Structure, Vol. 11, 1485–1498, December, 2003, ©2003 Elsevier Science Ltd. All rights reserved. DOI 10.1016/j.str.2003.11.008

# Nucleobase Pairing in Expanded Watson-Crick-like Genetic Information Systems

**C. Ronald Geyer,[1,2,*] Thomas R. Battersby,[1,3] and Steven A. Benner**
Department of Chemistry
Department of Anatomy and Cell Biology
University of Florida
Gainesville Florida 32611

## Summary

**To guide the design of alternative genetic systems, we measured melting temperatures of DNA duplexes containing matched and mismatched nucleobase pairs from natural and unnatural structures. The pairs were analyzed in terms of structural features, including nucleobase size, number of hydrogen bonds formed, the presence of uncompensated hydrogen bonding functional groups, the nature of the bond joining the nucleobase to the sugar, and nucleobase charge. The results suggest that stability of nucleobase pairs correlates with the number of H-bonds, size complementarity, the presence of uncompensated functional groups, and the presence of charge on a nucleobase. Each of these properties appear to be more significant than the nature of the glycosidic bond and sequence context. The results provide guidelines for constructing stable Watson-Crick like nucleobase pairs with unnatural nucleobases. The experiments also demonstrate that expanded genetic systems can be constructed using size complementary nucleobase pairs that contain three hydrogen bonds.**

## Introduction

The emerging power of synthetic organic chemistry in the 1980s provided the opportunity to attempt the redesign of genetic systems. Many groups are active in this area (Benner et al., 1998; Freier et al., 1997; Groebke et al., 1998; Hirao et al., 2002; Morales et al., 1998; Tae et al., 2001), and the effort has served to show that we understand far less about the rule-specific molecular recognition in DNA than we had thought. For example, modifications of the backbone that were conceived to be subtle have proven to change dramatically nucleobase pairing.

Opportunities for redesign are substantial even when the standard backbone is left unchanged. According to the Watson-Crick structural model for duplex DNA (Watson et al., 1953a, 1953b), nucleobase pairing follows two rules of complementarity: size and hydrogen bonding (H-bonding). Size complementarity pairs larger purines with smaller pyrimidines. H-bond complementarity pairs H-bond donors with H-bond acceptors. The four natural nucleobases (A, T, G, and C) exploit only two pairing possibilities inherent in these rules. Within the constraints of the Watson-Crick rules, however, six canonical pairing schemes exploiting three hydrogen bonds can be conceived using carbon/nitrogen ring systems that are isosteric to natural purines or pyrimidines (Figure 1) (Piccirilli et al., 1990).

Even within the six pairing schemes, the number of possibilities is enormous. Analogous nucleobase pairs can be joined by fewer than three H-bonds by omitting specific H-bonding functionality. This is the case with the natural **A-T** pair. Further, the addition or removal of heteroatoms can change the physical properties of the heterocycles (e.g., their acid-base properties), substituents can be added, and the nature of the nucleobase-sugar linkage can be changed.

Given the large structural diversity possible once strict conformity to the natural Watson-Crick nucleobase structure is abandoned, it is difficult to immediately construct tight rules, such as those provided by Tinoco, SantaLucia, and Turner for RNA and DNA association (Borer et al., 1974; SantaLucia et al., 1996; Xia et al., 1998). Initially descriptors are needed that highlight essential features of the basic molecular recognition rules for modified base pairs in duplex DNA. The language that is most appropriate to describe these features is not known a priori. Clearly, the terms used to relate noncanonical nucleobase components to duplex stability must be increased beyond those used to describe standard RNA and DNA. It is not clear, however, whether composite terms (e.g., "C-glycoside," "uncompensated functionality at the minor groove") will suffice, or whether the predictive language must capture more detailed concepts (electrostatic charge distribution on the nucleobase, for example, or the placement of single water molecules). This language must be developed in the course of experimental work, where a variety of structures are prepared, trends are observed, exceptions are rationalized by hypothesis, and the hypotheses are tested.

Last, the precision of the analysis needed to construct molecular recognition rules is not known a priori. In general, all molecular recognition systems other than nucleic acids are sensitive to context changes in the surrounding structure. The stability of nucleic acid duplexes, however, is remarkably context independent. Nevertheless, context is important at the second level of analysis. Any rules that seek to be applied to noncanonical information systems must capture the essential impact of the structural feature, rather than interpret the contextual "noise."

Unnatural genetic alphabets are now appearing in FDA-approved diagnostics products (Collins et al., 1997), and tools under development exploit expanded genetic alphabets for both research and diagnostics purposes. With these considerations in mind, we set out to establish basic molecular recognition rules for modified base pairs in duplex DNA. We have synthesized a large number of oligonucleotides containing the nucleobases, many

*Correspondence: ron.geyer@usask.ca
[1]These authors contributed equally to this work.
[2]Present address: Department of Biochemistry, University of Saskatchewan, 107 Wiggins Road, Saskatoon, SK, Canada, S7N 5E5.
[3]Present address: Bayer Diagnostics, 725 Potter Street, Berkeley, California 94710.

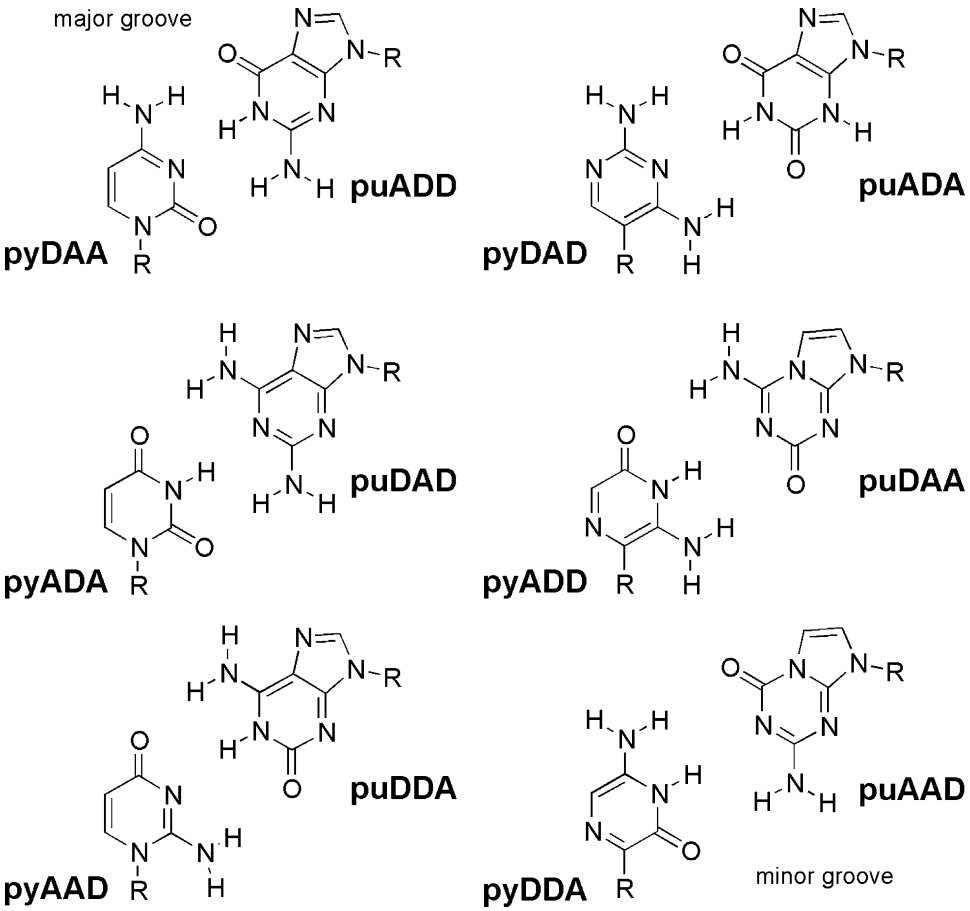


Figure 1. Expanded Watson-Crick-like "In-Line" Genetic System Containing Twelve Nucleobases Joined by Six Mutually Exclusive H-Bonding Patterns

Nucleobase pairs are complementary in both size (purine [**pu**] opposite–pyrimidine [**py**]) and H-bond donors and acceptors. The upper case letters following the **pu** or **py** designation indicate the hydrogen bonding pattern of acceptor (A) and donor (D) groups moving from the major to the minor groove. For example, the standard nucleobase cytosine is pyDAA and guanine is puADD. R indicates 2′-deoxyribose.

unnatural, shown in Figure 2. The nucleobases fit four of the six H-bonding schemes in Figure 1. Theoretically, these can be used to construct an eight nucleobase genetic system.

We used melting temperatures ($T_m$s) from thermal DNA duplex denaturation experiments as a convenient way to estimate the relative stabilities of nucleobase pairs. We then placed the pairs within conformational classes and used these to analyze $T_m$s for trends and construct hypotheses. By comparing $T_m$s of different nucleobase pairs, we identified characteristics that are important for nucleobase pair stability and propose a set of terms that have predictive value for the design of noncanonical genetic systems. The basic molecular recognition rules described here provide the groundwork for later determination of more detailed molecular recognition models similar to those that exist for DNA and RNA.

## Results

### Measuring the Relative Stability of Nucleobase Pairs

The free energy of nucleobase pair formation ($\Delta G^\circ_{T(pair)}$) at a given temperature (T) is used to compare stabilities of individual nucleobase pairs within a nucleic acid duplex. From Turner and coworkers (Kierzek et al., 1999), Equation 1 relates $\Delta G^\circ_{T(pair)}$ to the free energy of formation for a duplex that contains the nucleobase pair of interest ($\Delta G^\circ_{T(duplex)}$):

$$\Delta G^\circ_{T(pair)} = \Delta G^\circ_{T(duplex)} - \Delta G^\circ_{T(duplex\ w/o\ pair)} + \Delta G^\circ_{T(NN)} \quad (1)$$

$\Delta G^\circ_{T(duplex\ w/o\ pair)}$ is the free energy of duplex formation for a duplex without the nucleobase pair of interest. $\Delta G^\circ_{T(NN)}$ is the free energy for the nearest neighbor interaction in the duplex that is lost when the nucleobase pair of interest is inserted into the duplex (Kierzek et al., 1999). $\Delta G^\circ_{T(duplex\ w/o\ pair)}$ and $\Delta G^\circ_{T(NN)}$ remain constant for duplexes that are identical except for the nucleobase pair at a single variable position. Therefore, the differences in the free energies of nucleobase pair formation can be related to the free energies of DNA duplex formation using Equations 2a–2c:

$$\Delta G^\circ_{T(pairA)} = \Delta G^\circ_{T(duplexA)} - \Delta G^\circ_{T(duplex\ w/o\ pair)} + \Delta G^\circ_{T(NN)} \quad (2a)$$

$$\Delta G^\circ_{T(pairB)} = \Delta G^\circ_{T(duplexB)} - \Delta G^\circ_{T(duplex\ w/o\ pair)} + \Delta G^\circ_{T(NN)} \quad (2b)$$

$$\Delta\Delta G^\circ_{T(pairs\ A,B)} = \Delta\Delta G^\circ_{T(duplexes\ A,B)} \quad (2c)$$

**A**

5'--CACN$_1$ACTTTCTCCT-3'

3'-TGTGN$_2$TGAAAGAGG--5'

**B**

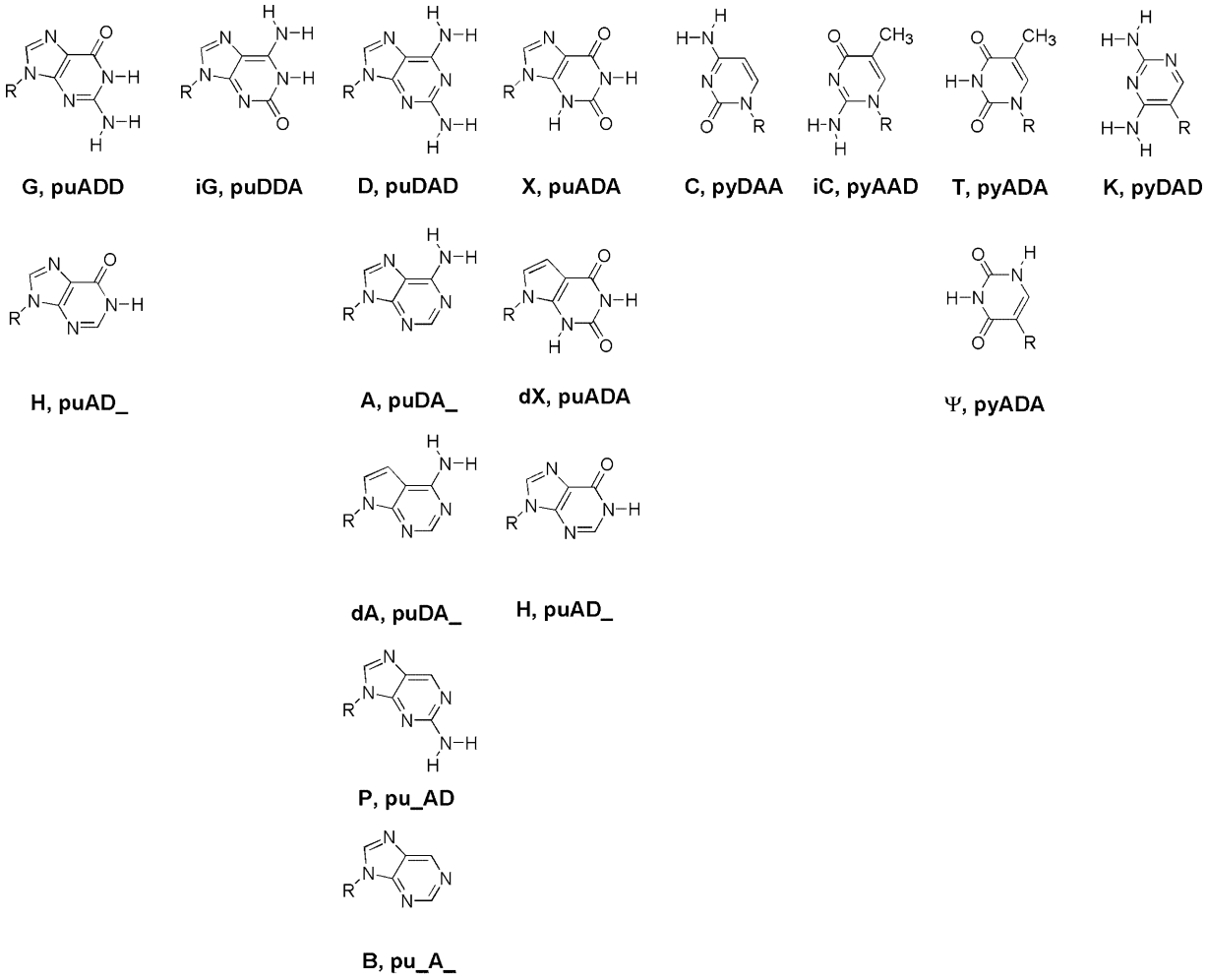


Figure 2. Synthesized Oligonucleotides Containing the Nucleobases

(A) The sequence of the DNA duplex used in the thermal denaturation experiments. $N_1$-$N_2$ represents the position of the variable nucleobase pair.

(B) The nucleobases used in the thermal denaturation experiments. All nucleobases are analogous to the Watson-Crick pairs, with modified carbon/nitrogen ring systems or rearranged or deleted functional groups. The nucleobases within each column have H-bonding functionality consistent with the recognition pattern of one or more of the components in the expanded genetic system in Figure 1. A single letter abbreviation, type of heterocycle, and H-bonding pattern are listed below each nucleobase.

To make feasible the examination of the many nucleobase pairs (matches and mismatches) possible using the nucleobases in Figure 2 (see partial list of these in Table 1), we first observed in previous studies that $T_m$ correlates with $\Delta G^\circ_{T(duplex)}$ when oligonucleotide and salt concentrations are held constant and when the duplex sequence, except for a single variable position, remains unchanged (Kierzek et al., 1999; Peyret et al., 1999). This supports the expedient, widely used in the field, of using $T_m$ values to rank the relative stabilities of nucleobase pairs, especially when large numbers of data are available, and when general trends are desired.

We used the DNA duplex (Figure 2) and thermal denaturation conditions reported by Horn and coworkers (Horn et al., 1995). Using this duplex allowed us to compare our results to Horn's experiments using **iG** and

Table 1. Melting Temperatures (°C) for Nucleobase Pairs

Purine-Pyrimidine Pairs

| In-Line | | Destabilized In-Line | | Wobble | | Reverse Wobble | |
|---|---|---|---|---|---|---|---|
| iG-iC | 63.3 | A-iC* | 50.5 | iG-C | 52.6 | iG-T | 52.3 |
| iC-iG | 61.5 | iC-A* | 48.8 | D-iC | 52.1 | T-iG | 51.7 |
| G-C | 59.5 | H-K | 47.5 | A-iC | 51.5 | D-K | 50.7 |
| C-G | 58.5 | X-K | 47.5 | G-T | 50.8 | G-iC | 49.8 |
| D-T | 58.4 | X-C* | 45.9 | D-K | 50.7 | A-C* | 48.7 |
| T-D | 56.9 | G-K | 45.6 | X-T* | 50.4 | C-P | 48.4 |
| dX-K* | 56.9 | iC-X* | 45.6 | T-X* | 50.2 | P-C | 48.3 |
| D-Ψ | 56.7 | iG-K | 45.4 | C-iG | 49.8 | iC-G | 47.8 |
| X-K* | 54.7 | C-X* | 44.5 | iC-D | 49.7 | C-D | 47.0 |
| dX-K | 52.8 | X-iC* | 43.1 | H-T | 49.7 | P-K | 46.8 |
| A-T | 55.9 | X-C | 41.8 | A-K | 49.5 | D-C | 46.2 |
| dA-T | 54.9 | iC-X | 39.9 | iC-A | 49.3 | B-K | 46.5 |
| P-T | 54.7 | C-X | 39.6 | X-T | 48.3 | C-A* | 44.8 |
| H-C | 54.6 | X-iC | 38.2 | T-X | 47.9 | B-C | 42.3 |
| A-Ψ | 54.4 | | | T-G | 47.9 | A-C | 42.3 |
| T-A | 52.9 | | | B-K | 46.5 | C-A | 41.8 |
| C-H | 52.8 | | | P-iC* | 45.7 | | |
| T-P | 52.1 | | | B-iC | 44.0 | | |
| T-dA | 51.6 | | | P-iC | 44.0 | | |
| B-T | 51.3 | | | | | | |
| T-B | 48.6 | | | | | | |

Pyrimidine-Pyrimidine Pairs

| In-line | | Wobble | |
|---|---|---|---|
| iC-C* | 54.5 | C-iC | 47.3 |
| T-K | 50.3 | iC-C | 47.0 |
| C-iC* | 50.1 | T-T | 46.4 |
| T-C | 45.1 | iC-K | 46.0 |
| C-T | 44.2 | C-C* | 45.6 |
| iC-T | 40.6 | iC-iC* | 45.2 |
| T-iC | 39.1 | C-K | 45.0 |
| | | iC-iC | 40.8 |
| | | C-C | 39.8 |

Purine-Purine Pairs

| Anti-Anti | | Anti-Syn | | Anti-Anti or Anti-Syn | | Unclassified | |
|---|---|---|---|---|---|---|---|
| H-A | 51.2 | A-H | 53.2 | X-D* | 50.5 | iG-iG | 45.8 |
| dA-H | 50.1 | iG-G | 52.7 | D-dX | 49.8 | dX-G | 44.7 |
| H-dA | 50.0 | dA-G* | 52.7 | dX-D | 49.6 | B-D | 44.3 |
| dX-A | 48.5 | G-iG | 52.2 | X-D | 48.1 | dX-dX | 44.2 |
| dA-dX | 48.3 | iG-H | 51.7 | iG-P | 48.1 | A-A | 44.0 |
| dX-dA | 48.2 | A-dX | 51.2 | H-P | 46.5 | X-X* | 43.7 |
| dA-G | 47.3 | H-iG | 50.8 | P-iG | 46.1 | A-iG | 43.5 |
| G-A | 46.9 | G-dA* | 50.6 | A-X | 46.0 | iG-A | 43.4 |
| A-G | 46.7 | A-X* | 50.5 | B-H | 45.3 | B-P | 43.2 |
| G-dA | 46.4 | X-A* | 50.4 | B-G | 44.7 | B-dX | 43.2 |
| P-dX | 45.9 | D-H | 49.5 | dX-B* | 44.4 | B-B | 43.1 |
| X-dA | 45.6 | A-G* | 49.5 | B-iG | 42.7 | X-X | 39.8 |
| X-A | 45.3 | X-dA* | 49.2 | dX-B | 42.5 | | |
| dX-P | 45.1 | iG-X | 48.8 | | | | |
| X-P | 42.2 | H-D | 48.5 | | | | |
| P-X | 40.9 | iG-dX | 48.1 | | | | |
| | | G-A* | 48.1 | | | | |
| | | dX-iG | 47.9 | | | | |
| | | X-iG | 47.5 | | | | |
| | | G-G | 47.4 | | | | |
| | | G-dX | 47.4 | | | | |
| | | X-G | 46.3 | | | | |
| | | H-G | 46.3 | | | | |
| | | D-D | 46.1 | | | | |
| | | A-D | 45.8 | | | | |

Abbreviations correspond to the nucleobases in Figure 2. The first nucleobase is $N_1$ and the second nucleobase is $N_2$ as shown in the DNA duplex in Figure 2. The $T_m$ values for the nucleobase pairs mentioned in the text were measured at pH 7.9 except for those pairs marked with an asterisk, which were measured at pH 5.4.

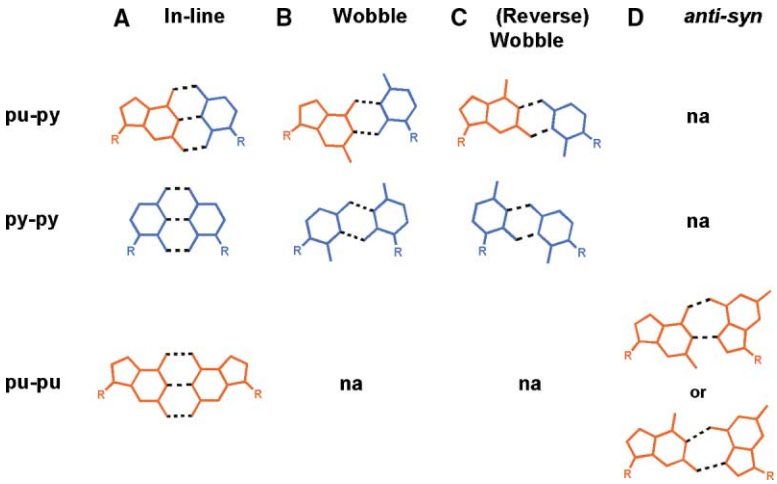


Figure 3. Conformational Classes in Accordance with H-Bonding Interactions of Purine-Pyrimidine (**pu-py**), Pyrimidine-Pyrimidine (**py-py**), and Purine-Purine (**pu-pu**) Nucleobase Pairs

In-line nucleobase pairs (A), defined by the alignment of purine N1 and pyrimidine N3 positions, were considered for all pairs. Conformations one step out of alignment from in-line pairs, wobble (B) and reverse wobble (C) **pu-py** pairs, and wobble **py-py** pairs (B) (C), were also considered. Finally, *anti-syn* **pu-pu** pairs (D) were considered. Additional conformations, including wobble geometries for **pu-pu** pairs and *syn-anti* conformations for **pu-py** pairs (Aishima et al., 2002), are possible but were deemed unlikely in the absence of intercalating agents or protein interactions.

**iC**, two components of our expanded genetic system (Figure 1). The variable position in the DNA duplex, denoted $\mathbf{N_1}$-$\mathbf{N_2}$, is the fourth nucleobase pair from the end of the duplex, so placed to eliminate end effects (Peyret et al. 1999). All DNA duplexes used in our experiments displayed smooth single-melting transitions, consistent with two-state melting behavior. In addition, individual oligonucleotides and DNA duplexes with natural nucleobases at the variable positions are not able to form significant alternative duplex or hairpin structures, minimizing the likelihood of non-two state behavior.

$T_m$s for DNA duplexes containing different nucleobase pairs were calculated at pH 7.9 using the maximum derivative of the $A_{260}$ melting curve (Table 1). In addition, the thermodynamic parameters $\Delta G^\circ_{50}$, $\Delta H^\circ$, and $\Delta S^\circ$ were calculated using a curve fit method (see Supplemental Data online [http://www.structure.org/cgi/content/full/11/12/1485/DC1]) (Marky et al., 1987). $T_m$ and $\Delta G^\circ_{50}$ lead to virtually identical rankings of the relative stabilities of individual nucleobase pairs. We also calculated $T_m$s for DNA duplexes containing different nucleobase pairs at pH 5.4 (see Supplemental Data). Paired nucleobases have perturbed $pK_a$ values (Knitt et al., 1994) and imino nitrogens on some nucleobases will be protonated at pH 5.4 (Boulard et al., 1992, 1997; Brown et al., 1990), while the pairs comprising the remainder of the duplex are relatively unaffected. Nucleobase protonation changes H-bonding functionality and allows nucleobase association that is not observed at pH 7.9.

### Structural Categories of Nucleobase Pairs

To determine characteristics important for forming stable nucleobase pairs, we examined the relationship between $T_m$ and nucleobase pair structure and conformation. We divided nucleobase pairs into structural categories based on the size of the nucleobase heterocycles present in each pair (Figure 3): purine-pyrimidine (**pu-py**) (top row, Figure 3), pyrimidine-pyrimidine (**py-py**) (middle row, Figure 3), and purine-purine (**pu-pu**) (bottom row, Figure 3).

Within each structural category, a systematic conformational analysis was performed for each possible nucleobase pair (Figure 3, columns). The analysis defined the following classes: (a) in-line, where the hexagons of the six-membered rings are aligned, (b) wobble, where one of the hexagons is displaced toward the minor groove, and (c) reverse wobble, where the other hexagon is displaced toward the minor groove (Figure 3). Conformational classes (b) and (c) are, of course, arbitrarily named, but are named for historical reasons. For **pu-pu** wobbling was not considered, but an additional class (d) involving *syn* purine conformation was considered (Figure 3, right column).

For each pair, we then evaluated the number of potential hydrogen bonding interactions for the appropriate class(es). Standard tautomeric forms were considered, with the notable exception of **iG**, where minor tautomeric forms are clearly possible (Sepiol et al., 1976; Roberts et al., 1997). We accepted a conformation for further consideration only if it joins the nucleobases by more attractive H-bonding interactions than repulsive H-bonding interactions. For example, if a conformation for a particular pair of nucleobases contains one attractive and one repulsive H-bonding interaction, then we did not consider this conformation. Then, conformations for individual nucleobase pairs were further characterized by the number of H-bonds and the number of uncompensated H-bonding functionalities or repulsive interactions.

### Analysis of Nucleobase Pairs

Dividing nucleobase pairs into systematic categories allowed us to correlate $T_m$ with nucleobase pair conformation. We asked whether nucleobase pairs assigned to the same conformational class have similar $T_m$s. If the DNA duplex $T_m$s cluster according to class, this suggests (but does not prove) that nucleobase pairs in the same cluster have similar conformations.

### Purine-Pyrimidine Pairs

In most cases, the **pu-py** pairs gave a single conformation with a net positive H-bonding count. Three pairs of **pu-py** pairs examined here, **D-K**, **B-K**, and **X-T** gave two conformations having net positive H-bonding, however. Each of these was considered.

### In-Line pu-py Pairs

As a general rule, the $T_m$ of in-line **pu-py** pairs correlates with the number of H-bonds formed in a nucleobase

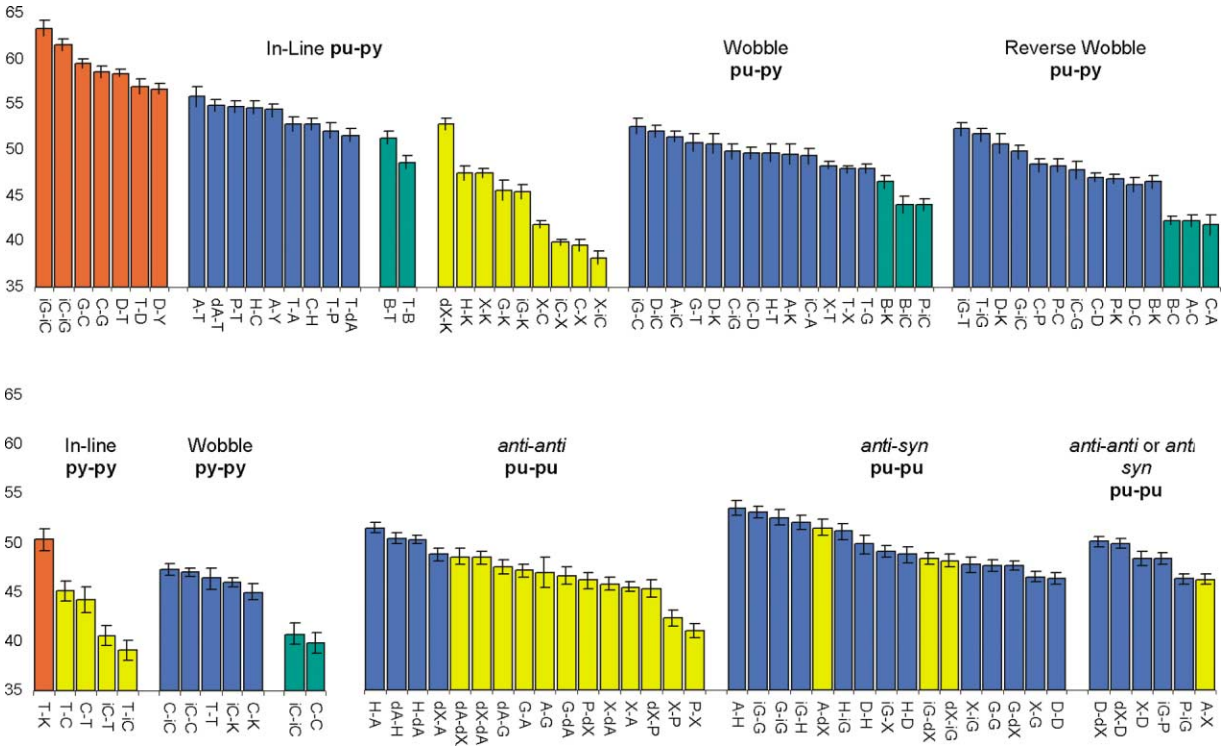


Figure 4. Comparison of $T_m$s (°C) of Nucleobase Pairs from Different Structural Categories

(A) In-line and wobble purine-pyrimidine (**pu-py**) pairs. The nucleobase pairs are grouped according to the number of H-bonds: three (red), two (blue), or one (green). Nucleobase pairs with destabilizing interactions (negative charge, electronic repulsion, or uncompensated amine) are grouped together (yellow). Wobble and reverse wobble **pu-py** pairs are colored according to the number of H-bonds: two (blue), or one (green).

(B) Pyrimidine-pyrimidine (**py-py**) pairs. **py-py** pairs are grouped according to conformation (in-line, wobble, or neither) and colored according to the number of H-bonds: three (red), two (blue), or one (green). **py-py** pairs with a destabilizing interaction are grouped together (yellow).

(C) Purine-purine (**pu-pu**) pairs. **pu-pu** pairs are grouped according to conformation (in-line or *anti-syn*) and colored according to the number of H-bonds: two (blue). **pu-pu** pairs with destabilizing negative charge are colored yellow.

pair (Figure 4). The subset of nucleobase pairs that contain only H-bonds or uncompensated carbonyls, such as in **A-T**, segregate into three $T_m$ ranges. Those formally joined by three H-bonds correspond to a $T_m$ range from 56.7°C to 63.3°C. Those joined by two H-bonds correspond to a $T_m$ range from 52.1°C to 55.9°C. Those joined by one H-bond correspond to a $T_m$ range from 48.6°C to 51.3°C.

Apparent exceptions to this rule are immediately obvious, however (in yellow, Figure 4), and we asked whether these shared one or more structural features that might cause them to violate the trend. Three immediately appeared: (i) the presence of a negative charge on the nucleobase, (ii) the presence of an uncompensated amino group on the nucleobase, and (iii) the presence of a repulsive electronic interaction between the paired nucleobases.

The first feature is represented by pairs that incorporate **X** and **dX**. The riboside analog of **X** has a $pK_a$ of 5.7 as a monomer (Christensen et al., 1970). At pH 7.9, therefore, **X** is expected to exist as an anion, with the negative charge presumably located at O2/N3 (Roy et al., 1983). The $pK_a$ of **dX** has been measured at 7.2 by Seela (Seela et al., 1985) and by us (H. Held, personal communication). Here, the $pK_a$ is close to the pH of the experiment; this requires us to consider that the local environment (in particular, the polyanionic backbone) could cause the predominant form to be protonated.

The second feature is represented by the **H-K** pair (47.5°C). This pair can be compared with the **H-C** pair, which has a $T_m$ = 54.6°C. The **H-K** and **H-C** pairs are particularly noteworthy because they are closely analogous in structure, and the $T_m$ for the **H-C** pair fits the trend with a large number of other pairs.

The third group is represented by **G-K** (45.6°C), and **iG-K** (45.4°C), each with a repulsive amine-amine interaction. These are compared with in-line structures joined by two hydrogen bonds that lack this repulsive interaction, each of which has a $T_m$ in the range of 52.1°C to 55.9°C.

Some structural features of the nucleobase did not evidently have a large impact on the stability of the duplex. For example, the C-glycosyl linkage does not substantially affect the stability of in-line nucleobase pairs. The N-glycoside pairs **D-T** and **A-T** have slightly higher $T_m$s than the C-glycoside pairs **D-Ψ** and **A-Ψ** ($\Delta T_m$s of 1.7 ± 0.9°C and 1.5 ± 1.8°C, respectively), despite the fact that a carbon-carbon bond joining the nucleobase to the deoxyribose is expected to perturb stereoelectronic distribution within the heterocycle (O'Leary et al., 1994). This observation is important for designing expanded genetic systems, as many of the

H-bonding patterns available to these nucleobases must be implemented on a C-glycosidic pyrimidine analog (Piccirilli et al., 1990).

### Testing Structural Hypotheses

Should the presence of a negative charge indeed be the reason why duplexes incorporating **X** and **dX** are less stable than expected based on their H-bond count, then those $T_m$s should rise when the pH is lowered and a proton is added to remove the negative charge. This was in fact observed. The $T_m$s of **X-K** and **dX-K** pairs increase when the pH is lowered to 5.4 ($\Delta T_m$s of 7.2 $\pm$ 1.4°C and 4.1 $\pm$ 1.2°C, respectively). The greater $\Delta T_m$ for the **X-K** pair is consistent with the lower $pK_a$ for **X** than **dX**. The $T_m$s of **X-C**, **C-X**, **X-iC**, and **iC-X** pairs also increase at pH 5.4 ($\Delta T_m$s of 4.1 $\pm$ 1.3°C, 4.9 $\pm$ 1.1°C, 4.9 $\pm$ 1.4°C, and 5.7 $\pm$ 1.0°C, respectively). The increased $T_m$s here are also explained by suppression of the destabilization caused by the negative charge on **X**. Additionally, protonation of **C** or **iC** changes the H-bonding pattern of **X-C** and **X-iC** allowing them to adopt wobble or reverse wobble conformations, respectively (see below).

### Wobble and Reverse Wobble pu-py Pairs

A major consideration in designing artificial genetic systems concerns the potential for mismatches. These can occur via a variety of mechanisms. Prominent among these are "wobble" mismatches. These involve a relative displacement of the nucleobases from the in-line geometry of the standard Watson-Crick pair (Figure 3). By convention, "wobble" and "reverse wobble" involve the displacement of the purine toward the minor and major groove respectively.

In the set of pairs examined here, the $T_m$ ranges for wobble and reverse-wobble pairs having the same H-bond count are very similar (Figure 4). Wobble pairs joined by two H-bonds have a higher $T_m$ range (47.9°C to 52.6°C) than wobble pairs joined by one H-bond (44.0°C to 46.5°C). Likewise, reverse wobble pairs with two H-bonds have a higher $T_m$ range (46.5°C to 52.3°C) than reverse wobble pairs with one H-bond (41.8°C to 46.5°C).

The pH was lowered from 7.9 to 5.4 to examine the relative stability of nucleobase pairs whose H-bonding pattern changes upon protonation. The $T_m$s of the **A-C** and **C-A** pairs significantly increase at pH 5.4 ($\Delta T_m$s of 6.4 $\pm$ 1.3°C and 3.0 $\pm$ 1.3°C, respectively), corresponding to a protonation that increased the hydrogen bond count of these "one H-bond reverse wobble" pairs from one to two, as N1 of **A** is protonated. Protonating **A** allows an H-bond to form between N1 of **A** and O2 of **C,** changing the conformation to a wobble pair joined by two H-bonds (Boulard et al., 1992).

The $T_m$s of the **A-iC** and **iC-A** pairs do not increase upon going to pH 5.4 ($\Delta T_m$s of 1.0 $\pm$ 1.4°C and 0.5 $\pm$ 1.4°C, respectively). At pH 7.9, **A-iC** forms a two H-bond wobble pair. Lowering the pH to 5.4 should protonate **A** and allow it to form an in-line two H-bond pair with **iC**. The $A^+$**-iC** in-line pair, however, has a destabilizing uncompensated amine at the minor groove. The $T_m$ of $A^+$**-iC** pair at pH 5.4 (50.5°C) is lower than other in-line two H-bond pairs ($T_m$ range: 52.1°C–55.9°C) and it is higher than the $T_m$ of a similar in-line nucleobase pair (**H-K**) (47.5°C), which also has an uncompensated amine at the minor groove. The lack of a $T_m$ increase at pH 5.4 indicates that the presence of a destabilizing uncompensated amine at the minor groove disfavors the formation of an in-line $A^+$**-iC** pair.

The clustering of $T_m$s with predicted conformations for **pu-py** pairs implies that there is physical relevance to the conformational analysis employed here. It is unknown, however, if the least stable nucleobase pairs from our analysis can also adopt alternative conformations. For example, unstable in-line pairs with repulsive interactions could potentially adopt other energetically unfavorable conformations such as a wobble or reverse wobble pair joined by one H-bond.

### Pyrimidine-Pyrimidine Pairs

The **py-py** pairs between standard nucleobases (**T-T**, **T-C**, and **C-C**) are not very stable in a DNA duplex when compared to **pu-py** and **pu-pu** pairs (Kierzek et al., 1999). This is perhaps not surprising, considering the fact that these pairs have no in-line conformations with more than one net H-bond.

The expanded set of pyrimidine analogs in this work enables us to examine **py-py** pairing between pyrimidines that can form three hydrogen bonds in an in-line conformation. All **py-py** pairs in our experiments, with the exception of **iC-C**, can adopt a single conformation with net positive H-bonding. These display a broad range of $T_m$s that segregate according to the preferred conformation and the number of H-bonds (Figure 4). The **T-K** in-line pair joined by three H-bonds has the highest $T_m$ at pH 7.9 (50.3°C). In-line pairs with two H-bonds and a repulsive H-bonding acceptor-acceptor interaction have lower $T_m$s ($T_m$ range: 39.1°C–45.1°C). Wobble **py-py** pairs with two H-bonds also have lower $T_m$s ($T_m$ range: 45.0°C–47.3°C). The remaining two **py-py** pairs, **C-C** (39.8°C) and **iC-iC** (40.8°C) cannot form in-line or wobble pairs, and are among the least stable nucleobase pairs examined.

In-line **py-py** pairs with two H-bonds and a repulsive interaction display an interesting positional effect with respect to the repulsive interaction. The **C-T** (44.2°C) and **T-C** (45.1°C) in-line pairs have a repulsive acceptor-acceptor interaction at O2 and higher $T_m$s than the **iC-T** (40.6°C) and **T-iC** (39.1°C) pairs, which have a similar repulsive interaction at O4. The difference in destabilization between repulsive interactions at the 2-position or the 4-position may be explained by the geometry of the in-line **py-py** pair. Strain in the backbone due to the short **py-py** C1′-C1′ distance and the electronic repulsion at the O2 sites of **C-T** can be relieved by rotating the O2 sites slightly apart, while largely maintaining favorable H-bonding interactions. An analogous rotation to relieve the repulsive interaction at the O4 site would require the C1′-C1′ distance to become even shorter and as a result, electronic repulsion cannot be similarly relieved at the O4 sites of **iC-T** pairs (Figure 5).

### Testing Structural Hypotheses

To test the hypothesis that the extreme stability of the **T-K** pair is due to its three H-bonds, we again turned

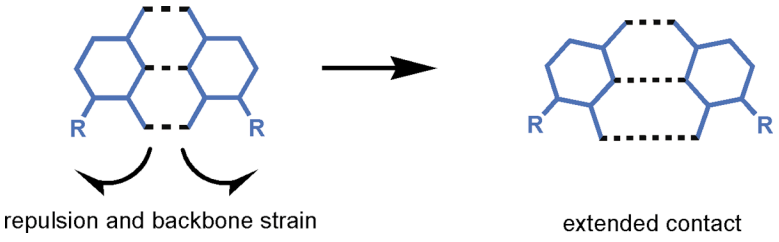


Figure 5. Destabilizing Interactions in **py-py** Pairs

Rotation can simultaneously relieve repulsive acceptor-acceptor interactions at O2 and backbone strain by increasing C1′-C1′ distance (**T-C**). Repulsive acceptor-acceptor interactions at O4 (**T-iC**) are not relieved by analogous nucleobase rotation.

to thermal DNA duplex denaturation experiments at pH 5.4. Either **iC** or **C** is expected to be easily protonated at 5.4, enabling **iC-C** and **C-iC** pairs to be joined by three H-bonds at pH 5.4. The H-bonding pattern of the protonated pair is very similar to that observed in the **C-C**$^+$ pair in parallel-stranded DNA, which has a p$K_a$ of ~6.8 (Leroy et al., 1993). Experimentally, a substantial increase was observed in the corresponding $T_m$s ($\Delta T_m$s of 7.2 ± 1.0°C and 2.8 ± 1.6°C, respectively), consistent with the hypothesis that **iC-C** pair transitions from a two H-bond wobble pair at high pH to a three H-bond in-line pair at low pH. The $T_m$s of **iC-iC** and **C-C** pairs also increase at pH 5.4 ($\Delta T_m$s of 4.4 ± 1.8°C and 5.8 ± 1.7°C, respectively). These $T_m$ increases are consistent with a conformational shift to a two H-bond wobble pair at low pH.

The clustering of $T_m$ with predicted conformations for **py-py** pairs suggests a physical relevance to the conformational analysis for these pairs. The extremely high $T_m$s for the three H-bond in-line **py-py** pairs (including protonated pairs) provides strong evidence that the stability of the **py-py** pairs is enhanced by a third H-bond interaction. The increase in relative stability for wobble **py-py** pairs over other **py-py** pairs classified as neither wobble nor in-line is clear.

There is less evidence to indicate conformations of the least stable **py-py** pairs. Further, **py-py** pairs have been depicted as involving bridging elements at the minor groove, such as water molecules (Boulard et al., 1997) or metal ions (Brown, 1995), to widen the C1′-C1′ distance and to allow minimal backbone distortion. Our results do not rule out interaction with bridging water molecules. However, the relatively high $T_m$s associated with in-line complementary **py-py** pairs are most straightforwardly explained by direct interaction of H-bonding functionality on opposing pyrimidines.

### Purine-Purine Pairs

The analysis of matches between purines is complicated by the presence of *syn* conformations that can present hydrogen bonds to the complementary strand. Pairings in *anti-anti* and *anti-syn* conformations are reported for the **pu-pu** pairs **A-G** (Webster et al., 1990; Prive et al., 1987; Leonard et al., 1990; Carbonnaux et al., 1991; Kan et al., 1983), **A-H** (Leonard et al., 1992; Corfield et al., 1987; Seela et al., 1999), **G-G** (Skelly et al., 1993; Lane et al., 1995) and **H-G** (Oda et al., 1991). The conformational difference between *anti-anti* and *anti-syn* **pu-pu** pairs is shown in Figure 3. When both purines are *anti*, the Watson-Crick recognition face is used, and the C1′-C1′ distance is long (Prive et al., 1987). When *anti-syn* pairing occurs, the Watson-Crick recognition face of the *anti* purine is used, and the Hoogsteen recognition face of the *syn* purine is used. The C1′-C1′ distance of an *anti-syn* pair is near that of an in-line **pu-py** pair (Corfield et al., 1987). The presence of multiple energetically similar conformations (Keepers et al., 1984) makes it more difficult to predict the conformation of **pu-pu** pairs.

We analyzed properties important for forming stable **pu-pu** pairs. **A-H** is our most stable **pu-pu** pair at pH 7.9. *Anti-anti* and *syn-anti* conformations are known for the **A-H** pair (Leonard et al., 1992; Corfield et al., 1987). To determine the conformation of **A-H** in the DNA duplex used here (Figure 2), **A** was replaced with 7-deazaadenine (**dA**). The **dA-H** pair should form the *anti-anti* conformation, as **dA**, in its *syn* conformation, can only form a single H-bond nucleobase pair conformation with **H** (Figure 6). The $T_m$ of the *anti-anti* **dA-H** pair is lower than the $T_m$ for the **A-H** pair ($\Delta T_m$ of 2.9 ± 1.3°C), suggesting that the **A-H** pair adopts a *syn-anti* conformation. The **H-A** and **H-dA** pairs have similar $T_m$s ($\Delta T_m$ of 1.2 ± 0.9°C), suggesting that both of these pairs are in the *anti-anti* conformation.

To confirm the conformation of the **A-dX** pair, **A** was replaced with **dA**. The **A-dX** pair has a higher $T_m$ than the *anti-anti* **dA-dX** pair ($\Delta T_m$ of 2.7 ± 1.5°C), suggesting that the **A-dX** pair adopts the *syn-anti* conformation. In contrast, the **dX-A** and **dX-dA** pairs have similar $T_m$s ($\Delta T_m$ of 0.3 ± 1.4°C), suggesting that both pairs adopt the *anti-anti* conformation. These results suggest an importance of sequence context in determining the conformation of **pu-pu** pairs, as has been reported previously (Webster et al., 1990).

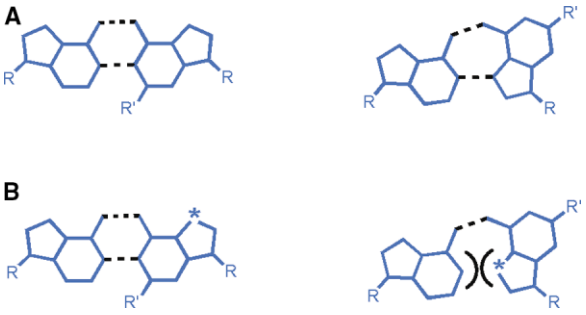


Figure 6. Conformation-Induced Stabilization of **pu-pu** Pairs Containing Potentially Destabilizing Functionalities

(A) Destabilization from unpaired amines or negative charge (represented by R′) in **pu-pu** pairs at the minor groove can be eliminated by adopting an *anti-syn* conformation, which shifts the destabilizing functionality into the major groove.

(B) An analogous *anti-syn* conformation with 7-deaza nucleobases lacking an H-bond acceptor at the 7-position (indicated by asterisk) is not as favorable as in (A).

The **G-A**, **A-G**, and **X-A** pairs at pH 7.9 have $T_m$s that are similar to pairs that have *A* replaced by its 7-deaza analog **dA** ($\Delta T_m$s of 0.5 ± 1.5°C, 0.6 ± 2.2°C, and 0.3 ± 1.1°C, respectively). Without a 7-position nitrogen, **dA** cannot participate in its *syn* conformation in a pair exploiting more than one hydrogen bond. This result, combined with the observation that no *anti-syn* pairs containing more than one H-bond are possible with **dA**, suggests that these pairs adopt the *anti-anti* conformation in our duplex.

Comparing the $T_m$s of the *anti-anti* **pu-pu** pairs described above suggest factors that might determine *anti-anti* **pu-pu** pair stability. The *anti-anti* **pu-pu** pairs with the lowest $T_m$s contain destabilizing functionalities at the minor groove. At the minor groove, pairs with anions (**X-A** pairs; $T_m$ range: 45.3°C–46.0°C) are less stable than pairs with uncompensated amines (**G-A** pairs; $T_m$ range: 46.4°C–47.3°C). Nucleobase pairs with uncompensated amines at the 2-position are less stable than nucleobase pairs lacking functional groups at the 2-positions (**H-A** pairs; $T_m$ range: 50.0°C–51.2°C). This result is similar to the destabilizing effects of uncompensated amines and anions observed in **pu-py** pairs.

### Testing Structural Hypotheses

If *anti-anti* **pu-pu** pair destabilization is caused by the negative charge on **X**, then we expect to restore duplex stability by replacing **X** with **dX**, which has a higher $pK_a$ (Seela et al., 1985). Consistent with this expectation, the $T_m$s of **dX-A** and **dX-dA** are higher than the corresponding nucleobase pairs **X-A** and **X-dA** at pH 7.9 ($\Delta T_m$s of 3.2 ± 1.2°C and 2.6 ± 1.3°C, respectively). Similarly, the $T_m$s of **P-X** and **X-P** are increased at pH 7.9 when **X** is replaced by **dX** ($\Delta T_m$s of 5.0 ± 1.5°C and 2.9 ± 1.6°C, respectively).

Duplexes incorporating purine-purine pairs involving **X** are also stabilized by reduction of pH, consistent with the hypothesis that the negative charge on **X** destabilizes the duplex. At pH 5.4, where **X** is expected to be largely protonated, the $T_m$s for **A-X**, **X-A**, **X-dA**, **X-D**, **P-X**, and **X-P** are higher ($\Delta T_m$s of 4.5 ± 1.3°C, 5.1 ± 1.1°C, 4.4 ± 1.5°C, 2.4 ± 1.4°C, 5.2 ± 1.4°C, and 1.8 ± 1.2°C, respectively). Interestingly, the $T_m$s for **X-G**, **iG-X**, and **X-iG** do not substantially change when the pH is lowered ($\Delta T_m$s of 0.3 ± 1.1°C, 1.0 ± 1.3°C, and 0.8 ± 1.7°C, respectively). The lack of $T_m$ increase at pH 5.4 can be explained by the ability of **G-X** and **iG-X** pairs to adopt a two H-bond *anti-syn* conformation at pH 7.9. In the *syn* conformation, the anion is shifted into the major groove, where it may be more easily solvated (Figure 7). **A-X** and **D-X** pairs on the other hand, cannot adopt an *anti-syn* conformation at pH 7.9 that shifts the anion into the major groove.

Shifting destabilizing functionalities into the major groove of *anti-syn* pairs can also be used to solvate uncompensated amines. For example, **G-A** and **D-H** pairs, in the *anti-anti* conformation, have an uncompensated amine at the minor groove. **D-H** and **H-D** pairs however, have higher $T_m$s than the corresponding **G-A** and **A-G** pairs ($\Delta T_m$s of 2.6 ± 1.6°C and 1.8 ± 2.3°C, respectively). A difference between these two **pu-pu** pairs is that **D-H** pairs can adopt a *syn-anti* conformation with two H-bonds at pH 7.9. **G-A** pairs however, cannot form two H-bonds in the *syn-anti* conformation at pH 7.9. In the *syn-anti* conformation, **D-H** pairs position their uncompensated amines into the major groove. **G-A** pairs can form a two H-bond *anti-syn* pair if the pH is lowered to 5.4 (Carbonnaux et al., 1991; Brown et al., 1990; Sau et al., 1995).

At pH 5.4, **A** can become protonated, allowing the formation of two H-bond *anti-syn* $\mathbf{A^+}$**-G** and **G-**$\mathbf{A^+}$ pairs that have higher $T_m$s ($\Delta T_m$s of 2.8 ± 1.4°C and 1.2 ± 1.3°C, respectively). The *anti-syn* $\mathbf{A^+}$**-G** pair is predicted to be further stabilized through a water-mediated interaction of the 2-amine of **G** with an adjacent phosphate (Leonard et al., 1992).

In general, the conformation adopted by **pu-pu** pairs appears to depend on the number of H-bonds formed and on the presence of destabilizing functionalities (anions and uncompensated amines). When no destabilizing groups are present, **pu-pu** pairs can adopt either the *anti-anti* or *anti-syn* conformation if two or three hydrogen bonds can form. If equivalent destabilizing functionality is present in *anti-anti* and *anti-syn* conformations, the conformation of a **pu-pu** pair can be dictated by sequence context (Webster et al., 1990). When destabilizing groups are present at the H-bonding interface of an in-line (*anti-anti*) conformation, **pu-pu** pairs will adopt an *anti-syn* conformation if two H-bonds can be formed and if the destabilizing functionality (anion or uncompensated amine) can be shifted into the major groove. If a nucleobase pair can form either an *anti-anti* or an *anti-syn* pair joined by two H-bonds, as is the case for **X-A** and **G-A** pairs, but has destabilizing functionality at the H-bonding interface of both the *anti-anti* and *anti-syn* conformations (anion or uncompensated amine), then the *anti-anti* conformation will be adopted.

## Discussion

Exploiting the power of modern synthetic organic chemistry to prepare artificial genetic systems requires guidelines describing how structural features of possible novel nucleobases might influence duplex stability. As noted in the introduction, we are seeking to understand the principal discriminants of stability, while recognizing that the details of duplex stability will vary with sequence context and other subtle features of the structural environment, as it does with standard nucleobases.

The most noteworthy result here is the degree to which simple descriptions of nucleobase structure serve as predictive tools for the range of synthetic nucleobases examined. H-bonding, nucleobase charge, and the presence of uncompensated amino groups have impacts that appear to be consistently large, whether in simple in-line Watson Crick conformations, in wobble size complementary conformations, in **py-py** pairs, and in **pu-pu** pairs, whether in the *anti-anti* or the *anti-syn* conformation. These appear to be more significant than sequence context and uncompensated carbonyl groups, and other unnatural structural features, such as a carbon-carbon glycosidic bond. Within the range of structural and sequence variation examined here, still higher order concepts and their associated language (such as

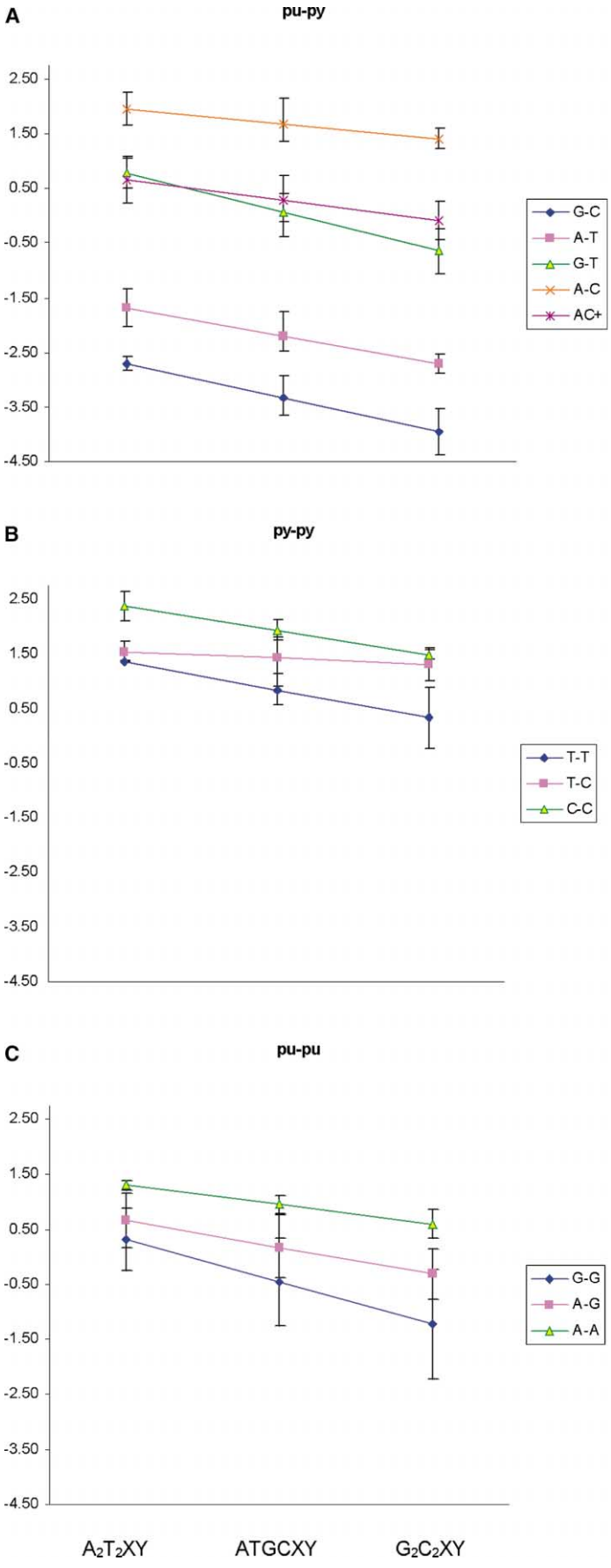


Figure 7. Nearest Neighbor Thermodynamic Parameters

Obtained by SantaLucia and coworkers for natural nucleobase pairs (SantaLucia et al., 1996; Allawi et al., 1997, 1998a, 1998b, 1998c). For internal nucleobase pairs **X-Y**, there are 16 possible trimer sequences when **X** ≠ **Y** and there are ten possible trimer sequences when **X** = **Y**. The trimers are separated into three groups based on the composition of flanking nucleobase pairs. In the **$A_2T_2$XY** group, the **X-Y** pair is flanked on both sides by **A-T** pairs. **In ATGCXY** group, the **X-Y** pair is flanked on one side by an **A-T** pair and on the other side by a **G-C** pair. In the **$G_2C_2$XY** group, the **X-Y** pair is flanked on both sides by **G-C** pairs. Average $\Delta G^\circ_{37}$ values for **X-Y** pairs in each group are represented by symbols. Error bars indicate the range of $\Delta G^\circ_{37}$ values for **X-Y** within each group. The $\Delta G^\circ_{37}$ ranges for **X-Y** pairs (A) **pu-py**, (B) **py-py**, and (C) **pu-pu** are displayed separately.

electrostatic charge distribution) do not appear to be necessary to make basic structural and stability predictions.

Within a structural category, it also appears that nucleobase pairs that have similar $T_m$s also adopt similar conformations. The clustering of $T_m$ according to conformation strongly implies that relative stability and nucleobase pair conformation are primarily determined by hydrogen bonding functionality for **pu-py** and **py-py** pairs.

While context is undoubtedly important, it appears that sequence dependence in our duplexes is a lesser contributor to nucleobase pair stability compared to conformation of **pu-py** and **py-py** pairs. $\Delta G^\circ_{T(pair)}$ in Equation 1 is not identical for a given pair in all sequence contexts. Yet, we observed only a small sequence dependence (average 1.5°C) in nucleobase pair stability by measuring the $T_m$ of many base pairs in two sequence contexts. Additionally, we examined a large variety of interactions between different electrostatic charge distributions (Hunter et al., 1993) by placing many different nucleobase pairs in the same sequence context. Although specific nearest neighbor interactions are different for each nucleobase pair examined, this variability is of insufficient magnitude to obscure the clustering of $T_m$s according to conformation mediated by H-bonding. However, whether the predictable stabilities found for nucleobase pairs in this study hold in all sequence contexts remains to be determined.

The $T_m$ data for **pu-pu** pairs are not easily correlated with conformation, as **pu-pu** pairs can adopt *anti-anti* and *anti-syn* conformations. The existence of multiple **pu-pu** conformations of similar energy confirms that $T_m$ differences of the magnitude arising from sequence context variation will be more likely to influence conformation of **pu-pu** pairs (Skelly et al., 1993).

Our experiments might have been quite difficult to interpret in the presence of tautomers not depicted in Figure 2. Two of our nucleobases, **iC** and **iG,** have been proposed to exist in alternative energetically accessible tautomeric structures. It has been proposed that **iC** has an additional imino-oxo tautomer. There is no evidence, however, of unusual stability and, hence, significant **iC** tautomerism in any of the **pu-py** or **py-py** pairs examined. Likewise, **iG** is proposed to exist in at least three additional tautomers (Roberts et al., 1997; Robinson et al., 1998; Krishnamurthy et al., 1996). Of the **pu-py** pairs examined here, only **iG-T** and **iG-K** can potentially form a three H-bond in-line pair with a hypothesized **iG** tautomer, which should result in an unexpectedly stable pair. **iG-T** can adopt a reverse wobble conformation with **iG** as the N1-H tautomer (as in Figure 2) and an in-line conformation with **iG** as an O2-H tautomer. Both conformations are observed in a DNA duplex crystal structure containing two isolated **iG-T** pairs (Robinson et al., 1998). The $T_m$s for **iG-T** and **T-iG** are somewhat high for pairs classified as wobble or reverse wobble, but are significantly lower than all in-line pairs with three H-bonds. An imino-oxo tautomer of **iG** (Roberts et al., 1997) is complementary to **K**. The low $T_m$ of the **iG-K** duplex, however, suggests that this imino-oxo tautomer is not significantly present. Among the **pu-pu** pairs, **iG-G** pairs have quite high $T_m$s. The N1-H tautomer of **iG** can adopt a favorable *anti-syn* conformation with two H-bonds. However, **iG** may also adopt the N3-H tautomer, which can form an *anti-anti* in-line three H-bond pair with **G.** Because these *anti-anti* and *anti-syn* conformations of **iG-G** are each predicted to be very stable, we cannot use $T_m$s to determine which **iG** tautomer is present in **iG-G** and **G-iG** pairs. In summary, there is no unambiguous evidence of atypical stability due to tautomerism in any of the nucleobase pairs examined.

Previous studies on nucleobase pair thermodynamics in natural DNA support our observations of predictable relative stabilities and conformations for **pu-py** and **py-py** pairs. SantaLucia and coworkers have obtained nearest neighbor thermodynamic parameters for all natural nucleobase pairs (Peyret et al., 1999; SantaLucia et al., 1996; Allawi et al., 1997, 1998a, 1998b, 1998c) (including mismatches) within Watson-Crick DNA duplexes (Figure 7). $\Delta G^\circ_{37}$ values for **pu-py** nucleobase pairs classified as three H-bond in-line (**G-C**), two H-bond in-line (**A-T**), two H-bond wobble (**G-T** and $\mathbf{A^+}$**-C**), and one H-bond reverse wobble (**A-C**) are, in fact, distinctly separated from each other in all sequence contexts when composition of flanking base pairs is taken into account. The A-C pair ($pK_a$ 7.2-7.5 [Boulard et al., 1992, 1995]) was examined at pH 7.0 and $\Delta G^\circ_{37}$ would have probably been even further separated from $\Delta G^\circ_{37}$ values for wobble pairs with two H-bonds if measured, as in our work, at pH 7.9. $\Delta G^\circ_{37}$ values for the two H-bond **py-py** wobble pair **T-T** are always lower than the comparable $\Delta G^\circ_{37}$ values for **C-C**, a **py-py** pair unable to adopt either in-line or wobble conformations. The separation between $\Delta G^\circ_{37}$ for **C-C** ($pK_a$ 7.0 [Boulard et al., 1997]) and **T-T** should also increase at higher pH.

The **T-C** pair, classified in our analysis as an unusual two H-bond in-line pair with an acceptor-acceptor repulsive interaction, has sequence context behavior that is unique among the nucleobase pairs examined by SantaLucia and coworkers (Allawi et al., 1998c) (Figure 7). Except for **T-C**, the stability of all other nucleobase pairs increases when the flanking nucleobase pairs are changed from **A-T** to **G-C** pairs. There is very little difference in the stability of **T-C** pairs when flanked by either **A-T** or **G-C** pairs. This insensitivity to sequence context may be a consequence of the in-line **T-C** pair conformation, which has a repulsive H-bond acceptor-acceptor interaction (Figure 5). Flanking **G-C** pairs more rigidly juxtapose the repulsive acceptor-acceptor interactions in **C-T**, offsetting the duplex stability that could otherwise be gained from flanking **G-C** pairs.

Turner and coworkers have done similar thermodynamic studies (Kierzek et al., 1999) with RNA duplexes at pH 7.0. These experiments yielded results similar to the experiments with DNA duplexes.

While crystallographic and nmr data are sparse with unnatural nucleobases, data confirm the conformational analyses where they are available. The **iC-iG** pair exists as an in-line three H-bond pair in an RNA duplex (Chen et al., 2001). The **iG-T** pair is observed in either the reverse wobble conformation or the in-line conformation pair through tautomerization (Robinson et al., 2001). The wobble conformation of the **iC-D** pair has been inferred through relative thermodynamic properties of association (Strobel et al., 1994).

The stability of complementary nucleobase pairs rela-

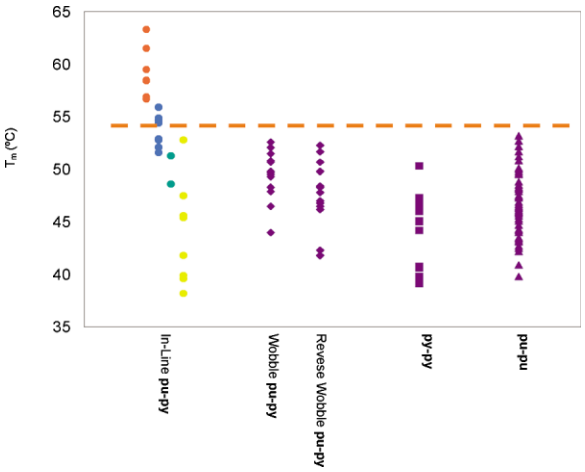


Figure 8. Comparison of Pairs

Comparison of the $T_m$s of complementary in-line purine-pyrimidine (**pu-py**) pairs (circles) to noncomplementary **pu-py** wobble and reverse wobble pairs (diamonds), **py-py** (diamonds), and **pu-pu** (triangles) pairs. The number of H-bonds in **pu-py** in-line pairs is represented by the following colors: three (red), two (blue), and one (green). **pu-py** pairs with destabilizing interactions are colored yellow. The dashed line indicates the high $T_m$ barrier for noncomplementary nucleobase pairs.

tive to noncomplementary pairs is a key thermodynamic requirement for designing an expanded genetic system. Increasing the number of nucleobase pairs in a genetic system requires that all possible noncomplementary pairs, including **pu-py** wobble and reverse wobble pairs and all **py-py** and **pu-pu** pairs, confer less stability to a duplex than the complementary pairs. Discrimination problems as a result of **pu-pu** or **py-py** pairs are not significant in natural DNA duplexes, as all **py-py** and **pu-pu** mismatches possible with standard nucleobases are less stable than complementary **pu-py** pairs.

This may not always be the case for expanded genetic systems containing pairs joined by only two H-bonds, where relatively stable **pu-pu** and **py-py** pairs may be just as stable as complementary nucleobase pairs. Thus, if a new genetic system retains the two H-bond pair between **A** and **T**, it appears that some additional nucleobase pairs will not meet this thermodynamic requirement (Figure 8). From our data, however, nucleobase pairs joined by three H-bonds (replacing, for example, **A** by **D**) yield alternative genetic systems that meet this thermodynamic requirement.

The results from our experiments suggest that a twelve-base Watson-Crick like genetic system employing purines and pyrimidines capable of forming in-line pairs with three H-bonds should have specific and predictable pairing properties. Such expanded genetic systems will have a variety of applications including the development of artificial genetic systems, microarrays and nanostructures with enhanced specificity.

It remains to be seen whether an expanded genetic system will be faithfully replicated using polymerases. Organization of individual nucleobase pair structures by H-bonds is certainly an important factor in polymerases, although polymerases undoubtedly provide additional constraints (Kool, 2002). What is clear from many polymerase studies, however, is that different polymerases interact with unnatural nucleobases differently (Horlacher et al., 1995), even within an evolutionary class. Therefore, it is likely that polymerase behavior reflects more the idiosyncrasies of evolutionary history than underlying features of duplex stability.

## Experimental Procedures

Oligonucleotides (5′-CAC$N_1$ACTTTCTCCT-3′ and 5′-GGAGAAAGT $N_2$GTGT-3′) containing the following nucleobases at the variable positions ($N_1$ or $N_2$) were prepared as follows: Oligonucleotides containing **iG** (Jurczyk et al., 1998), **iC** (Jurczyk et al., 1998), **X** (Jurczyk et al., 2000), and **K** (Piccirilli et al., 1991) were prepared as described. Nucleobases **A**, **T**, **G**, **C**, **P**, **D**, **dX**, **dA**, **B**, **H**, and **Ψ**, were purchased as phosphoramidites from Glen Research (Virginia) and used according to the manufacturer's instructions. Oligonucleotides were synthesized by phosphoramidite chemistry on an Expedite 8900 automated DNA synthesizer and then purified (Jurczyk et al., 2000) by anion exchange HPLC followed by reversed-phase HPLC. The purified oligonucleotides were enzymatically digested and the component nucleosides were analyzed (Jurczyk et al., 1998, 2000) using HPLC to verify composition (oligonucleotides containing **iG**, **iC**, **X**, **K**) and analyzed by MALDI-TOF mass spectrometry (all oligonucleotides).

Thermal DNA duplex denaturation experiments with the above oligonucleotides at pH 7.9 were done in 0.45 M NaCl, 0.045 M sodium citrate, (conditions from Horn et al. [1995]). Each oligonucleotide was present at 1.6 μM. Absorbance was monitored at 260 nm over a range of 10°C–80°C with a change in temperature of 1°C $min^{-1}$ for five heating and cooling cycles on a Varian Cary 300 (Palo Alto, CA). Initial heating and cooling cycles were discarded and the $T_m$ was determined by averaging the temperatures of the maximum derivatives of each cooling and heating cycle. Thermodynamic parameters were estimated using a curve fit method (Marky et al., 1987). ΔGs are reported at 50°C (see Supplemental Data online [http://www.structure.org/cgi/content/full/11/12/1485/DC1]). After measuring the melting curves at pH 7.9, the samples were adjusted to pH 5.4 by adding three parts of 0.45 M NaCl, 0.045 M citric acid to 17 parts of the original sample at pH 7.9. $T_m$ measurements were repeated at pH 5.4.

## Acknowledgments

We gratefully thank Dr. Stefan Lutz for providing K phosphoramidites. NASA Astrobiology Institute, the National Institutes of Health (NIH GM54048), the National Science Foundation (NSF CTS 9871880), the Office of Naval Research (ONR N00014-9610362). C.R.G was supported by a Fellowship from the National Science and Engineering Research Council (Canada).

Received: June 25, 2003
Revised: August 8, 2003
Accepted: August 8, 2003
Published: December 2, 2003

## References

Aishima, J., Gitti, R.K., Noah, J.E., Gan, H.H., Schlick, T., and Wolberger, C. (2002). A Hoogsteen base-pair embedded in undistorted DNA. Nucleic Acids Res. *30*, 5244–5252.

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

Allawi, H.T., and SantaLucia, J., Jr. (1998a). Nearest neighbor thermodynamic parameters for internal G·A mismatches in DNA. Biochemistry *37*, 2170–2179.

Allawi, H.T., and SantaLucia, J., Jr. (1998b). Nearest neighbor thermodynamics of internal A·C mismatches in DNA: sequence dependence and pH effects. Biochemistry *37*, 9435–9444.

Allawi, H.T., and SantaLucia, J., Jr. (1998c). Thermodynamics of internal C·T mismatches in DNA. Nucleic Acids Res. *26*, 2694–2701.

Benner, S.A., Battersby, T.R., Eschgfäller, B., Hutter, D., Kodra, J.T., Lutz, S., Arslan, T., Bäschlin, D.K., Blättler, M., Egli, M., et al. (1998). Redesigning nucleic acids. Pure Appl. Chem. *70*, 263–266.

Boulard, Y., Cognet, J.A.H., Gabarro-Arpa, J., Le Bret, M., Sowers, L.C., and Fazakerly, G.V. (1992). The pH dependent configurations of the C·A mispair in DNA. Nucleic Acids Res. *20*, 1933–1941.

Boulard, Y., Cognet, J.A.H., Gabarro-Arpa, J., Le Bret, M., Carbonnaux, C., and Fazakerly, G.V. (1995). Solution structure of an oncogenic DNA duplex, the K-*ras* gene and the sequence containing a central C·A or A·G mismatch as a function of pH: nuclear magnetic resonance and molecular dynamics studies. J. Mol. Biol. *246*, 194–208.

Boulard, Y., Cognet, J.A.H., and Fazakerly, G.V. (1997). Solution structure as a function of pH of two central mismatches, C·T and C·C, in the 29 to 39 K-*ras* gene sequence, by nuclear magnetic resonance and molecular dynamics. J. Mol. Biol. *268*, 331–347.

Borer, P.N., Dengler, B., Tinoco, I., Jr., and Uhlenbeck, O.C. (1974). Stability of Ribonucleic acid Double-stranded Helices. J. Mol. Biol. *86*, 843–853.

Brown, T. (1995). Mismatches and mutagenic lesions in nucleic acids. Aldrichimica Acta *28*, 15–20.

Brown, T., Leonard, G.A., and Booth, E.D. (1990). Influence of pH on the Conformation and Stability of Mismatch Base-pairs in DNA. J. Mol. Biol. *212*, 437–440.

Carbonnaux, C., van der Marel, G.A., van Boom, J.H., Guschlbauer, W., and Fazakerley, G.V. (1991). Solution structure of an oncogenic DNA duplex containing a G-A mismatch. Biochemistry *30*, 5449–5458.

Chen, X., Kierzek, R., and Turner, D.H. (2001). Stability and structure of RNA duplexes containing isoguanosine and isocytidine. J. Am. Chem. Soc. *123*, 1267–1274.

Christensen, J.J., Rytting, J.H., and Izatt, R.M. (1970). Thermodynamic pK, $\Delta H^\circ$, $\Delta S^\circ$, and $\Delta Cp^\circ$values for proton dissociation from several purines and their nucleosides in aqueous solution. Biochemistry *9*, 4907–4913.

Collins, M.L., Irvine, B., Tyner, D., Fine, E., Zayati, C., Chang, C.-A., Horn, T., Ahle, D., Detmer, J., Shen, L.-P., et al. (1997). A branched DNA signal amplification assay for quantification of nucleic acid targets below 100 molecules/mL. Nucleic Acids Res. *25*, 2979–2984.

Corfield, P.W., Hunter, W.N., Brown, T., Robinson, P., and Kennard, O. (1987). Inosine-adenine base pairs in a B-DNA duplex. Nucleic Acids Res. *15*, 57935–57949.

Freier, S.M., and Altmann, K.-H. (1997). The ups and downs of nucleic acid duplex stability: structure-stability studies on chemically-modified DNA:RNA duplexes. Nucleic Acids Res. *25*, 4429–4443.

Groebke, K., Hunziker, J., Fraser, W., Peng, L., Diederichsen, U., Zimmermann, K., Holzner, A., Leumann, C., and Eschenmoser, A. (1998). Warum pentose- und nicht hexose-nucleinsäuren? (Purin-purin)-basenpaarung in der homo-DNS-reihe: guanin, isoguanin, 2,6-diaminopurin und xanthin. Helv. Chim. Acta *81*, 375–474.

Hirao, I., Ohtsuki, T., Mitsui, T., Yokogawa, T., Okuni, T., Nakayama, H., Takio, K., Yabuki, T., Kigawa, T., Kodama, K., et al. (2002). An unnatural base pair for incorporating amino acid analogs into proteins. Nat. Biotechnol. *20*, 177–182.

Horlacher, J., Hottiger, M., Podust, V.N., Hübscher, U., and Benner, S.A. (1995). Recognition by viral and cellular DNA polymerases of nucleosides bearing bases with nonstandard hydrogen bonding patterns. Proc. Natl. Acad. Sci. USA *92*, 6329–6333.

Horn, T., Chang, C.A., and Collins, M.L. (1995). Hybridization properties of the 5-methyl-isocytidine/isoguanosine base pair in synthetic oligonucleotides. Tetrahedron Lett. *36*, 2033–2036.

Hunter, C.A. (1993). Sequence-dependent DNA structure the role of base stacking interactions. J. Mol. Biol. *230*, 1025–1054.

Jurczyk, S.C., Kodra, J.T., Rozzell, J.D., Benner, S.A., and Battersby, T.R. (1998). Synthesis of oligonucleotides containing 2′-deoxyisoguanosine and 2′-deoxy-5-methylisocytidine using phosphoramidite chemistry. Helv. Chim. Acta *81*, 793–811.

Jurczyk, S.C., Horlacher, J., Devine, K.G., Benner, S.A., and Battersby, T.R. (2000). Synthesis and characterization of oligonucleotides containing 2′-deoxyxanthosine using phosphoramidite chemistry. Helv. Chim. Acta *83*, 1517–1524.

Kan, L.S., Chandrasegaran, S., Pulford, S.M., and Miller, P.S. (1983). Detection of a guanine-adenine base pair in a decadeoxyribonucleotide by proton magnetic resonance spectroscopy. Proc. Natl. Acad. Sci. USA *80*, 4263–4265.

Keepers, J.W., Schmidt, P., James, T.L., and Kollman, P.A. (1984). Molecular-mechanical studies of the mismatched base analogs of d(CGCGAATTCGCG)2:d(CGTGAATTCGCG)2, d(CGAGAATTCGCG)2, d(CGCGAATTCACG)2, d(CGCGAATTCTCG)2, and d(CGCAGAATTCGCG).d(CGCGAATTCGCG). Biopolymers *23*, 2901–2929.

Kierzek, R., Burkard, M.E., and Turner, D.H. (1999). Thermodynamics of single mismatches in RNA duplexes. Biochemistry *38*, 14214–14223.

Knitt, D.S., Narlikar, G.J., and Herschlag, D. (1994). Dissection of the role of the conserved G·U pair in group I RNA self-splicing. Biochemistry *33*, 13864–13879.

Kool, E.T. (2002). Active site tightness and substrate fit in DNA replication. Annu. Rev. Biochem. *71*, 191–219.

Krishnamurthy, R., Pitsch, S., Minton, M., Miculka, C., Windhab, N., and Eschenmoser, A. (1996). Pyranosyl-RNA: base pairing between homochiral oligonucleotide strands of opposite sense of chirality. Angew. Chem. Int. Ed. Engl. *35*, 1537–1541.

Lane, A.N., and Peck, B. (1995). Conformational flexibility in DNA duplexes containing single G-G mismatches. Eur. J. Biochem. *230*, 1073–1087.

Leonard, G.A., Booth, E.D., and Brown, T. (1990). Structural and thermodynamic studies on the adenine-guanine mismatch in B-DNA. Nucleic Acids Res. *18*, 5617–5623.

Leonard, G.A., Booth, E.D., Hunter, W.N., and Brown, T. (1992). The conformational variability of an adenosine-inosine base-pair in a synthetic DNA dodecamer. Nucleic Acids Res. *18*, 4753–4759.

Leroy, J.-L., Gehring, K., Kettani, A., and Guéron, M. (1993). Acid multimers of oligodeoxycytidine strands: stoichiometry, base-pair characterization, and proton exchange properties. Biochemistry *32*, 6019–6031.

Marky, L.A., and Breslauer, K.J. (1987). Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. Biopolymers *26*, 1601–1620.

Morales, J.C., and Kool, E.T. (1998). Efficient replication between non-hydrogen-bonded nucleoside shape analogs. Nat. Struct. Biol. *5*, 950–954.

Oda, Y., Uesugi, S., Ikehara, M., Kawase, Y., and Ohtsuka, E. (1991). NMR studies for identification of dI-dG mismatch base-pairing structure in DNA. Nucleic Acids Res. *19*, 5263–5267.

O'Leary, D.J., and Kishi, Y. (1994). Preferred D conformation of C-glycosides 13. A comparison of the conformational behavior of several C-furanosides, N-furanosides, and O-furanosides. J. Org. Chem. *59*, 6629–6636.

Peyret, N., Seneviratne, P.A., Allawi, H.T., and SantaLucia, J., Jr. (1999). Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A·A, C·C, G·G, and T·T mismatches. Biochemistry *38*, 3468–3477.

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

Piccirilli, J.A., Krauch, T., MacPherson, L.J., and Benner, S.A. (1991). A direct route to 3-(D-ribofuranosyl)pyridine nucleosides. Helv. Chim. Acta *74*, 397–406.

Prive, G.G., Heinemann, U., Chandrasegaran, S., Kan, L.S., Kopka, M.L., and Dickerson, R.E. (1987). Helix geometry, hydration, and G.A mismatch in a B-DNA decamer. Science *238*, 498–504.

Roberts, C., Bandaru, R., and Switzer, C. (1997). Theoretical and experimental study of isoguanine and isocytosine: base pairing in an expanded genetic system. J. Am. Chem. Soc. *119*, 4640–4649.

Robinson, H., Gao, Y.-G., Bauer, C., Roberts, C., Switzer, C., and

Wang, A.H.-J. (1998). 2′-Deoxyisoguanosine adopts more than one tautomer to form base pairs with thymidine observed by high-resolution crystal structure analysis. Biochemistry *37*, 10897–10905.

Roy, K.B., and Miles, H.T. (1983). Tautomerism and ionization of xanthosine. Nucleosides Nucleotides *2*, 231–242.

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

Sau, A.K., Chary, K.V., Govil, G., Chen, C.Q., Howard, F.B., and Miles, H.T. (1995). Evidence for A+(anti)-G(syn) mismatched base-pairing in d-GGTAAGCGTACC. FEBS Lett. *377*, 301–305.

Seela, F., Driler, H., and Liman, U. (1985). 7-deaza-isotere von 2′-desoxyxanthosin und 2′-desoxyspongosin—synthese via glycosylierung von 2,4-dichlor-7*H*-pyrrolo[2,3-*d*]pyrimidin. Liebigs Ann. Chem., 312–320.

Seela, F., and Mittelbach, C. (1999). 7-deaza-2′-deoxyinosine: A stable nucleoside with the ambiguous base pairing properties of 2′-deoxyinosine. Nuclosides Nucleotides *18*, 425–441.

Sepiol, J., Kazimierczuk, Z., and Shugar, D. (1976). Tautomerism of isoguanosine and solvent-induced keto-enol equilibrium. Z. Naturforsch. *31C*, 361–370.

Skelly, J.V., Edwards, K.J., Jenkins, T.C., and Neidle, S. (1993). Crystal structure of an oligonucleotide duplex containing G-G base pairs: Influence of mispairing on DNA backbone conformation. Proc. Natl. Acad. Sci. USA *90*, 804–808.

Strobel, S.A., Cech, T.R., Usman, N., and Beigelman, L. (1994). The 2,6 diaminopurine riboside·5-methylisocytidine wobble base pair: an isoenergetic substitution for the study of G·U pairs in RNA. Biochemistry *33*, 13824–13835.

Sun, Z., Chen, D., Lan, T., and McLaughlin, L.W. (2002). Importance of minor groove functional groups for the stability of DNA duplexes. Biopolymers *65*, 211–217.

Tae, E.L., Wu, Y., Xia, G., Schultz, P.G., and Romesberg, F.E. (2001). Efforts toward expansion of the genetic alphabet: replication of DNA with three base pairs. J. Am. Chem. Soc. *123*, 7439–7440.

Watson, J.D., and Crick, F.H.C. (1953a). Molecular structure of nucleic acids. Nature *171*, 737–738.

Watson, J.D., and Crick, F.H.C. (1953b). General implications of the structure of deoxyribonucleic acid. Nature *171*, 964–967.

Webster, G.D., Sanderson, M.R., Skelly, J.V., Neidle, S., Swann, P.F., Li, B.F., and Tickle, I.J. (1990). Crystal structure and sequence-dependent conformation of the A-G mispaired oligonucleotide d(CGCAAGCTGGCT). Proc. Natl. Acad. Sci. USA *87*, 6693–6697.

Xia, T., SantaLucia, J., Jr., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., and Turner, D.H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry *37*, 14719–14735.