Theoretical Investigations of Novel Nucleic Acid Bases

Andrew R. Leach[§,†] **and Peter A. Kollman***[,†]

Contribution from the Computer Graphics Laboratory and the Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143-0446.
Received September 13, 1991

Abstract: A recent paper (Piccirilli et al. *Nature* **1990**, *343*, 33) reported the design, synthesis, and enzymatic incorporation into DNA and RNA of a new base-pair with a hydrogen bonding pattern different from those used by the GC and AT base-pairs. We describe a variety of calculations on the new bases, both in isolation and in DNA duplexes. Semiempirical and molecular mechanics calculations predict the formation of a stable base-paired structure in the gas phase. Molecular mechanics and energy component analysis of 11 duplexes d(CAAAMAAAG)·d(CTTTNTTTG), where **M** and **N** are chosen from the four naturally occurring and two new bases, provide a partial explanation for the experimentally determined melting temperatures. The model can be qualitatively improved by taking account of the relative free energies of solvation of the bases **M** and **N**. We report relative free energies of solvation of the new bases calculated using the free energy perturbation technique which extend previous calculations (Bash et al. *Science* **1987**, *236*, 564) on the naturally occurring bases. These calculations demonstrate the synergistic relationship between theory and experiment, leading to insights into nucleic acid structure and stability.

Introduction

Molecular recognition lies at the heart of many chemical and biochemical processes and has been the subject of considerable experimental and theoretical investigation in recent years. Base pairing in DNA and RNA is one of the most important examples of molecular recognition in the biological sciences. The hydrogen bond plays a major role in the mechanism by which a base recognizes its complement, yet, as has been noted on many occasions, the naturally occurring bases only use a subset of the possible hydrogen bonding patterns. There has consequently been some speculation into the possibility of alternative patterns. A recent paper[1] described the design and synthesis of some new bases which were designed to form mutually compatible hydrogen bonding patterns with each other and unstable structures with the other naturally occurring bases. They could, therefore (in theory at least), "extend the genetic alphabet". Such systems present interesting challenges for theoretical chemists because they permit existing methodologies to be tested on systems for which they have not been parameterized and may help to further our understanding of the process of molecular recognition. In this paper we describe a number of studies on these molecules in isolated base-pairs and in DNA duplexes using a variety of theoretical methods including semiempirical quantum mechanics, molecular mechanics, molecular dynamics, and free-energy perturbation calculations.

A maximum of three hydrogen bonds can be accommodated per base-pair in a Watson–Crick double helix, giving rise to six mutually exclusive hydrogen bonding schemes (see Figure 1 of Piccirilli et al.), of which the naturally occurring bases use just two (one incompletely as adenine makes only two hydrogen bonds with thymine and uracil). The molecules which are the main focus of this paper were designed κ and π by Piccirilli et al.; κ is a pyrimidine (2,6-diaminopyrimidine) and its complement π is a purine analogue (1-methylpyrazolo[4,3]pyrimidine-5,7-(4*H*,6*H*)-dione, Figure 1. Derivatives of π and κ form standard Watson–Crick base-pairs in chloroform using NMR. In addition, duplexes containing the $\pi\kappa$ base-pair are only slightly less stable than analogous duplexes containing the natural bases. Mismatched duplexes containing π or κ and one of the four natural bases were comparable in stability to their natural mismatched analogues. We also report some calculations on a purine analogue of π named χ by Piccirilli et al., which for synthetic reasons was not pursued further and on two other bases, termed IsoC and IsoG by Piccirilli et al., which are capable of forming a fourth hydrogen bonding pattern (Figure 1).

Energetics of Base Pairing

Initially, semiempirical and molecular mechanics methods were employed to investigate the energetics of base-pair formation. We assume that the major tautomers present for each base are as shown in Figure 1; a recent ab initio study of the base κ[2] determined that the diamino tautomer is more stable than the amino–imino and diimino forms by approximately 75 kJ/mol at the MP2/6-31G**/HF/3-21G+ZPE level. The AM1[3] and PM3[4] Hamiltonians within the MOPAC program[5] were used to determine how the gas-phase interaction energies of the base-pairs $\pi\kappa$, $\chi\kappa$, and IsoC–IsoG compare with those of CG and AT. In each case the energy of base-pair formation was calculated by subtracting the sum of the energies of the individual minimized bases from that of the minimized base-pair. The PRECISE option was used throughout. The results for both AM1 and PM3 are shown in Table I and predict that in the gas phase π and κ would be expected to form a stable base-pair. With the exception of IsoG–IsoC, both PM3 and AM1 predict GC to be more stable than any other base-pair, although, whereas PM3 predicts both $\pi\kappa$ and $\chi\kappa$ to be comparable in stability to AT, AM1 predicts AT to be rather less stable than either $\pi\kappa$ or $\chi\kappa$. We can find no explanation for the seemingly anomalous PM3 result for IsoC–IsoG. Relevant hydrogen bonding distances in the minimized structures of the five base-pairs are shown in Table II. In most cases these are quite reasonable, although the geometries obtained with PM3 are noticeably superior to those obtained with AM1 as we have found for hydrogen bonding in models of the serine protease active site.[6] A further measure of the geometry of the optimized base-pair structures is listed in Table III, which gives the tilt angle between each base-pair, a value of 0° corresponding to exact coplanarity. Some deviation from coplanarity is observed in most cases. However, one rather unusual result is obtained

Table I. Energetics of Base-Pair Formation Using PM3, AM1, and AMBER (Constant Dielectric and Distance-Dependent Dielectric)[a]

base	PM3	AM1	AMBER	AMBER ($\epsilon \propto 1/r$)
GC	−11.7	−13.7	−20.3	−21.6
AT	−5.6	−4.9	−11.2	−13.3
IsoC–IsoG	−3.5	−15.6	−21.2	−21.7
$\pi\kappa$	−5.9	−7.4	−14.5	−17.1
$\chi\kappa$	−5.8	−7.4	−14.8	−17.4

[a] The energy of base-pair formation is calculated in each case by subtracting the energy of the base-pair from the energies of the individual bases.

[§] Computer Graphics Laboratory.
[†] Department of Pharmaceutical Chemistry.

(1) Piccirilli, J. A.; Krauch, T.; Moroney, S. E.; Benner, S. A. *Nature* **1990**, *343*, 33.
(2) Leszczynski, J. *Chem. Phys. Lett.* **1990**, *173*, 371.
(3) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(4) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221.
(5) Stewart, J. J. P. MOPAC Version 5.0: *QCPE Bull.* **1989**, *9*, 98.
(6) Schröder, S.; Daggett, V.; Kollman, P. A. *J. Am. Chem. Soc.* **1991**, *113*, 8922.

0002-7863/92/1514-3675$03.00/0

Table II. Hydrogen Bonding Geometries in the Minimized Base-Pair Structures Using PM3, AM1, and AMBER ($\epsilon \propto 1/r$)[a]

	AM1			PM3			AMB		
	r_{ha}	r_{da}	θ_{dha}	r_{ha}	r_{da}	θ_{dha}	r_{ha}	r_{da}	θ_{dha}
				π–κ					
O6(π)–N2(κ)	2.12	3.10	172.7	1.85	2.86	179.2	1.90	2.90	171.6
N1(π)–N1(κ)	2.05	3.06	173.1	1.77	2.81	178.4	1.89	2.91	179.5
O2(π)–N6(κ)	2.12	3.11	168.5	1.84	2.85	176.7	1.89	2.89	169.0
				χ–κ					
N6(χ)–O2(κ)	2.12	3.10	169.9	1.84	2.85	178.5	1.89	2.89	170.4
N1(χ)–N1(κ)	2.06	3.06	171.1	1.77	2.81	178.4	1.89	2.91	179.1
O2(χ)–N6(κ)	2.12	3.10	167.7	1.84	2.85	176.8	1.89	2.89	168.8
				IsoG–IsoC					
O6(χ)–N2(κ)	2.08	3.06	166.6	1.82	2.83	174.0	1.88	2.82	165.0
N2(χ)–N1(κ)	2.05	3.03	166.2	1.79	2.81	172.4	1.88	2.89	172.9
O2(χ)–N6(κ)	2.06	3.00	155.9	1.79	2.79	170.3	1.88	2.82	168.2
				G–C					
O6(G)–N4(C)	2.08	3.06	166.1	1.80	2.81	178.4	1.87	2.88	176.5
N1(G)–N3(C)	2.32	3.07	131.1	1.79	2.81	173.2	1.90	2.91	175.2
N2(G)–O2(C)	2.11	3.05	155.6	1.83	2.84	175.4	1.87	2.88	173.8
				A–T					
N6(A)–O4(T)	2.10	3.10	179.9	1.82	2.83	175.4	1.89	2.89	168.8
N1(A)–N3(T)	2.46	3.46	170.4	1.78	2.82	175.8	1.89	2.90	175.2

[a] r_{ha} is the distance between the hydrogen and the acceptor atom; r_{da} is the distance between the heavy donor atom and the acceptor; θ_{dha} is the angle subtended at the hydrogen. Distances are in angstroms, angles in degrees.

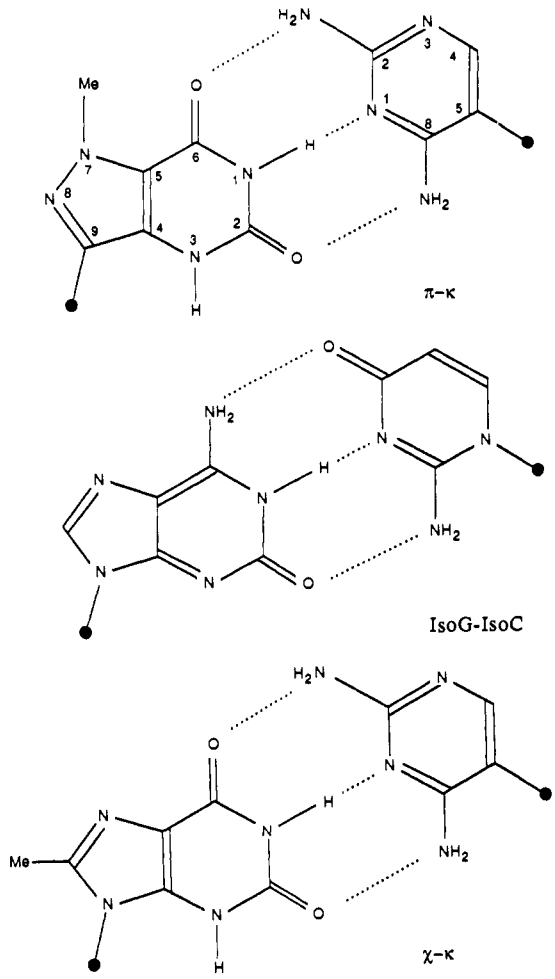

Figure 1. The five novel bases considered in this paper: π, κ, χ, IsoG, and IsoC.

from the AM1 calculation on the GC base-pair in which the optimized structure shows substantial deviation from coplanarity (Figure 2) yet still achieves three hydrogen bonding interactions.

Analogous molecular mechanics calculations were performed using the molecular simulation package AMBER.[7,8] Partial atomic charges for 9-methyl-π, 9-methyl-χ, 9-methyl-IsoG, 5-methyl-κ, and 1-methyl-IsoC were derived from ab initio calculations at the STO-3G level on the PM3-optimized geometries of the individual bases using Gaussian 80 UCSF[9] followed by molecular electrostatic potential fitting[10] and are shown in Figure 3. Bond, angle, and torsion parameters were taken from the AMBER all-atom force field[11,12] or were assigned equilibrium values and force constants based upon the PM3 geometry of the isolated base and analogous AMBER force constants. The energies of base-pair formation (obtained, as before, by subtracting the intramolecular energies of the individual bases from the energy of the base-pair) using both constant and distance dependent dielectric models are shown in Table I, together with previously reported values for the GC and AT base-pairs.[12] Both the $\pi\kappa$ and $\chi\kappa$ base-pairs are predicted to be more stable than the AT base-pair but less stable than the

Table III. Tilt Angles (in degrees) of the Minimized Base-Pair Structures for AM1, PM3, and AMBER

	AM1	PM3	AMBER
A–T	6	2	10
G–C	45	6	8
IsoG–IsoC	23	15	14
χ–κ	20	10	9
π–κ	17	10	8

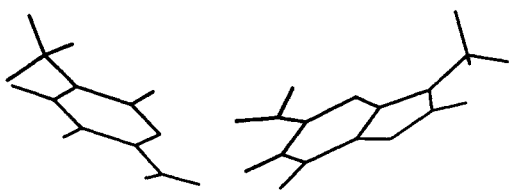

Figure 2. AM1 optimized structure of the GC base-pair.

(7) Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Caldwell, J. C.; Kollman, P. A. *Amber 3.0, Revision A*; University of California: San Francisco, CA, 1989.
(8) Pearlman, D. A.; Case, D. A.; Caldwell, J. C.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *Amber 4.0*; University of California: San Francisco, CA, 1991.
(9) Singh, U. C.; Kollman, P. A. *UCSF* Gaussian-80. *QCPE Bull.* **1982**, *2*, 17, Program 446.
(10) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.
(11) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.
(12) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230.

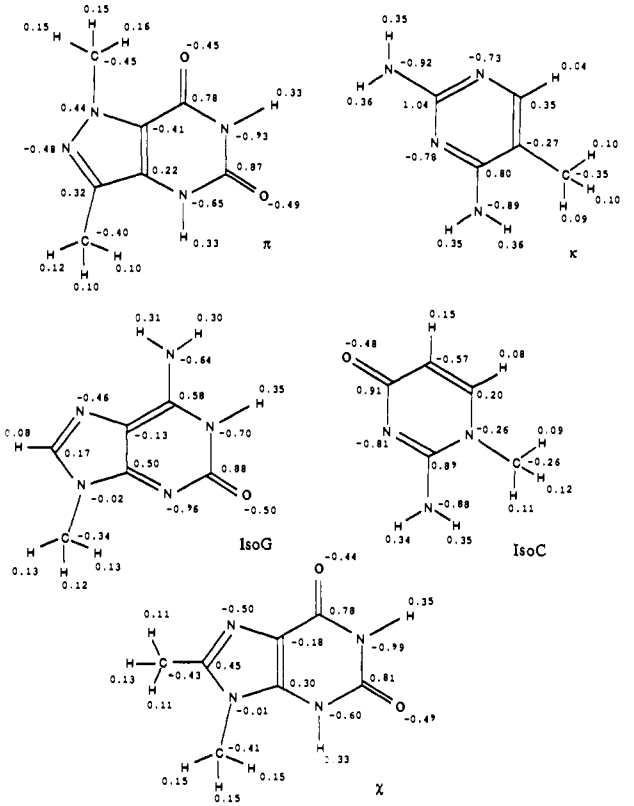

Figure 3. Partial atomic charges for the bases π, κ, χ, IsoG, and IsoC used in the AMBER calculations.

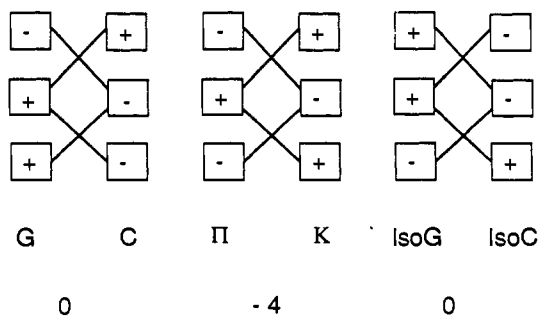

Figure 4. Secondary interaction model of Jorgensen applied to the π–κ and IsoG–IsoC base-pairs to explain the low base-pairing energy of πκ, despite its three hydrogen bonds. Counting +1 for each secondary interaction between opposite charges and −1 for an interaction between charges of the same sign gives a total of 0 for both GC and IsoG–IsoC and −4 for πκ.

GC base-pair. The IsoG–IsoC base-pair is comparable in stability to GC.

The decreased stability for the πκ and χκ base-pairs compared to GC and IsoG–IsoC (despite the fact that all have three hydrogen bonds) can be understood using the secondary interaction model introduced by Jorgensen[13,14] to explain the interaction between 1-methyluracil (U) and 2,6-diaminopyridine (DAP). In common with the GC and πκ/χκ systems, U-DAP is capable of forming three hydrogen bonds yet was calculated to have an interaction energy of −11.4 kcal/mol using the OPLS potentials (cf. GC −22.1 kcal/mol and AT −10.5 kcal/mol). The apparent anomaly was explained by invoking the presence of secondary electrostatic interactions between the donor hydrogens and the negatively charged oxygen and nitrogen atoms. In common with U-DAP, in both the πκ and χκ systems all four secondary interactions are destabilizing, whereas in IsoC–IsoG the net effect is zero (Figure 4). Jorgensen has estimated the destabilizing effect at between 2 and 3 kcal/mol per interaction; for the πκ and χκ base-pairs and the AMBER force field, a somewhat smaller value seems more appropriate. The reasonable agreement between the calculated energies of the AT and GC base-pairs with AMBER and the experimentally measured gas-phase enthalpies of interaction[15] means that the calculated energies of IsoC–IsoG, πκ, and χκ are likely to be reasonable predictions for the results of mass spectroscopic experiments on these systems.

Table IV. Experimentally Determined Melting Temperatures of the 11 Duplexes d(CAAAMAAAG)·d(GTTTNTTTC)[a]

M, N	experimentally determined melting temp (K)	M, N	experimentally determined melting temp (K)
CG	315	TG	299
AT	313	κA	295
GC	311	GT	294
TA	310	πC	292
πκ	308	κG	291
πT	302		

[a] From ref 1.

Duplex Stability

Piccirilli et al. incorporated the bases π and κ into DNA duplexes of sequence d(CAAAMAAAG)·d(CTTTNTTTG) where **M** and **N** were chosen from π, κ, and the four naturally occurring bases. They measured the melting temperatures of 11 base-paired and mismatched duplexes which are given in Table IV. In the following, each duplex will be referenced by the identities of the bases **M** and **N** (e.g., the "AT duplex" has **M** = A and **N** = T). The AMBER parameters described above were used to investigate the effectiveness of molecular mechanical models in explaining the trends in melting temperature in a manner analogous to some earlier studies.[16] For each of the 11 duplexes in Table IV, an "ideal" B-DNA structure was generated using the NUCGEN program.[7] Each model was then minimized until the norm of the energy gradient was less than 0.01 kcal/mol Å. A distance-dependent dielectric function and an infinite cutoff were employed together with 16 "hydrated" counterions (ϵ = 0.1 kcal/mol; R^* = 5.0 Å[17]) which were initially positioned on the bisectors of each phosphate group.

The 11 duplexes comprise 5 matches (AT, TA, GC, CG, πκ) and 6 mismatches (GT, TG, πT, κA, κG, πC). All the minimized structures suffered some compression at the ends of the helix, due to the imbalance in forces on the terminal base-pairs. Another general feature was that in the A- and T-rich regions of the duplexes the AT base-pairs show quite large propeller twist angles (25°–29°). The tendency of AT base-pairs to adopt a larger twist than GC base-pairs has been observed in the crystal structure of d(CGCGAATTCGCG),[18] it being suggested that the smaller twist in GC base-pairs is due to steric interactions between the guanine amino group and the flanking bases rather than because of the extra hydrogen bond.[19] One effect of such a propeller twist in $d(A_n)\cdot d(T_n)$ regions is to permit the O4 oxygen of thymine to interact not only with the 6-amino group of the adenine with which it is paired but also with the 6-amino group of one of the flanking adenines. The average distance between the thymine O4 and nearest hydrogen of the 6-NH_2 group of the flanking adenine is around 3 Å in the minimized structure of the AT duplex which contains a sequence of seven AT pairs (Figure 5). This is too large to be counted as a "real" hydrogen bond, but is an interaction which may contribute to the structures of such sequences and has

(13) Jorgensen, W. L.; Pranata, J. *J. Am. Chem. Soc.* **1990**, *112*, 2008.
(14) Pranata, J.; Wierschke, S. G.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1991**, *113*, 2810.
(15) Yanson, I.; Teplitsky, A.; Sukhodur, L. *Biopolymers* **1979**, *18*, 1149.
(16) Keepers, J. W.; Schmidt, P.; James, T. L.; Kollman, P. A. *Biopolymers* **1984**, *23*, 2901.
(17) Singh, U. C.; Weiner, S. J.; Kollman, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 755.
(18) Wing, R.; Drew, H.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. *Nature* **1980**, *287*, 755.
(19) Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984; Chapter 11.

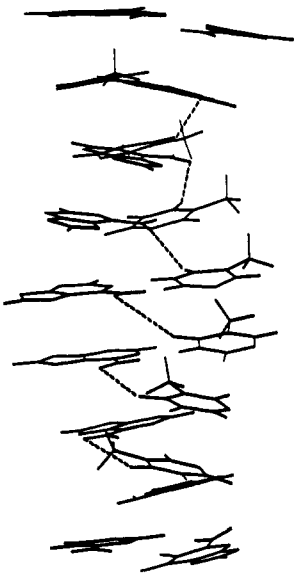

Figure 5. Minimized structure of the AT duplex in which the large propeller twist angles of the AT base-pairs leads to a secondary interaction between adjacent bases.

also been observed in other X-ray crystallography[20,21] and NMR[22,23] studies.

The minimized structures fall into three categories according to the behavior of the central base-pair MN. The first category contains the five matched pairs, in all of which the central base-pair adopts a structure similar to that found for the isolated geometry. The second category, comprising the mismatched GT, TG, πT, and κA duplexes, form "wobble" structures. In these, the initial (Watson–Crick) duplex structures have two or more primary interactions between noncomplementary hydrogen bonding groups, which thus experience a large repulsion. A lateral movement of the bases perpendicular to the helical axis eliminates this strain and enables the base to form two hydrogen bonds. The third category contains the two remaining duplexes (πC and κG). In the initial B-DNA model two of the hydrogen bonding groups are complementary but the third is not. A different mechanism occurs here, namely a twist about the nucleotidic bond which permits the two hydrogen bonds to be maintained while moving the third pair of functional groups further apart.

The furanose rings in an ideal Watson–Crick DNA helix adopt a C2′-endo conformation, characterized by a phase angle[24] of approximately 190°. The sugar phase angles are smaller than this in the minimized structures, corresponding in most cases to a conformation intermediate between C2′-endo and C1′-endo. In addition, between one and four of the sugars in each of the 11 duplexes was distorted further from the initial C2′-endo conformation to give an O4′-endo conformation. In two of the wobble duplexes (the πT and GT mismatches), the sugar bonded to the thymine adopted a C3′-endo conformation, with a phase angle around 15°. By contrast, in the other two wobble duplexes (TG and κA) the sugar on the base N (G or A) is C2′-endo. In the κA duplex the sugar bonded to κ adopts a C2′-endo conformation, but in TG the analogous T sugar conformation is intermediate between C4′-endo and O4′-endo, with a phase angle of 69°. The propeller twist angles between the bases M and N in these four cases are 25° (πT), 12° (TG), 17° (κA), and 26° (GT). The twist angles are thus rather larger for wobble bases MN, where M is a purine or the purine analogue π. Examination of the minimized structures indicates a possible reason for this behavior. In the GT and πT mismatches, the base M (G or π) adopts an orientation coplanar with the flanking adenines. This not only enables the interactions between unpaired adenine and thymine bases described above to be maintained but also retains the favorable purine–purine stacking interactions between the base M (G or π) and the flanking adenines. To form the wobble structure, the thymine sugar is forced into the C3′-endo conformation. In the κA and TG mismatches, the base M is a pyrimidine. The stacking interaction with the flanking adenines is smaller in these cases, and it is energetically less important that coplanarity with the flanking adenines is attained. The difference in sugar pucker of the base M in the TG and κA duplexes arises from an interaction between the base κ with the O4 oxygen of a thymine on the opposite strand (Figure 6).

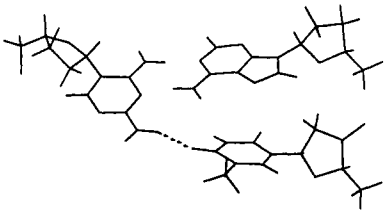

Figure 6. Minimized structure of the κA duplex showing the interaction between the base κ and a thymine on the opposite strand.

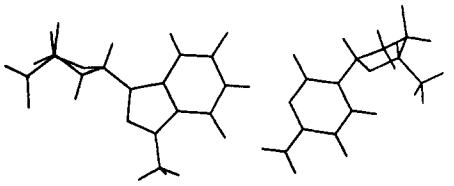

Figure 7. Structure of the mismatched base-pair observed in the minimized structure of the πC duplex.

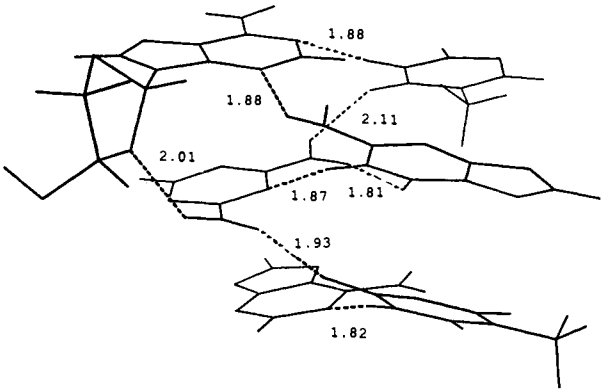

Figure 8. Interactions present in the central d(AκA)·d(TGT) segment of the minimized structure of the κG duplex; possible hydrogen bonds are indicated (distances in Å).

In the πC mismatch, two of the three hydrogen bonding groups are complementary when paired in a Watson–Crick helix while the third pair (O2(π)–O2(C)) is not. To resolve this strain the bases twist (the propeller twist angle is in fact only about 21°, comparable to that observed in matched structures), reducing the interaction between the two oxygens while maintaining the favorable interactions between the other two hydrogen bonding groups (Figure 7). A similar situation arises in the κG mixmatch insofar as two of the three hydrogen bonding groups are complementary. However, the third (unfavorable) interaction is between two NH_2 groups. These have a greater bulk than the carbonyl oxygens in the πC mismatch, and consequently a much more significant twist (41°) about the nucleotidic bond is required to resolve the strain. The minimized structure appears to be stabilized by additional interactions made by the two amino groups of κ with the flanking bases. As shown in Figure 8, in the minimized structure of the κG duplex the 2-amino functionality of κ interacts not only with the O6 of the complementary guanine but also with O4 of a flanking thymine in the opposite strand.

(20) Coll, M.; Frederick, C. A.; Wang, A. H.-J.; Rich, A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 8385.
(21) Nelson, H. C. M.; Finch, J. T.; Luisi, B. F.; Klug, A. *Nature* **1987**, *330*, 221.
(22) Sarma, M. H.; Gupta, G.; Sarma, R. H. *Biochemistry* **1988**, *27*, 3423.
(23) Gupta, S.; Sarma, M. H.; Sarma, R. H. *Biochemistry* **1988**, *27*, 7909.
(24) Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984; Chapter 3.

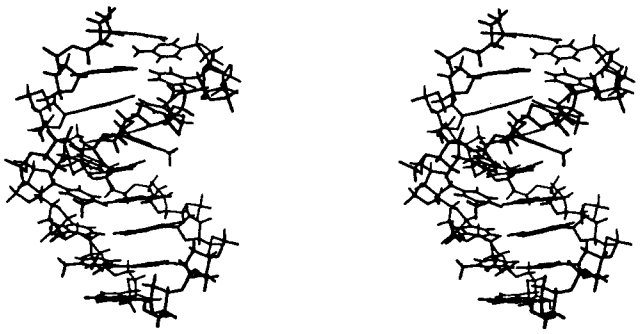

Figure 9. Stereodiagram of the minimized κG duplex illustrating the bend induced in the helix.

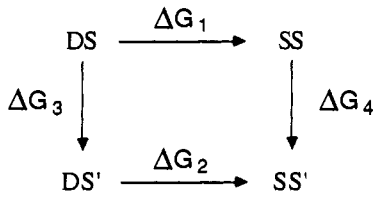

Figure 10. Thermodynamic cycle illustrating the relative stabilities of two duplexes DS and DS′ which dissociate to give single strands SS and SS′. The relative stability of the two duplexes is given by $\Delta G_4 - \Delta G_3$.

The 6-amino functionality of κ interacts with the O2 of the other flanking thymine and with the O4′ sugar atom from an adjacent adenine. The 2-amino group of the guanine interacts with the N3 of a flanking adenine. One consequence of these varied interactions is to induce a severe bend in the helix (Figure 9).

Energy Component Analysis

One objective of the duplex minimization studies was to determine the extent to which the calculations can explain the experimentally observed stability. A rigorous theoretical model to predict the relative stabilities of two sequences of DNA (DS and DS′, Figure 10) would require calculation of the relative free energies of duplex (DS) and single strand (SS), including all solvation and counterion effects (i.e., $\Delta G_4 - \Delta G_3$, Figure 10). The greatest difficulty in doing so is our lack of knowledge of the structures of the single stranded forms. Moreover, such calculations of ΔG_4 or ΔG_3 are quite time-consuming.[25] Thus we sought a simple model which might, qualitatively at least, explain the relative stabilities. Earlier calculations on homo- and heteropolymers[26,27] showed that the relative molecular mechanical energies of duplexes after minimization in vacuo could often qualitatively reproduce the relative stabilities. These studies suggested that energy differences in the duplexes dominated differences in the single strands, which is understandable given the greater conformational heterogeneity of the single strands. However, changes in stability due to changes in single bases or base-pairs in a large duplex are harder to rationalize using total energies, due to end effects and the numerical problem of taking small differences between large numbers. In such cases, we showed that energy component analysis, focusing on the site of mutation, is a useful approach.[16] Here, a small part of the system containing the region of interest is divided into groups whose interaction energies in the minimized structure are calculated. Combinations of the group interactions are examined to determine if the sum of the energies can explain the observed behavior. By this means we aimed to derive a model which can reproduce the trends in duplex melting behavior for the duplexes studied here.

We defined the region of interest to contain the central five base-pairs and their adjoining sugars and phosphates (Figure 11).

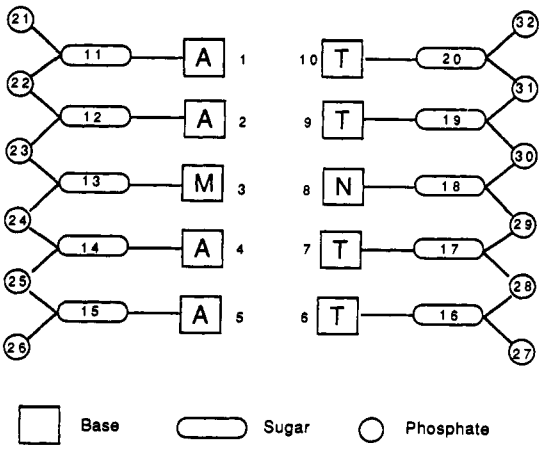

Figure 11. The groups considered for the energy component analysis. The 12 models examined were (1) 5 base-pair segment, considering all base, sugar, and phosphate interactions (groups 1–27); (2) 5 base-pair segment, considering all base and sugar interactions (groups 1–20); (3) 5 base-pair segment, considering all base interactions (groups 1–10); (4) 5 base-pair segment, considering base, sugar, and phosphate interactions between groups on different strands (groups 1–27); (5) 5 base-pair segment, considering base and sugar interactions between groups on different strands (groups 1–20); (6) 5 base-pair segment, considering base interactions between groups on different strands (groups 1–10); (7) 3 base-pair segment, considering all base, sugar, and phosphate interactions (groups 2–4, 7–9, 12–14, 17–19, 22–25, 28–31); (8) 3 base-pair segment, considering all base and sugar interactions (groups 2–4, 7–9, 12–14, 17–19); (9) 3 base-pair segment, considering all base interactions (groups 2–4, 7–9); (10) 3 base-pair segment, considering base, sugar, and phosphate interactions between groups on different strands (groups 2–4, 7–9, 12–14, 17–19, 22–25, 28–31); (11) 3 base-pair segment, considering base and sugar interactions between groups on different strands (groups 2–4, 7–9, 12–14, 17–19); (12) 3 base-pair segment, considering base interactions between groups on different strands (groups 2–4, 7–9).

The nonbonded and electrostatic interaction energies between these groups in the minimized duplex structures were calculated. Combinations of these component energies were then examined to determine how well their energy sum reproduced the trends in duplex melting temperature. The 12 models investigated are indicated in Figure 11 and can be categorized as follows. Segments containing five base-pairs and segments containing three base-pairs were considered. For a given segment size (i.e., five or three base-pairs), we considered one model which contained bases, sugars, and phosphates, one model containing just bases and sugars, and one model consisting of the bases alone. Within each of these three subcategories we calculated both the energies of all intergroup interactions and the energies between groups on different strands which would arise solely as a result of duplex formation. The resulting energies for these various models are shown in Table V.

None of the models is capable of reproducing the trends in melting temperature for all 11 duplexes. In particular we note that the values for the AT and TA duplexes are consistently lower than would be expected from their melting temperatures. In order to try and identify the "next-best" candidates, we considered how well the various models were able to explain the trends in melting temperature for subsets of the duplexes, concentrating on the three pairs of duplexes which have the same composition (i.e., CG vs GC, AT vs TA, TG vs GT). In these cases one would expect the greatest cancellation of other factors not considered in this analysis, the most important of which is solvation, which one would expect to be significantly more favorable for single strands containing GC base-pairs than AT base-pairs (vide infra). Three of the 12 models (numbers 10, 11, and 12) gave the correct relative ordering for each pair; Figure 12 graphically shows the relationship between the interaction energy and the experimental melting temperatures for model 10. This graph illustrates that a reasonable correlation is obtained for some of the duplexes, but there are some anomalous values which deviate from the general trend. In particular, the calculated enthalpy values for the AT, TA, πκ, and πT duplexes

(25) Dang, L. X.; Pearlman, D. A.; Kollman, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 4630.
(26) Kollman, P. A.; Weiner, P. K.; Dearing, A. *Biopolymers* **1981**, *20*, 2583.
(27) Tilton, R. F., Jr.; Weiner, P. K.; Kollman, P. A. *Biopolymers* **1983**, *22*, 969.

Table V. Results of Energy Component Analysis[a]

	1	2	3	4	5	6
CG	−188.8	−103.9	−150.6	−55.1	−71.9	−91.4
AT	−177.1	−95.8	−138.8	−41.4	−58.3	−78.3
GC	−199.2	−100.2	−145.7	−59.6	−73.8	−84.9
TA	−179.6	−94.1	−141.0	−41.1	−57.9	−79.0
πκ	−193.9	−113.0	−145.8	−54.3	−70.5	−84.1
πT	−189.5	−108.7	−139.9	−48.3	−62.7	−80.8
TG	−180.9	−97.4	−139.1	−49.4	−64.9	−83.2
κA	−185.5	−103.5	−145.5	−52.9	−66.5	−84.4
GT	−188.7	−102.2	−141.6	−48.3	−63.4	−81.2
πC	−190.2	−108.9	−139.7	−47.0	−62.5	−77.4
κG	−185.2	−107.2	−145.6	−51.2	−71.8	−87.5
	7	8	9	10	11	12
CG	−119.7	−89.7	−84.2	−45.1	−50.7	−56.7
AT	−112.2	−81.8	−76.1	−33.8	−39.6	−46.1
GC	−120.3	−90.6	−84.7	−42.6	−47.7	−52.8
TA	−111.4	−80.7	−73.5	−31.6	−37.1	−43.9
πκ	−118.2	−89.5	−81.2	−37.8	−45.4	−50.6
πT	−117.1	−85.5	−73.0	−34.2	−39.5	−45.1
TG	−111.1	−83.0	−73.5	−37.6	−43.1	−47.9
κA	−112.9	−83.3	−78.0	−36.9	−41.8	−47.6
GT	−117.7	−86.6	−75.7	−35.4	−41.1	−46.3
πC	−114.5	−84.2	−75.7	−33.6	−38.8	−44.5
κG	−113.9	−85.0	−80.2	−39.7	−46.5	−51.9

[a] Values are in kcal/mol for the sum of the vdw (van der Waals) and electrostatic interaction energy, for the combinations of groups shown in Figure 10.

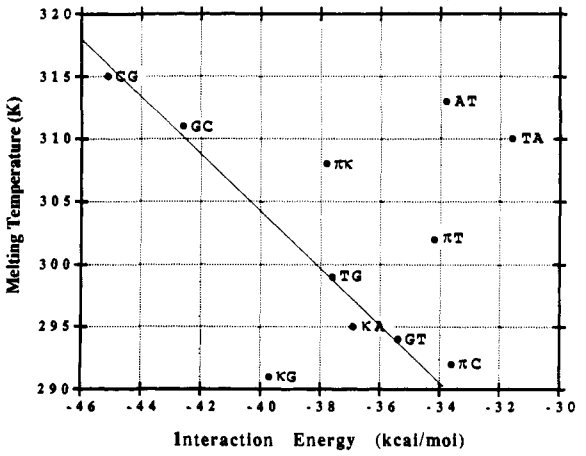

Figure 12. Graphical illustration of the correlation between the energy sum for model number 10 (see caption to Figure 10) and the experimentally determined melting temperatures for duplexes d(CAAAMAAAG)·d(GTTTNTTTC).

are too small to explain their observed stability, and the value for the κG duplex is too large.

It is worth noting that models 10–12 reproduce the trend in melting temperature whereas models 7–9 do not. The difference between these models (7–9 include intrastrand effects, whereas 10–12 do not) suggests that there may be considerable stacking in the single stranded forms of these duplexes, such that these intrastrand stacking energies occur in both single and double stranded forms and should not be included when comparing *relative* double stranded stabilities. All six models (7–9, 10–12) correctly predict the AT duplex to be more stable than TA, consistent with experiment and theory[26] on homo- versus heteropolymer AT stabilities. As noted there, poly(dA·dT) is more stable than poly(d(AT))·poly(d(AT)) because the van der Waals stacking is more favorable in the homopolymer duplexes due to purine stacking. Even though such stacking is likely to be present in the single polypurine single strand, it is not as stabilizing as in the duplex. However, such van der Waals effects cannot rationalize the greater stability of CG than that of GC or of TG than that of GT; in fact, one would expect the opposite trend based on van der Waals energies alone, since these would be most favorable for all purine–all pyrimidine duplexes. A particularly favorable electrostatic interaction that can rationalize these stabilities[16] is the interstrand major groove interaction between guanine O6 and adenine 6-NH_2; this interaction would be present in CG and TG but not in GC or GT. Calladine's rules[28] have emphasized steric clashes involving cross-strand NH_2 interactions involving adenine; we suggest that cross-strand attractive effects between G and A can also be important.

Incorporation of Solvation Effects

In an attempt to improve the purely enthalpic model, we sought to include some measure of the effect of solvation. For duplexes which differ by a single base-pair (as is the case here), one of the major contributions to differences in the solvation free energy for the duplex to single strand transition would be due to differences in free energy of solvation of the two variable bases M and N. Such free energies can be calculated by the free energy perturbation technique, and values of −14.7, −9.6, −14.9, and −21.7 kcal/mol have previously been reported[29] for the free energy of solvation (relative to methane) of the bases A, T, C, and G, respectively. The effect of single strand solvation on the sequences studied here would thus be to decrease the melting temperature of strands containing CG base-pairs more than those containing GT base-pairs which in turn would be affected more than AT base-pairs. Whereas the enthalpic term favors duplexes containing GC base-pairs over those containing AT base-pairs, the more negative free energy of solvation of G and C compared to A and T would tend to lessen this effect. In extremis, the effect of incorporating such a solvation term would be to "penalize" a GC duplex by 12.3 kcal/mol relative to one where the GC base-pair is substituted by AT.

The effect of solvation was incorporated in a more quantitative fashion by expressing the melting temperature as a linear function of the component analysis enthalpy (ΔH) and the sum of the relative free energies of solvation of the two single strands (each of which is approximated by the relative free energies of the bases M and N, $\sum(\Delta\Delta G_{solv})$):

$$T_m = \alpha\Delta H + \beta\sum(\Delta\Delta G_{solv}) + \gamma$$

Least-squares fitting of the six data points for the duplexes containing the naturally occurring bases using relative free energies of solvation from ref 29 gave values for α, β, and γ of −4.1 (1.3), 3.6 (1.2), and 265 (19), respectively, the numbers in parentheses being the standard deviations for the coefficients. The overall standard deviation for the fit was 5.3. While the significance of these values should not be over-interpreted, we note that the coefficients have physically meaningful signs (i.e., a more negative duplex interaction energy leads to a higher melting temperature while a more negative relative free energy of solvation results in a lower melting temperature).

To extend the scope of the analysis to the bases π and κ, we have performed free energy perturbation calculations to derive their relative free energies of solvation. κ was mutated into C and π into G (and vice versa) using statistical perturbation theory and molecular dynamics simulations. In this technique the free energy change between two states (0 and 1) is determined by generating a molecular dynamics trajectory along which the system is perturbed from the initial state to the final one. As free energy is a state function, the pathway for the perturbation need not be a chemically meaningful one. A number of methods are available for performing such calculations. We used the recently developed dynamically modified windows method[30] which calculates the free energy at discrete values of λ (the value of which varies between 0 and 1 according to the current state of the system) using the formulas:

$$G_{\lambda(i+1)} - G_{\lambda(i)} = -RT \ln \langle \exp\{-[(V_{\lambda(i+1)} - V_{\lambda(i)})/RT]\}\rangle_{\lambda(i)}$$

$$\Delta G = G_1 - G_0 = \sum G_{\lambda(i+1)} - G_{\lambda(i)}$$

(28) Calladine, C. R. *J. Mol. Biol.* **1982**, *61*, 343.
(29) Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. *Science* **1987**, *236*, 564.
(30) Pearlman, D. A.; Kollman, P. A. *J. Chem. Phys.* **1989**, *90*, 2460.

Dynamically modified windows differ from the traditional windows approach through the use of a variable window size ($\delta\lambda$) which is automatically changed according to the value of the slope $\partial G/\partial\lambda$. This allows the simulation to run more quickly when the free energy is changing slowly and more slowly when it is not. Calculations in both the forward and reverse directions were performed. Each simulation was performed in a periodic box of side approximately 25 Å; 457 TIP3P water molecules were used to solvate the base π and 421 for κ. Each system was first minimized until the maximum gradient was less than 0.1 kcal/mol·Å and then equilibrated for 10 ps using a time step of 1 fs during which the temperature was raised to 298 K. An 8.0-Å cutoff was employed and SHAKE[31] was used to constrain bonds involving hydrogen. During the perturbation calculations 40 steps of equilibration were performed for each value of λ before collecting the statistics over a subsequent 40 steps. Double-wide sampling was used throughout. For the mutation of π ($\lambda = 1$) into G ($\lambda = 0$), the target free energy change was 0.01 kcal/mol for each window. A total of 2136 windows were used, the calculation requiring 170 ps. The total change in free energy was determined to be 12.3 kcal/mol for ($\lambda = 0 \rightarrow 1$) sampling and −12.4 kcal/mol for ($\lambda = 1 \rightarrow 0$) sampling. To reduce the amount of simulation time, a larger target free energy change per window (0.025 kcal/mol) was used for the reverse perturbation. (This illustrates one of the problems with the dynamically modified window approach in that the amount of simulation cannot be accurately predicted until the calculation is complete.) For the G $\rightarrow \pi$ perturbation 867 windows were used and the free energy values were 13.6 kcal/mol for ($\lambda = 0 \rightarrow 1$) sampling and −15.6 kcal/mol for the reverse. The effect of using a larger target free energy per window is clearly manifested in the larger difference in the two sampling directions which is, however, adequate for the present purposes. The relative free energy of π relative to G is thus calculated to be +13.5 with a lower bound on the error of 2.1 kcal/mol. For the mutation of κ into C (and C $\rightarrow \kappa$), the 0.025-kcal/mol target free energy change per window was retained; 337 windows were required (27 ps) for the mutation of κ into C with the ($\lambda = 0 \rightarrow 1$) free energy change being 2.8 kcal/mol and the ($\lambda = 1 \rightarrow 0$) value −2.8 kcal/mol. For the C $\rightarrow \kappa$ perturbation 345 windows were required with a total simulation time of 27.7 ps. The free energy change for ($\lambda = 0 \rightarrow 1$) sampling was 2.7 kcal/mol and −2.6 kcal/mol for ($\lambda = 1 \rightarrow 0$). The overall relative free energy of solvation for $\kappa \rightarrow$ C is thus calculated to be −2.7 kcal/mol with a lower bound on the error of 0.1 kcal/mol. Relative to methane this gives values of −12.2 kcal/mol and −8.2 kcal/mol for κ and π, respectively, placing κ intermediate in solvation free energy between C and T, whereas π has a lower value than either G and A.

It is instructive to compare these values with the dipole and quadrupole moments of the molecules. As noted previously,[29] adenine has a larger electrostatic contribution to the free energy of solvation than does thymine and yet has a somewhat smaller dipole moment, the apparent anomaly being accounted for by adenine's higher quadrupole moment. The dipole moments of π and κ (calculated from the STO-3G partial charges) are 1.5 D and 2.4 D, respectively. The principal values of the quadrupole moments for π and κ (Q_a, Q_b, Q_c) are −11.0, 6.3, 4.7 and 8.8, −7.9, −1.0. The low dipole and quadrupole moments of π may thus explain its low free energy of solvation relative to the other nucleic acid bases. However, the dipole and quadrupole moments of κ are both smaller than those of thymine, yet κ has a more favorable free energy of solvation. While the first two multipole moments appear to be able to explain the solvation behavior of the naturally occurring bases and the base π, such a model appears not to extend to the base κ.

From the free energy calculation described above, the effect of solvation on the duplex melting temperature of the five duplexes involving the bases κ and π would be exepected to follow the order πG > πA > πC > $\pi\pi$ > πT. The high solvation penalty for the κG duplex could thus explain to some degree the discrepancy between the calculated enthalpy value and the melting temperature. Similarly, the low solvation penalty for πT could explain why this base-pair is more stable than the enthalpic calculation would suggest. Applying the least-squares fitting procedure as described above to the five duplexes containing the novel bases gave values for the coefficients α, β, and γ of −2.8 (0.1), 1.6 (0.03), and 237 (2) with an overall standard deviation of 0.2. The coefficients do therefore have the same signs as those for the naturally occurring bases. In addition, both the α and β coefficients are much smaller in size than for the naturally occurring bases, indicating both a weaker enthalpic contribution from the base-pairs and that the solvation term is less important in these cases. This could be due to the greater tendency of duplexes involving the bases π and κ to adopt distorted structures in which the bases are not in the canonical form suggested by energy minimization, thereby reducing the magnitude of the favorable enthalpic interactions in the double helix and the difference in solvation free energy between the duplex and the single strands. These differences could explain why, when the fitting procedure was applied to the entire set of 11 duplexes, the correlation was essentially nonexistent ($\alpha = -1.2$ (1.1), $\beta = 0.5$ (0.7), $\gamma = 272$ (30); overall standard deviation = 9.4). Clearly, while the model is able to explain in a qualitative sense the influence of solvation effects on duplex dissociation, other factors need to be taken into account to completely understand the observed behavior for all 11 duplexes.

Molecular Dynamics Simulations of κG Duplex

The structure of the minimized κG duplex is so unlike any others obtained in these and earlier studies that we were concerned it might simply be an artifact of the minimization. In this section we therefore report the results of a molecular dynamics simulation which was performed to provide a more realistic picture of the behavior of this particular system. The κG duplex was enclosed by a 6-Å solvent shell containing 612 water molecules. A positively charged counterion ($\epsilon = 0.1$ kcal/mol; $R^* = 1.6$ Å) was positioned near each phosphate group. The water molecules and counterions were first minimized until the maximum gradient was lower than 0.5 kcal/mol·Å. They were then allowed to reorient in a 10-ps simulation during which the DNA was kept fixed. The whole system (duplex, waters, and ions) was then minimized to a gradient of 0.3 kcal/mol·Å. Finally, 40 ps of molecular dynamics was run using a nonbonded cutoff of 10 Å and a time step of 1 fs during which the temperature was raised from 10 K to 298 K over approximately the first 2.2 ps. SHAKE was applied to all bonds involving hydrogen. Coordinate sets were saved every 100 steps. A control experiment using an identical protocol was also performed for the CG duplex.

Our primary interest in the molecular dynamics simulation was to predict the structure of the κG duplex in solution and to compare it with that obtained by minimization. Visual observation of the trajectory indicated that the duplex remained intact during the simulation with no noticeable fraying of the helix ends. In light of the interactions made by κ in the minimized structure, a hydrogen bonding analysis was performed to determine the interactions made by this base during the simulation. Interactions between all possible hydrogen donor and acceptor atoms in the middle three base-pairs d(5′ AκA 3′)·d(3′ TGT 5′) were monitored for the presence of hydrogen bonds, the criteria for acceptance being a maximum donor–acceptor distance of 3.0 Å and an angle at the hydrogen greater than 90°. The results of this analysis are graphically illustrated in Figure 13. The graph shows the hydrogen bonds which are present for more than 20% of the simulation. One of the AT base-pairs exhibits the expected hydrogen bonding behavior with two hydrogen bonds being consistently present. However, no hydrogen bonds are formed in the other AT base-pair according to these criteria. To compensate, there is a consistent interaction between the thymine of this base-pair and one of the amino groups on the base κ. The remaining two hydrogen bonding functional groups on κ interact with the complementary functionalities on the base G, giving the hydrogen bonds N1(G)–N1(κ) and N2(κ)–O6(G). The large twist angle

(31) van Gunsteren, W. F.; Berendsen, H. J. C. *Mol. Phys.* **1977**, *34*, 1311.

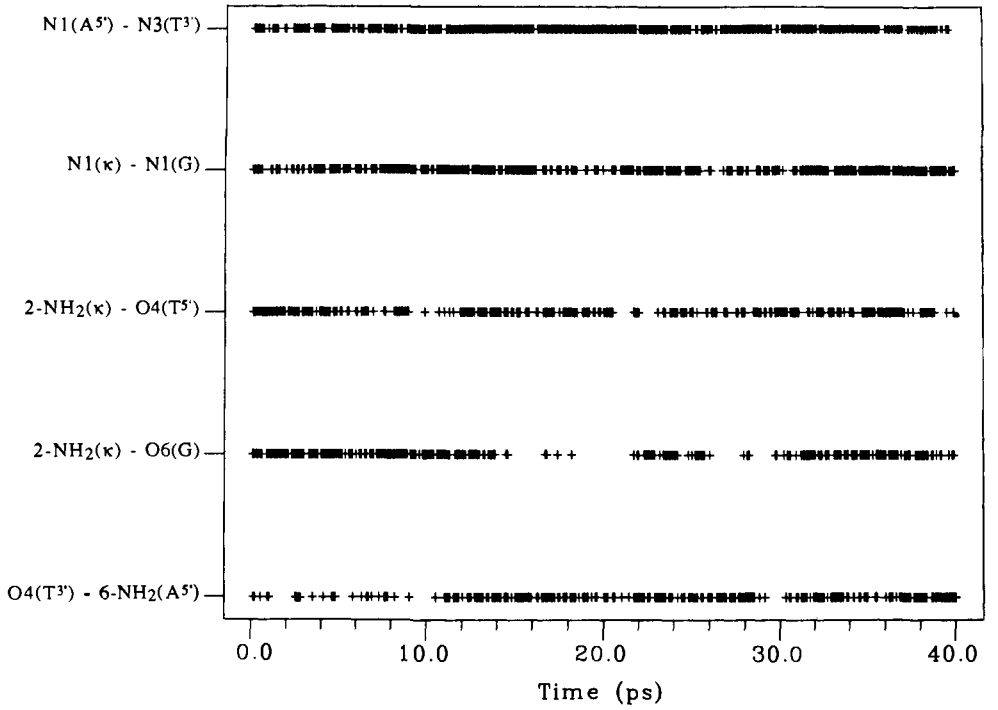

Figure 13. Hydrogen bonding analysis of the central three base fragment d(AκA)·d(TGT) for the 40-ps molecular dynamics simulation of the κG duplex in water. The presence of a hydrogen bond in a given coordinate set is indicated by a "+" symbol. The analysis was performed from the 400 sets of coordinates obtained at 0.1-ps intervals. The A and T bases are distinguished according to which end (5′ or 3′) of the appropriate strand they are closest to.

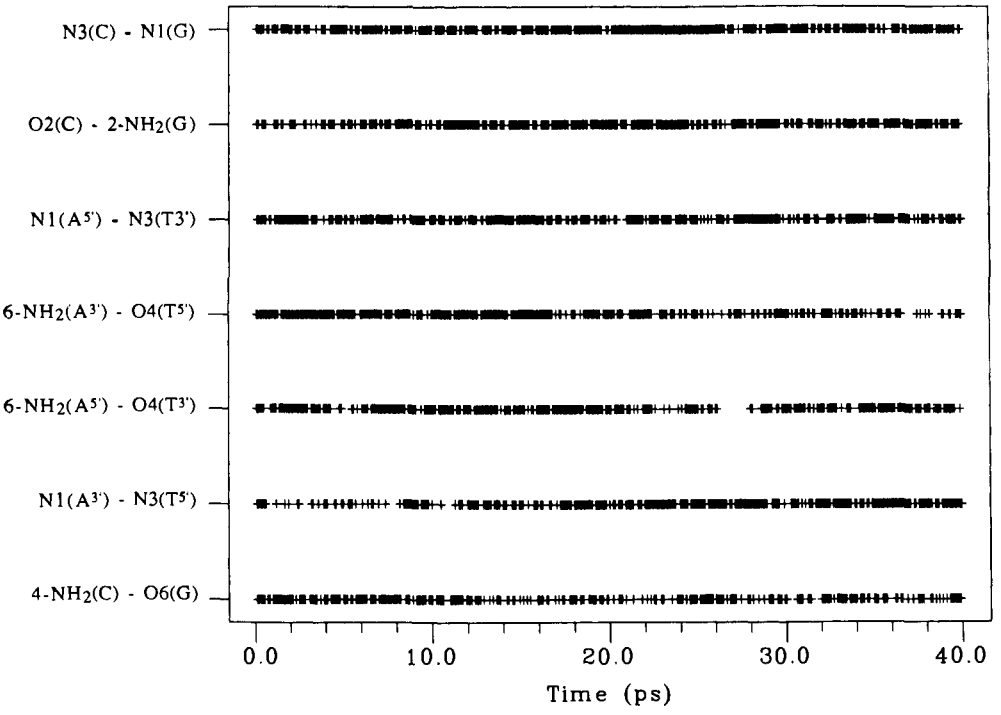

Figure 14. Hydrogen bonding analysis of central three base-pair fragment d(AGA)·d(TCT) for 40-ps simulation of the CG duplex.

between κ and G observed in the minimized structure is a consistent feature throughout the simulation, with a mean value of 52°. A comparable hydrogen bond analysis was performed on the CG duplex, the results of which are displayed in Figure 14. The seven hydrogen bonds which would be expected for the central ACA·TGT fragment remain essentially intact through the entire course of the simulation.

Discussion and Conclusions

With the recent demonstration by Piccirilli et al. that a novel base-pair can be enzymatically incorporated into both DNA and RNA, it is a propitious time to speculate again why nature chose to use the two particular base-pairing schemes represented by adenine–thymine and guanine–cytosine. While many factors may, of course, be involved (not the least of which is chance), contemporary theoretical techniques such as those described above can provide some insight into the behavior and properties of such systems.

The semiempirical and molecular mechanics calculations on the isolated base-pairs indicate that the πκ base-pair appears to be comparable in stability with its natural counterparts. The molecular mechanical duplex calculations demonstrate that simple

models can provide information that may help to understand the behavior of complex systems. Nevertheless, a purely enthalpic model is unable to provide a complete explanation for the variation in duplex melting temperature for all 11 duplexes studied. While we are still far from being able to provide a total free energy analysis of duplex melting, incorporation of a solvation term provides a qualitative explanation for the anomalous results and does indicate possible areas for improving such models. In addition, the fact that an energy component analysis which includes only interstrand interactions gives the best qualitative agreement with relative stabilities of the duplexes suggests considerable intrastrand stacking in the single stranded forms of these molecules and that it would be interesting to study the relative stabilities of duplexes of the type d(CAAAMTTTG)·d(CAAANTTTG), which may have different properties and relative stabilities. The nature of the κG base-pair in such a duplex, where there is an adenine 5′ to the G and which would not allow the same unusual hydrogen bonding pattern as found in d(CAAAκAAAG)·d(CTTTGTTTG), would be very interesting from both a theoretical and experimental point of view.

In many respects some of the most interesting questions raised by our analysis concern the mismatched structures formed by the new bases. Two mismatches are possible in a Watson–Crick double helix with the four naturally occurring bases (ignoring mismatching due to tautomerism). The GT mismatch appears to be the most prevalent, due in part to its ability to form two hydrogen bonds by wobbling. With the introduction of the bases π and κ, the number of possible mismatched base-pairs increases to six. In the Aκ and πT mismatches the arrangement of hydrogen bonding groups is such that there is a net destabilizing effect of the primary interactions (i.e., there are more noncomplementary interactions than complementary ones), and similar wobble behavior to GT would be expected. In the other two mismatches (πC and κG), however, two of the three interactions in the Watson–Crick structure are between complementary hydrogen bonding groups while the third is destabilizing. There is no analogous situation for mismatches between the dominant tautomers of the naturally occurring bases. In order to maintain the two favorable interactions while reducing the effect of the third unfavorable one, an alternative, twisting mechanism is observed. The base κ in particular appears from both molecular mechanics and molecular dynamics calculations to be able to form additional hydrogen bonding interactions when thymines flank the base to which it is paired, as demonstrated by our results for the κG mismatch. However, it should be pointed out that molecular dynamics simulations in solution are weak controls for molecular models predicted to be stable by in-vacuo energy minimization. If such models are trapped in a shallow minimum, the simulation may change their structure in a few picoseconds. The fact that the κG duplex was stable for 40 ps is no guarantee that it is correct, only that further experimental (e.g., 2D NMR) and theoretical study is warranted.

In summary, the application of molecular mechanics, molecular dynamics, and free energy perturbation calculations have given useful insight into the nature of duplexes involving new and unusual base-pairs. A model using the results of the molecular mechanics and free energy perturbation calculations could explain many but not all trends in relative duplex stability. The incorporation of solvation free energies enable us to go one step beyond the simple model of Keepers et al.[16] in rationalizing duplex stability. Energy minimizations led to an unusual structure in the κG duplex, and molecular dynamics simulations in water confirmed that it was at least a viable candidate for further experimental and theoretical study.

Acknowledgment. A.R.L. thanks the Science and Engineering Research Council (U.K.) and NATO for a postdoctoral fellowship and the Computer Graphics Laboratory for their kind hospitality. The Computer Graphics Laboratory is supported by the National Institutes of Health (Grant R1081; R. Langridge PI). Molecular graphics images were produced using the MidasPlus software system.[32] Support was also provided by the Defence Advanced Research Projects Agency under Contract N00014-86-K-0757 administered by the Office of Naval Research. P.A.K. acknowledges research support from the National Institutes of Health (Grant CA-25644). We thank the San Diego Supercomputing Centre for the provision of some of the computing facilities used in this work and Dr. C. A. Leach for assistance with the fitting part of the analysis.

Registry No. π, 109421-37-6; κ, 18588-37-9; IsoG, 54746-36-0; IsoC, 2080-17-3; χ, 91746-29-1; d(CAAACAAAG)·d(GTTTGTTTC), 97937-46-7; d(CAAAAAAAG)·d(GTTTTTTTC), 97937-43-4; d(CAAAGAAAG)·d(GTTTCTTTC), 97937-49-0; d(CAAATAAAG)·d(GTTTATTTC), 97937-52-5; d(CAAAπAAAG)·d(GTTTκTTTC), 139913-80-7; d(CAAAπAAAG)·d(GTTTTTTTC), 139913-81-8; d(CAAATAAAG)·d(GTTTGTTTC), 97937-53-6; d(CAAAκAAAG)·d(GTTTATTTC), 139895-58-2; d(CAAAGAAAG)·d(GTTTTTTTC), 97937-56-9; d(CAAAπAAAG)·d(GTTTCTTTC), 139895-60-6; d(CAAAκAAAG)·d(GTTTGTTTC), 139895-61-7.

(32) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. *J. Mol. Graph.* **1988**, *6*, 13.