures and the chemical shifts change with temperature (see Supporting Information), suggesting that the melting processes for the two duplexes are non-two state.

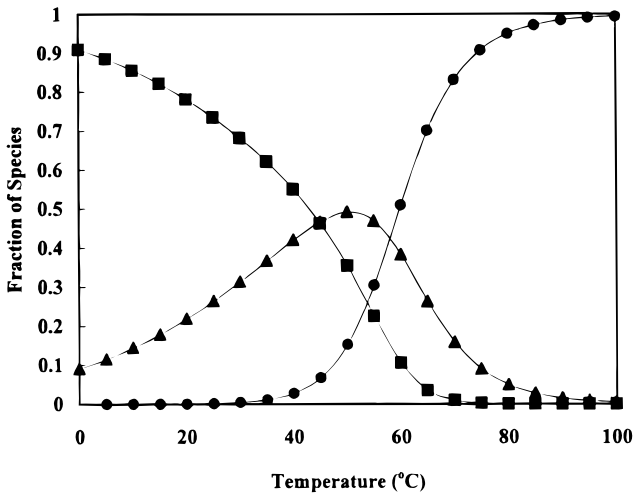

FIGURE 2: Calculated fraction of species formed by CGTTGCGTAACG vs temperature at total strand concentration of 1×10^{-4} M. The species represented are duplex (■), hairpin (▲), and random coil (●).

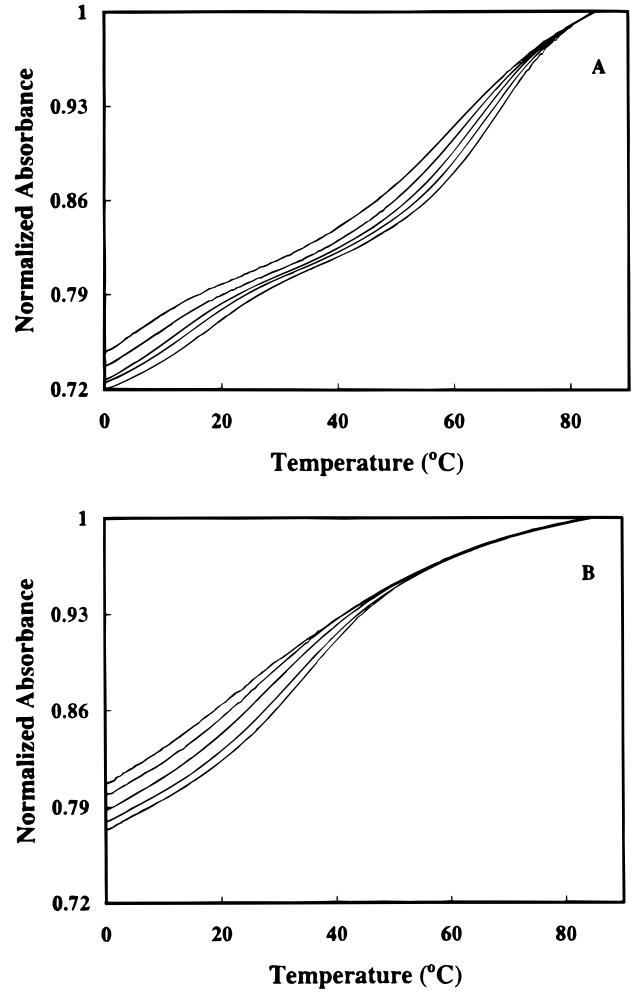

FIGURE 3: Normalized absorbance curves for the single strands of NTS-2 (A) GCGTACGCATGCG, and (B) CGCATGTGTACGC in 1 M NaCl, 10 mM sodium cacodylate, and 0.5 mM Na_2EDTA, pH 7.0. For plot A, the curves shown are at concentrations of 2.1×10^{-4}, 1.2×10^{-4}, 6.5×10^{-5}, 3.5×10^{-5}, and 1.7×10^{-5} M. For plot B, the concentrations are 1.8×10^{-4}, 1.0×10^{-4}, 5.2×10^{-5}, 3.0×10^{-5}, and 1.4×1^{-5} M.

NMR Spectroscopy of Molecules with Two-State Thermodynamics. Figure 5 shows the exchangeable imino region (9−15 ppm) of the 1D proton NMR spectrum of 10

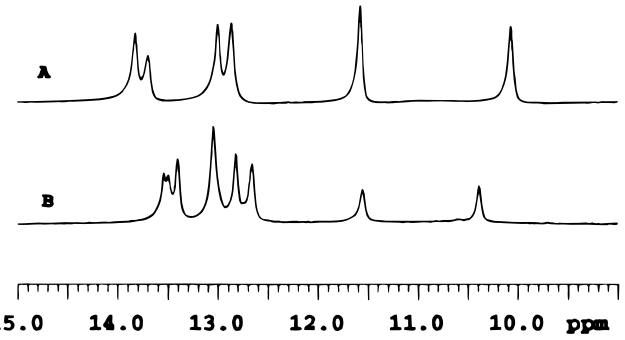

FIGURE 4: 500-MHz ^1H NMR spectra of the exchangeable imino region (9−15 ppm) at 10 °C in 1 M NaCl, 10 mM disodium phosphate, and 0.1 mM Na_2EDTA at pH 7.0 in 90% H_2O/10% D_2O of (A) CGTTGCGTAACG and (B) GCGTACGCATGCG/CGCATGTGTACGC. Assignments are given in Table 6.

sequences. Resonances were assigned using 1D-NOE difference spectroscopy and the temperature dependent broadening of imino protons from terminal base pairs (Figure 6 and Supporting Information). Figure 6 shows the 1D-NOE difference spectra used to assign the sequence $(CGTGACGTTACG)_2$. The assignments for the other sequences listed in Table 6 were determined by the same methods (see Supporting Information). We have assumed that the T imino proton resonance is downfield of the G imino proton of the G·T mismatch (Patel et al., 1984; Hare et al., 1986). In general, the imino protons from the G·T mismatches resonate between 10 and 12 ppm and are as sharp as imino protons in Watson−Crick pairs. G·T mismatches with different surrounding base pairs all show sharp imino resonances consistent with the formation of wobble pairs with stable hydrogen bonding in diverse contexts.

Imino Proton Chemical Shift Predictions. We used chemical shift data in Table 6 to test whether the shielding parameters of Arter and Schmidt (1976) could be extended to apply to sequences with G·T mismatches. We find that the Arter & Schmidt parameters make good predictions for the imino protons of DNA sequences with G·T mismatches if the following assumptions are applied: (1) B-form structure is formed (the use of A-form parameters does not give good chemical shift predictions). (2) G·T mismatches have the same shielding parameters as G·C base pairs. (3) The unperturbed shifts of imino protons in G·C and A·T base pairs are 13.67 and 14.57 ppm, respectively. The unperturbed shifts of G and T imino protons in G·T mismatches are 11.49 and 12.86 ppm, respectively. These numbers are based on a best fit analysis of the data in Table 6. Overall, the Arter & Schmidt shielding parameters with the above modifications predict the chemical shifts of G·C and A·T pairs with average deviations of 0.2 and 0.1 ppm, respectively. For G·T mismatches, the average deviations for G and T chemical shifts are 0.3 and 0.5 ppm, respectively.

DISCUSSION

Applicability of the Nearest-Neighbor Model to G·T Mismatches. If the nearest-neighbor model was not appropriate for G·T mismatches, a single set of energies that predict all sequences could not be found. Table 3 compares the experimental results for 45 G·T mismatch containing oligonucleotides with those predicted using G·T mismatch parameters listed in Table 4 or 5 in conjunction with the Watson−Crick nearest-neighbors (Table 1). The parameters

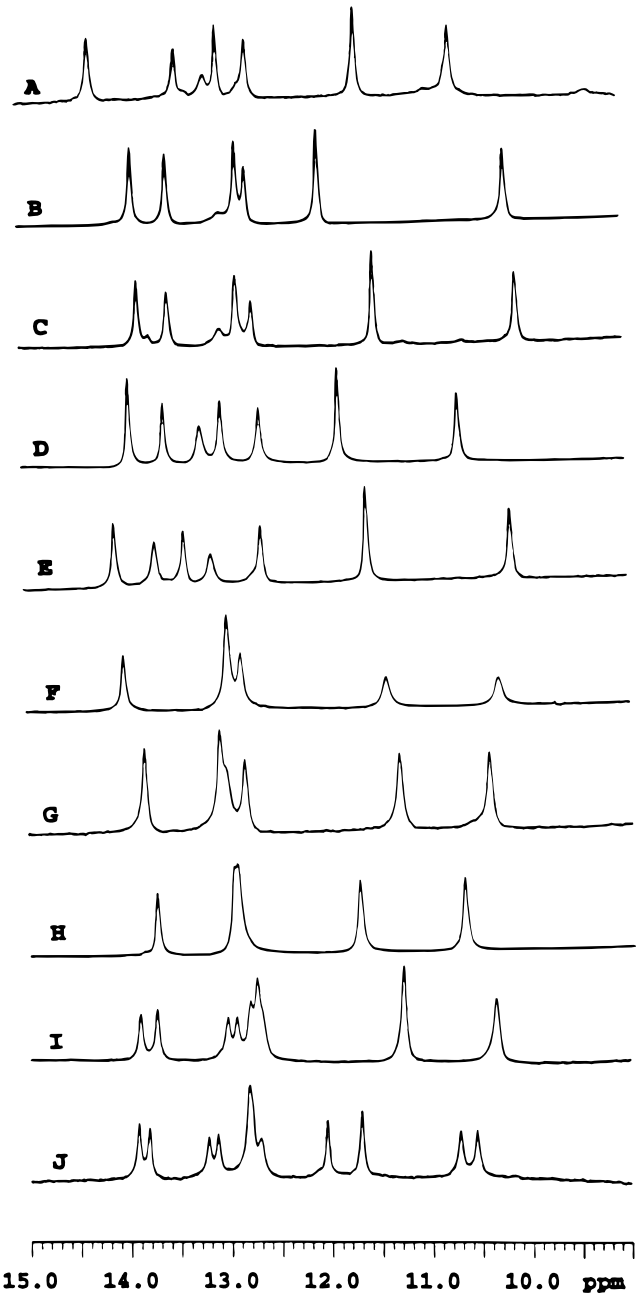

FIGURE 5: 500-MHz ^{1}H NMR spectra of the exchangeable imino region (9−15 ppm) at 10 °C in 1 M NaCl, 10 mM disodium phosphate, and 0.1 mM Na_2EDTA at pH 7.0 in 90% H_2O/10% D_2O of (A) CGAGCATGTTCG, (B) CGTGACGTTACG, (C) CGTGTCGATACG, (D) CTCGGATCTGAG, (E) CTTGCATGTAAG, (F) GCAGGTCTGC, (G) GCGATGTCGC, (H) GGAGTGCTCC, (I) GACCGTGCAC/CTGGTGCGTG, and (J) GACGTTGGAC/CTGCGGCCTG. Assignments are given in Table 6 (see also Supporting Information).

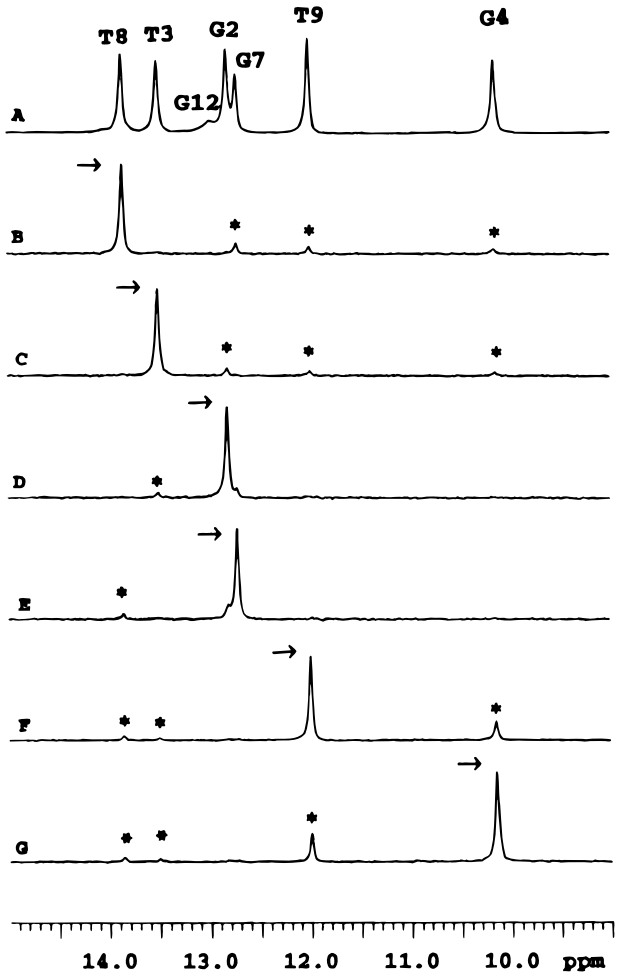

FIGURE 6: Imino proton region (9−15 ppm) of the 1D-NOE difference spectra of CGTGACGTTACG at 10 °C in 1 M NaCl, 10 mM disodium phosphate, and 0.1 mM Na_2EDTA at pH 7.0 in 90% H_2O/10% D_2O. (A) Control spectrum (off-resonance irradiation at 15.0 ppm); (B−G) difference spectra between a control spectrum and spectra obtained with 1 s saturation at 13.9, 13.5, 13.0, 12.8, 12.7, 12.0, and 10.2 ppm, respectively. The saturated resonances are indicated by arrows while the observed NOEs are designated by asterisks. Assignments are shown above spectrum A.

listed in Tables 4 and 5 predict sequences with two-state transitions with average deviations of $\Delta G°_{37} = 5.1\%$, $\Delta H° = 7.5\%$, $\Delta S° = 8.0\%$, and $T_M = 1.4$ °C. Previously, we found that a nearest-neighbor model predicted DNA oligonucleotide thermodynamics with average deviations in $\Delta G°_{37}$, $\Delta H°$, $\Delta S°$, and T_M of 4%, 7%, 8%, and 2 °C, respectively (SantaLucia et al., 1996). This indicates that the nearest-neighbor model applies equally well to oligonucleotides with only Watson−Crick pairs and those with both Watson−Crick and G·T mismatches. The validity of the nearest neighbor model for G·T mismatches is consistent with previous NMR work on G·T mismatches that showed that structural perturbations from the mismatch are mainly in the vicinity of the mismatch and the nearest base pair (Patel et al., 1984; Hare et al., 1986). The agreement between the observed and predicted imino proton shifts (Table 6) also supports the nearest-neighbor model, since the predictions essentially assume a nearest-neighbor model. X-ray crystallographic data for G·T mismatch containing sequences also support the notion that structural perturbations from the mismatch are localized (Hunter et al., 1987).

Another way to test the applicability of the nearest neighbor model for nucleic acid thermodynamics is to synthesize oligonucleotides with different sequences but with the same nearest-neighbor composition (Kierzek et al., 1986; Sugimoto et al., 1994, 1995). Five pairs of sequences listed as melting with two-state thermodynamics in Table 3 have this property. For example, the duplexes CGTCTGTCC/GCAGGCAGG and CGTCGGTCC/GCAGTCAGG have different sequences but the same nearest-neighbor composition and their $\Delta G°_{37}$, $\Delta H°$, $\Delta S°$, and T_M values agree within 0.41 kcal/mol, 6.0 kcal/mol, 18.4 eu, and 0.9 °C, respectively. The five pairs of sequences with the same nearest neighbors have average deviations from the mean for $\Delta G°_{37}$, $\Delta H°$, and

Table 6: Observed and Predicted Chemical Shifts (ppm) of Exchangeable Imino Protons of Oligonucleotides with G·T Mismatches[a]

Sequence[b]								
		G_2	T_3	G_4	G_7	T_8	T_9	G_{12}
$C_1G_2T_3\underline{G_4}A_5C_6G_7T_8\underline{T_9}A_{10}C_{11}G_{12}$	Obs.	12.8	13.5	10.2	12.7	13.9	12.0	13.0
	Pred.	12.8	13.6	9.8	12.7	13.9	11.6	13.2
		G_2	T_3	G_4	T_5	G_7	T_9	G_{12}
$C_1G_2T_3\underline{G_4}T_5C_6G_7A_8\underline{T_9}A_{10}C_{11}G_{12}$	Obs.	12.8	13.5	10.1	13.8	12.7	11.5	13.0
	Pred.	12.8	13.7	10.0	13.8	12.4	11.1	13.2
		G_2	T_3	T_4	G_5	G_7	T_8	G_{12}
$C_1G_2T_3T_4\underline{G_5}C_6G_7\underline{T_8}A_9A_{10}C_{11}G_{12}$	Obs.	12.9	13.8	13.7	10.1	13.0	11.6	(~13.1)
	Pred.	12.8	14.0	13.7	10.1	12.8	11.5	13.2
		T_2	G_4	G_5	T_7	T_9	G_{10}	G_{12}
$C_1T_2C_3\underline{G_4}G_5A_6T_7C_8\underline{T_9}G_{10}A_{11}G_{12}$	Obs.	13.9	10.7	12.7	13.6	11.9	13.0	13.2
	Pred.	14.2	10.5	12.6	13.5	12.0	12.4	13.4
		T_2	T_3	G_4	T_7	G_8	T_9	G_{12}
$C_1T_2T_3\underline{G_4}C_5A_6T_7G_8\underline{T_9}A_{10}A_{11}G_{12}$	Obs.	14.1	13.7	10.2	13.4	12.6	11.6	13.1
	Pred.	14.2	13.7	10.1	13.2	12.3	11.4	13.4
		G_2	G_4	T_7	G_8	T_9	T_{10}	G_{12}
$C_1G_2A_3\underline{G_4}C_5A_6T_7G_8\underline{T_9}T_{10}C_{11}G_{12}$	Obs.	13.0	10.7	13.4	12.7	11.6	14.3	13.1
	Pred.	12.5	10.8	13.2	12.4	12.2	14.1	13.1
		G_2	T_3	G_4	T_7	T_8	G_{10}	G_{12}
$C_1G_2\underline{T_3}G_4A_5A_6T_7T_8C_9\underline{G_{10}}C_{11}G_{12}$[c]	Obs.	(~13.2)	11.7	12.9	13.6	13.7	10.5	13.3
	Pred.	13.0	11.9	12.4	13.5	13.9	10.6	13.2
		G_1	G_4	G_5	T_6	T_8	G_9	
$G_1C_2A_3G_4\underline{G_5}\underline{T_6}C_7T_8G_9C_{10}$	Obs.	(~13.1)	13.0	10.3	11.4	14.0	12.9	
	Pred.	13.3	13.0	11.0	12.2	13.9	12.5	
		G_1	G_3	T_5	G_6	T_7	G_9	
$G_1C_2G_3A_4\underline{T_5}\underline{G_6}T_7C_8G_9C_{10}$	Obs.	13.0	12.8	11.3	10.4	13.8	13.1	
	Pred.	12.4	12.7	11.5	10.5	13.9	12.9	
		G_1	G_2	G_4	T_5	G_6	T_8	
$G_1G_2A_3G_4\underline{T_5}\underline{G_6}C_7T_8C_9C_{10}$	Obs.	(~13.0)	12.9	13.0	11.7	10.7	13.7	
	Pred.	13.2	12.7	13.1	11.9	10.7	14.1	
		G_1	G_5	T_6	G_7	G_{11}	T_{12}	G_{13}	G_{15}	T_{16}	G_{17}	G_{18}	T_{19}
$G_1\ A_2\ C_3\ C_4\ \underline{G_5}\ \underline{T_6}\ G_7\ C_8\ A_9\ C_{10}$	Obs.	(~12.9)	10.4	11.3	13.0	(~12.9)	13.7	12.8	10.4	11.3	12.9	12.8	13.9
$C_{20}\ T_{19}G_{18}\ G_{17}\underline{T_{16}}\ \underline{G_{15}}C_{14}\ G_{13}\ T_{12}G_{11}$	Pred.	12.9	10.7	12.0	12.9	13.2	13.8	12.4	10.8	12.0	12.8	12.9	14.0
		G_1	G_4	T_5	T_6	G_7	G_8	G_{11}	T_{12}	G_{15}	G_{16}	G_{18}	T_{19}
$G_1\ A_2\ C_3\ G_4\ \underline{T_5}\ \underline{T_6}\ G_7\ G_8\ A_9\ C_{10}$	Obs.	(~13.0)	13.2	12.1	11.7	13.1	12.7	(~13.0)	13.9	10.6	10.7	12.8	13.8
$C_{20}T_{19}G_{18}C_{17}\underline{G_{16}}\underline{G_{15}}C_{14}C_{13}T_{12}G_{11}$	Pred.	12.9	12.9	12.2	12.0	12.7	12.7	13.2	14.0	10.7	11.0	12.7	14.0
		G_1	G_3	T_4	G_7	T_{10}	G_{11}	G_{13}	G_{15}	T_{18}	G_{19}	T_{20}	G_{21}	T_{22}	G_{25}
$G_1\ C_2\ G_3\ T_4\ A_5\ C_6\ \underline{G_7}\ C_8\ A_9\ T_{10}G_{11}C_{12}G_{13}$	Obs.	13.1	12.8	13.4	10.4	13.4	12.7	13.1	13.1	13.5	12.7	11.6	12.8	13.5	13.1
$C_{26}G_{25}C_{24}A_{23}T_{22}G_{21}\underline{T_{20}}G_{19}T_{18}A_{17}C_{16}G_{15}C_{14}$	Pred.	13.4	12.6	13.3	10.6	13.2	12.4	13.3	12.9	13.2	12.4	11.9	12.6	13.3	13.0

[a] Observed chemical shifts in 90% H_2O, 10% D_2O, 1 M NaCl, 10 mM disodium phosphate, 0.1 mM Na_2EDTA, pH 7.0, at 10 °C. The predicted shifts using the parameters of Arter and Schmidt (1976) and assuming that the unperturbed chemical shift of imino protons in G−C and A−T pairs are 13.7 and 14.6 ppm, respectively. For G·T mismatches, the unperturbed chemical shift of the imino protons for G and T were assumed to be 11.5 and 12.9 ppm, respectively. In addition, parameters for shielding by G·T mismatches were assumed to be the same as those for G−C base pairs (see text). Chemical shift assignments given in parentheses are tentative. [b] For self-complementary sequences only the top strand is shown. For non-self-complementary sequences the two strands are listed in antiparallel orientation. [c] Observed chemical shifts in 80% H_2O, 20% D_2O, 0.1 M phsophate, 2.5 mM EDTA, pH 6.37, at −5 °C (Patel et al., 1984).

$\Delta S°$ of 3%, 7%, and 8%, and an average difference in T_M of 1.2 °C. These deviations are similar to those observed in RNA (Kierzek et al., 1986) and DNA (Sugimoto et al., 1994). The largest differences are observed for the duplexes CGTGTCTCC/GCACGGAGG and GGAGTCACG/CCTCGGTGC which show deviations from the mean for $\Delta G°_{37}$, $\Delta H°$, and $\Delta S°$ of 6%, 11%, and 12%, and a T_M difference of 2.6 °C. These data indicate the limits of what can be expected from a nearest-neighbor model. Thus, our mismatch parameters in Tables 4 and 5 make predictions within the limits of the nearest-neighbor model.

Context Dependence of G·T Mismatch Thermodynamics. The data in Tables 4 or 5 can be used to predict the thermodynamics of G·T mismatches in all 16 different nearest-neighbor contexts. The most stable context is CGC/GTG which contributes −1.05 kcal/mol to duplex free energy at 37 °C. The least stable context is AGA/TTT which contributes +1.05 kcal/mol. The general trend for the nucleotide at the 5′ side of the G of a G·T mismatch in order of decreasing stability is C > G > T ≥ A. On the 3′ side of the G, the stability order is: C ≥ G > T ≥ A. Interestingly, these trends are reflected in the parameters in

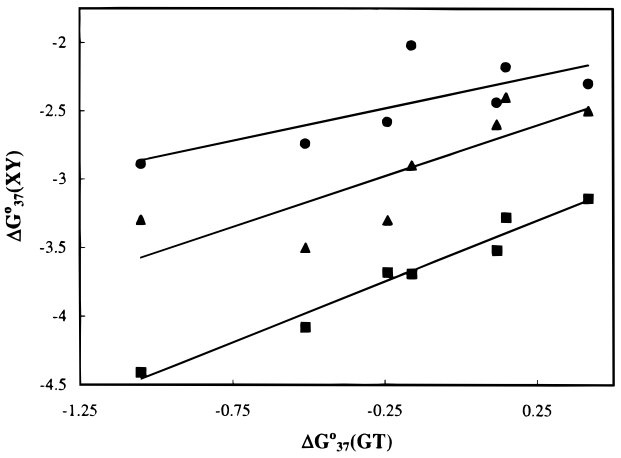

FIGURE 7: Free energy comparison of seven linearly independent single G·T mismatch sequences (Table 4) vs the equivalent sequences with G·T mismatches replaced by G−C base pairs (■), A−T base pairs (●), and RNA G·U mismatches (▲). The lines shown are the least-squares fit of the data ($R^2 = 0.97$ for G−C, 0.56 for A−T, and 0.68 for G·U) (see text for equations).

Table 5, despite the fact that the parameters in Table 5 should not, strictly speaking, be interpreted physically (see Results).

Effects of Terminal and Penultimate G·T Mismatches. Preliminary data from our lab indicate that terminal G·T mismatches contribute between −0.4 and −1.0 kcal/mol, depending on the neighboring base pair (Jenkins and SantaLucia, unpublished results). On the other hand, a single G·T mismatch in the interior of a duplex contributes between +1.05 and −1.05 kcal/mol depending on the context. These data suggest that caution is required when predicting the thermodynamics of duplexes that have mismatches at the penultimate position, particularly when the terminal base pair is A−T. Consider the following self-complementary duplex structures:

```
                                    A           T
AGTGAGCTCATT        vs               GTGAGCTCAT
TTACTCGAGTGA                         TACTCGAGTG
                                    T           A
```

Our data predict that the structure on the right, without terminal A−T hydrogen bonding, is approximately 2.5 kcal/mol *more stable* than the structure on the left which has terminal A−T hydrogen bonding. This effect is presumably due to unfavorable steric interactions that occur when a G·T mismatch is placed in the interior of a duplex.

Comparison of G·T Mismatch and Watson−Crick Base Pair Thermodynamics. A comparison between the 10 linearly independent sequences with G·T mismatches with sequences with only Watson−Crick pairs in which the T in each G·T mismatch is replaced by a C (i.e., G·T versus G−C), revealed that there is a relationship that can be drawn (Figure 7). When double-mismatched nearest neighbors (GG/TT, GT/TG, and TG/GT) are excluded, a line with a correlation coefficient of 0.97 can be fit to the equation

$$\Delta G^\circ_{37}(\mathrm{GC}) = 0.89\Delta G^\circ_{37}(\mathrm{GT}) - 3.52 \qquad (18)$$

where ΔG°_{37}(GC) and ΔG°_{37}(GT) are the Watson−Crick and G·T mismatch nearest neighbor free energies, respectively. On average, a G·T mismatch contributes 3.5 kcal/mol less to duplex stability than an equivalent duplex with a G−C base pair. A poorer correlation is observed when comparing G·T with A−T base pairs ($R^2 = 0.56$). One interpretation of this result is that the guanine stacking plays a more significant thermodynamic role in the G·T mismatch than thymine stacking does. The agreement between experimental and predicted NMR chemical shifts provides evidence that stacking in G−C base pairs is similar to G·T mismatches (see Results). Tandem G·T mismatches do not correlate with Watson−Crick base pairs (not shown); this suggests that unique stacking interactions are present in tandem G·T mismatches. Note that the imino proton chemical shifts of G·T mismatches are predicted better for sequences with single mismatches than for sequences with tandem mismatches (Table 6).

Comparison of DNA G·T Mismatch with RNA G·U Mismatch Nearest-Neighbors. He et al. (1991) reported nearest-neighbor analysis of G·U mismatches in RNA. They found that with the exception of sequences containing GGUC, that the nearest-neighbor analysis applied. Interestingly, we find that GGTC in DNA is not exceptional, and the sequence is well predicted by a single set of nearest-neighbor parameters.

Figure 7 shows a plot of free energies of 7 linearly independent trimer sequences with G·T mismatches in DNA vs the equivalent sequence with G·U mismatches in RNA (excluding RNA G·U tandem mismatches) (He et al., 1991). The data in Figure 7 can be fitted to a line ($R^2 = 0.68$) with the following equation:

$$\Delta G^\circ_{37}(\mathrm{GU}) = 0.75\Delta G^\circ_{37}(\mathrm{GT}) - 2.79 \qquad (19)$$

where ΔG°_{37}(GU) and ΔG°_{37}(GT) are the RNA and DNA trimer free energies, respectively. On average, a G·T mismatch contributes 2.7 kcal/mol less to DNA duplex stability than an analogous RNA duplex with a G·U mismatch. Differences observed for DNA versus RNA thermodynamics are most likely due to different stacking interactions observed in B-form versus A-form structures. For comparison, Watson−Crick G−C and A−T pairs in B-form DNA (Table 1) are also less stable than G−C and A-U pairs in A-form RNA (Freier et al., 1986) by 1.02 kcal/mol, on average. Since G·T and G·U form similar hydrogen-bonded wobble pairs it is somewhat surprising how destabilizing the G·T mismatches are in DNA. One possible explanation of why G·U mismatches are stable in RNA is the presence of a water mediated hydrogen bond between the G-2-amino and the U-2′-hydroxyl oxygen, as revealed by X-ray crystallography (Holbrook et al., 1978, 1991; Hingerty et al., 1978). Another possible explanation for the instability of G·T is unfavorable steric interactions involving the thymine methyl group. The relative instability of G·T mismatches in DNA compared with G·U mismatches in RNA suggests that in addition to DNA's superior hydrolytic stability compared with RNA, DNA may also be inherently better suited than RNA for high-fidelity replication.

ACKNOWLEDGMENT

We thank David Hyndman (Advanced Gene Computing Technologies), Christine Chow, and Sandra Shaner for stimulating conversations and Meiko Ogura (Hitachi Chemical Research) for synthesizing oligonucleotides. We thank Jeff McDowell and Douglas H. Turner for providing the program MELTWIN v2.1 for analysis of melting curves.

SUPPORTING INFORMATION AVAILABLE

One table showing experimental versus predicted (using Table 1) thermodynamics of 131 sequences with Watson−Crick base pairs; 10 figures showing $1/T_M$ vs ln C_T plots for the 33 sequences presented in Table 2 which are not shown in Figure 1; 11 figures showing the 1D-NOE difference spectra used for peak assignments listed in Table 6; two figures showing the 1D-NMR spectra at different temperatures for NTS-1 and NTS-2; one figure showing normalized melting curves for HP-1; and two figures showing plots of $\Delta H°$ vs $\Delta S°$ (31 pages). Ordering information is given on any current masthead page.

REFERENCES

Aboul-ela, F., Koh, D., Tinoco, I., Jr., & Martin, F. H. (1985) *Nucleic Acids Res. 13*, 4811−4824.

Albergo, D. D., Marky, L. A., Breslauer, K. J., & Turner, D. H. (1981) *Biochemistry 20*, 1409−1413.

Arter, D. B., & Schmidt, P. G. (1976) *Nucleic Acids Res. 6*, 1437−1447.

Bevington, P. R. (1969) *Data Reduction and Error Analysis for the Physical Sciences*, pp 164−186, 187−203, McGraw-Hill, New York.

Bhattacharyya, A., & Lilley, D. M. J. (1989) *J. Mol. Biol. 209,* 583−597.

Borer, P. N., Dengler, B., Tinoco, I., Jr., & Uhlenbeck, O. C. (1974) *J. Mol. Biol. 86*, 843−853.

Brown, T. (1995) *Aldrichimica Acta 28*, 15−20.

Brown, T., & Brown, D. J. S. (1991) in *Oligonucleotides and Analogues* (Eckstein, F., Ed.), pp 1−24, IRL Press, New York.

Callidine, C. R., & Drew, H. R. (1984) *J. Mol. Biol. 178,* 773−782.

Chou, S.-H., Flynn, P., & Reid, B. (1989) *Biochemistry 28*, 2422−2435.

Doktycz, M. J., Morris, M. D., Dormandy, S. J., Beattie, K. L., & Jacobson, K. B. (1995) *J. Biol. Chem. 270,* 8439−8445.

Efron, B., & Tibshirani, R. (1993) *An Intoduction to the Bootstrap*, Chapman & Hall, London.

Fodor, S. P. A., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P., & Adams, C. L. (1993) *Nature 364*, 555−556.

Freier, S. M. (1993) in *Antisense Research and Applications* (Crooke, S. T., & Lebleu, B., Eds.) pp 67−82, CRC Press, Boca Raton, FL.

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T., & Turner, D. H. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 9373−9377.

Goldstein, R. F., & Benight, A. S. (1992) *Biopolymers 32*, 1679−1693.

Goodman, M. F., Creighton, S., Bloom, L. B., & Petruska, J. (1993) *Crit. Rev. Biochem. Mol. Biol. 28*, 83−126.

Gray, D. M., & Tinoco, I., Jr. (1970) *Biopolymers 9*, 223−244.

Hare, D., Shapiro, L., & Patel, D. J. (1986) *Biochemistry 25*, 7445−7456.

He, L., Kierzek, R., SantaLucia, J., Jr., Walter, A. E., & Turner, D. H. (1991) *Biochemistry 30*, 11124−11132.

Hingerty, B., Brown, R. S., & Jack, A. (1978) *J. Mol. Biol. 124*, 523−534.

Holbrook, S., Sussman, J. L., Warrant, R. W., & Kim, S.-H. (1978) *J. Mol. Biol. 123*, 631−660.

Holbrook, S. R., Cheong, C., Tinoco, I., Jr., & Kim, S.-H. (1991) *Nature 353*, 579−581.

Hunter, C. A. (1993) *J. Mol. Biol. 230*, 1025−1054.

Hunter, W. N., Brown, T,. Kneale, G., Anand, N. N., Rabinovich, D., & Kennard, O. (1987) *J. Biol. Chem. 262*, 9962−9970.

Ikuta, S., Takagi, K., Wallace, R. B., & Itakura, K. (1987) *Nucleic Acids Res. 15*, 797−811.

Johnson, K. A. (1993) *Annu. Rev. Biochem. 62*, 685−713.

Kawase, Y., Iwai, S., Inoue, H., Miura, K., & Ohtsuka, E. (1986) *Nucleic Acids Res. 14*, 7727−7736.

Kierzek, R., Caruthers, M. H., Longfellow, C. E., Swinton, D., Turner, D. H., & Freier, S. M. (1986) *Biochemistry 25*, 7840−7846.

Klump, H. H. (1990) in *Landolt-Bornstein, New Series, VII Biophysics, Vol. 1, Nucleic Acids, Subvol. c, Spectroscopic and Kinetic Data, Physical Data I* (Saenger, W., Ed.) pp 241−256, Springer-Verlag, Berlin.

Kunkel, T. A., Roberts, J. D., & Zakour, R. A. (1987) *Methods Enzymol. 154*, 367−382.

Leonard, G. A., Booth, E. D., & Brown, T. (1990a) *Nucleic Acids Res. 18*, 5617−5623.

Leonard, G. A., Thomson, J., Watson, W. P., & Brown, T. (1990b) *Proc. Natl. Acad. Sci. U.S.A. 87*, 9573−9576.

Lippens, G., Dhalluin, C., & Wieruszeski, J.-M. (1995) *J. Biomol. NMR 5*, 327−331.

Longfellow, C. E., Kierzek, R., & Turner, D. H. (1990) *Biochemistry 29*, 278−285.

Marky, L. A., & Breslauer, K. J. (1987) *Biopolymers 26*, 1601−1620.

McDowell, J. A., & Turner, D. H. (1996) *Biochemistry 35*, 14077−14089.

Mendelman, L. V., Boosalis, M. S., Petruska, J., & Goodman, M. F. (1989) *J. Biol. Chem. 264*, 14415−14423.

Modrich, P., & Lahue, R. (1996) *Annu. Rev. Biochem. 65*, 101−133.

Patel, D. J., Kozlowski, S. A., Ikuta, S., & Itakura, K. (1984) *Fed. Proc. Fed. Am. Soc. Exp. Biol. 43*, 2663−2670.

Petersheim, M., & Turner, D. H. (1983) *Biochemistry 22*, 256−263.

Petruska, J., Goodman, M. F., Boosalis, M. S., Sowers, L. C., Cheong, C., & Tinoco, I., Jr. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 6252−6256.

Piotto, M., Saudek, V., & Sklenar, V. (1992) *J. Biomol. NMR 2*, 661−665.

Plum, G. E., Grollman, A. P., Johnson, F., & Breslauer, K. J. (1995) *Biochemistry 34*, 16148−16160.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989) *Numerical Recipes*, pp 52−64, 498−520, Cambridge University Press, New York.

Raap, J., van der Marel, G. A., van Boom, J. H., Joordens, J. J. M., & Hilbers, C. W. (1985) *Fourth Conversation in Biomolecular Stereodynamics* (Sarma, R. H., Ed.), p 122a, Adenine Press, Guilderland, NY.

Ratmeyer, L., Vinayak, R., Zhong, Y. Y., Zon, G., & Wilson, W. D. (1994) *Biochemistry 33*, 5298−5304.

Richards, E. G. (1975) in *Handbook of Biochemistry and Molecular Biology: Nucleic Acids* (Fasman, G. D., Ed.) 3rd ed., Vol. 1, p 597, CRC Press, Cleveland, OH.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S., Higuchi, R. H., Horn, G. T., Mullis, K. B., & Erlich, H. A. (1988) *Science 239*, 487−494.

SantaLucia, J., Jr., Kierzek, R., & Turner, D. H. (1990) *Biochemistry 29*, 8813−8819.

SantaLucia, J., Jr., Kierzek, R., & Turner, D. H. (1991) *J. Am. Chem. Soc. 113*, 4313−4322.

SantaLucia, J., Jr., Allawi, H. T., & Seneviratne, P. A. (1996) *Biochemistry 35*, 3555−3562.

Steger, G. (1994) *Nucleic Acids Res. 22*, 2760−2768.

Sugimoto, N., Honda, K., & Sasaki, M. (1994) *Nucleosides Nucleotides 13*, 1311−1317.

Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yonegama, M., & Sasaki, M. (1995) *Biochemistry 34*, 11211−11216.

Sugimoto, N., Nakano, S., Yoneyama, M., & Honda, K. (1996) *Nucleic Acids Res. 24*, 4501−4505.

Tibanyenda, N., De Bruin, S. H., Haasnoot, C. A. G., van der Marel, G. A., van Boom, J. H., & Hilbers, C. W. (1984) *Eur. J. Biochem. 139*, 19−27.

Vologodskii, A. V., Amirikyan, B. R., Lyubchenko, Y. L., & Frank-Kamenetskii, M. D. (1984) *J. Biomol. Struct. Dyn. 2*, 131−148.

Wallace, R. B., Shaffer, J., Murphy, R. F., Bonner, J., Hirose, T., & Itakura, K. (1979) *Nucleic Acids Res. 6*, 3543−3557.

Werntges, H., Steger, G., Riesner, D., & Fritz, H.-J. (1986) *Nucleic Acids Res. 14*, 3773−3790.

Wolfram, S. (1992) *MATHEMATICA*, Version 2.1, Wolfram Research, Inc.

Wu, M., McDowell, J. A., & Turner, D. H. (1995) *Biochemistry 34*, 3204−3211.

BI962590C