Guanine−Cytosine Base Pairs in Parallel-Stranded DNA: An ab Initio Study of the Keto−Amino Wobble Pair versus the Enol−Imino Minor Tautomer Pair

Daniel Barsky* and Michael E. Colvin

Biology and Biotechnology Research Program, Lawrence Livermore National Lab, L-448, Livermore, California 94550

Received: April 14, 2000; In Final Form: June 22, 2000

Doubly hydrogen bonded, "reverse Watson−Crick" thymine−adenine base pairs make possible the formation of parallel-stranded DNA double helices. Although the presence of guanine and cytosine reduces the stability of parallel-stranded DNA, the rather modest experimentally determined reduction in stability (less than 1 kcal/mol for each C−G pair) has been ascribed separately to favorable amino−amino contacts, tautomerizations, and a wobble pair geometry. Earlier studies predicted that favorable amino−amino contacts could yield an interaction energy (gas phase) of about −5 kcal/mol for a twisted, reverse Watson−Crick C−G base pair. It is shown here that either a minor tautomer pairing or a wobble pairing can much more strongly stabilize reversed C−G base pairs. The calculated gas-phase interaction energies of −14 kcal/mol each are comparable to the gas-phase stability of a T−A base pair. Aqueous-phase calculations, however, greatly favor the wobble pair geometry by 9 kcal/mol.

1. Introduction

In 1986 Pattabiraman reported model-building studies which demonstrated the feasibility of parallel DNA double helices composed of poly d(A)·poly d(T) strands.[1] The symmetry of the Watson−Crick (WC) hydrogen-bond donors and acceptors about the central thymine-N3−adenine-N1 axis of a T−A base pair allows hydrogen bonds to be formed for each T−A base pair in almost the same place for parallel-stranded DNA as for antiparallel-stranded DNA. The hydrogen-bonding arrangement for T−A pairs in parallel-stranded DNA has thus been called "reverse Watson−Crick".[2]

In 1988 Jovin and co-workers reported experimental observations of parallel-stranded hairpin DNA (containing a 3′-p-3′ linkage)[3] and parallel-stranded DNA duplexes consisting of alternating poly d(A)·poly d(T) tracts.[4] It was later verified by Raman spectroscopy that the base-pair hydrogen bonding structure is reverse WC.[5] It was not known what effect the presence of reverse C−G pairs (Figure 1, IV) would have on the formation, stability, and structure of parallel-stranded DNA. A WC C−G base pair (Figure 1, I), unlike T−A, is not symmetric about a central axis with respect to hydrogen-bond donors and acceptors.

In 1990 Rippe et al. reported the stability of a 25-nt parallel-stranded DNA duplex containing four C−G pairs. Each C−G pair lowered the melting temperature by about 3 °C and thus destabilized the parallel duplex by 0.7 kcal/mol.[6] For a single base pair this destabilization is mild—comparable to a hydrogen-bonded G−A mismatch in the least favorable context.[7] Rippe et al. also proposed a model, based on molecular mechanical energy minimization, where the reverse C−G pairs are propeller-twisted and form a single hydrogen bond about the central base-pair axis [(C)N3···H1−N1(G)][6]. This is depicted schematically in conformation IV in Figure 1.

The relative stability of the twisted, reverse C−G pair is surprising because of the unfavorable carbonyl−carbonyl and amino−amino contacts and the presence of only a single hydrogen bond. Šponer and Hobza published ab initio quantum chemical calculations showing that the amino−amino contacts could actually be energetically favorable.[8] They showed that, for a propeller-twisted geometry similar to that of the Rippe et al. model,[6] a twisted, reverse WC C−G base pairing was mildly favorable in energetic terms; they found an electronic interaction energy of −4.8 kcal/mol [MP2/6-31G(d,p)//HF/6-31G(d), counterpoise corrected] using constrained optimization techniques.[8] The equivalent calculation on the optimized geometry for WC C−G yields −24.4 kcal/mol. The authors conclude that the stabilization energy of the reverse WC pairing, due to the amino−amino bifurcated hydrogen bonds, is sufficient to explain the incorporation of the C−G pair into parallel-stranded DNA. Importantly, this predicted twisted, reverse WC C−G configuration is not a local minimum for the isolated base pair. Also important is that this is a gas-phase result; effects of the solvent were not included.

Two alternative arrangements leading to the relative stability of C−G in parallel-stranded DNA are the appearance of the minor tautomer forms of C and G (Figure 1, VI) and the formation of a wobble base pair (Fig. 1, V).[9−11] As we will show here a favorable feature of the minor tautomer reverse C−G base pair (VI) is that the shape is very close to that of reverse T−A and can be accommodated into the parallel double helix without distortion of the DNA backbone. In contrast, the reverse wobble C−G (V) requires a small distortion of the backbone. In antiparallel DNA occurrence of minor tautomers has long been postulated as a source of substitution mutations,[12] and based on the experimentally observed frequency of the minor tautomers, the predicted frequency of occurrence of non-Watson−Crick base pairs including A−C, G−T, A−A, G−G, and G−A seems to correlate well with frequencies of spontaneous substitution mutations.[13] Recently, an investigation of DNA helices containing deoxyisoguanosine (iG) revealed both minor tautomers and wobble pairs for the iG−T pairing by both X-ray crystallography and NMR,[14] and in RNA helices G−U base pairs were observed in wobble form.[15]

* Corresponding author. E-mail: barsky@llnl.gov.

10.1021/jp001420d CCC: $19.00
Published on Web 08/19/2000

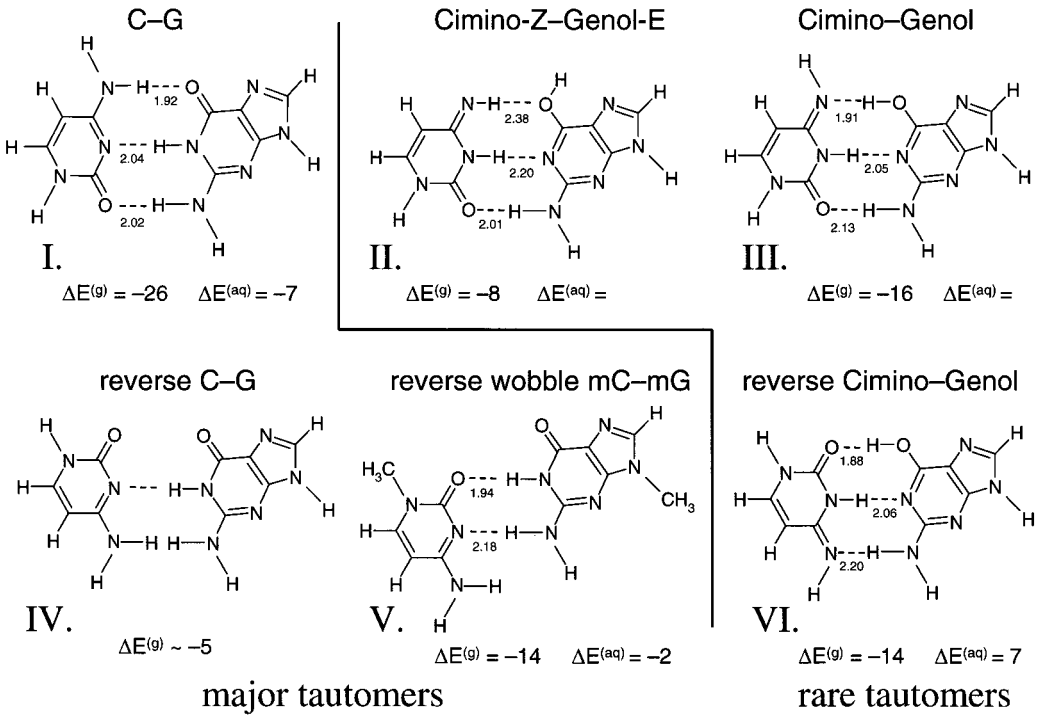

Figure 1. Top row: Watson−Crick (antiparallel) orientation for C−G (I) and two rotamers of the Cimino−Genol minor tautomer base-pairs(II & III). Bottom row: Reverse Watson−Crick (parallel) orientation for reverse C−G (IV), reverse wobble mC−mG (V), and reverse Cimino−Genol (VI) base pairs. Below each base pair are the gas phase ($\Delta E^{(g)}$) and solvent phase ($\Delta E^{(aq)}$) interaction energies (kcal/mol), taken from Tables 1 (column 3) and 3 (column 3), respectively.

A third possibility, other than minor tautomer or wobble pairs, is that the G and C bases could rotate out of the helix altogether. In light of the following evidence, we consider this possibility very remote. In 1960 Fresco and Alberts showed that in pairings of long poly(AU) tracks with poly U oligomers, the "extra" U bases of the poly(AU) are looped (bulged) out of the helix.[16] It turns out that the U−U (or T−T) base pairings are among the least stable base pairings,[17,18] presumably because pyrimidine−pyrimidine (pyr−pyr) pairs do not stack as favorably as purine−pyrimidine (pu−pyr) pairs or because they distort the backbone of the helix. Even so, in one of the few crystal structures containing mismatched bases, that of an RNA double helix containing both C−U (which is pyr−pyr) and G−U (pu−pyr) mismatches, all bases remained inside the helix and formed "intrahelical" base pairs.[19] The least stable base pairings are those that can form only one hydrogen bond: A−A, A−C, and C−C. The least stable of these, C−C, has combined the disadvantages of a single hydrogen bond and weak stacking interactions, and is the only "base pair" so far observed to exist outside the helix.[20] Since an intrahelical C−G pair in parallel-stranded DNA would have both the favorable pu−pyr stacking and, in the two arrangements mentioned above, at least two hydrogen bonds, we conclude that these arrangements are far better candidates than an extrahelical arrangement. This view seems even more reasonable when it is realized that the C−G pair destabilizes the parallel stranded duplex by much less than the removal of even one A−T base pair in anti-parallel DNA (1.5 kcal/mol). Finally, the recent study of parallel-stranded DNA by FTIR and UV spectroscopy gives strong evidence of C−G base pairing.[11]

Using isocytosine (iC) as a model for guanine (iC and G have the same polar groups at the WC-like interface), Zhanpeisov et al.[21] have calculated association energies for a reverse WC-like enol-isocytosine−keto-cytosine ("iCCc") base pair that is analogous to the minor tautomer pair VI in Figure 1, and a WC-like base pair ("iCC1") that is analogous to the WC C−G usual base pair I in Figure 1. They showed that in dual tautomerizations the minor tautomer pair (iCCc) is only 9 kcal/mol less stable than the major tautomer base pair (iCC1) in gas-phase MP2/6-31G(d)//HF/6-31G(d) calculations and 15 kcal/mol less stable in supermolecule calculations involving six explicit water molecules to simulate the first layer of hydration.[21,22]

Considering C and G directly and employing higher level optimizations together with a polarizable continuum solvent model (see Methods below), we investigate here the relative stability of the various forms possible for a C−G base pair in parallel-stranded DNA as discussed above and shown in Figure 1.

2. Methods

Because the computational cost of ab initio quantum chemical optimizations goes up rapidly with the number of atoms, we have simulated the bases with a hydrogen replacing the sugar (deoxyribose), except in a few cases where we represent the sugar as a methyl group.

We calculated the association energies of various orientations and chemical forms of the base pairs by unconstrained geometry optimizations in the gas phase. Solvation effects were included in single point energy calculations (see below). In addition to the unconstrained optimizations, we have in some cases performed geometry optimizations in C_s symmetry which constrains the bases to be coplanar. We have then compared the C_s symmetry minimum energies with the energies of the unconstrained minima to estimate the energy required to maintain the planarity of the base pairs.

Using Gaussian 94 and Gaussian 98,[23] we fully optimized the structures of the following compounds and their base pairs: adenine (A), thymine (T), 9-methyl-A (mA), 1-methyl-T (mT), guanine (G), cytosine (C), 9-methyl-G (mG), 1-methyl-C (mC), enol-guanine (Genol), imino-cytosine (Cimino), enol-(*E*)-

guanine (Genol-*E*), and imino-(*Z*)-cytosine (Cimino-*Z*), where "*E*" and "*Z*" designate the opposite rotamers of the hydroxyl or imino group. The base pairs evaluated are shown in Figure 1. For every monomer and base pair, we have done a geometry optimization using the Hartree−Fock (HF) method[24] with a 6-31G(d) basis set [HF/6-31G(d)], and also density functional theory (DFT) optimizations with larger 6-31G(d,p) basis sets, using the Becke three parameter exchange[25] and Lee−Yang−Parr correlation functionals[26] (B3LYP). For the base-pair optimizations we started with the geometries analogous to antiparallel B-DNA or the parallel analogue, obtained by rotating ("reversing") the pyrimidine about the N3−C6 axis. The HF calculations have been followed by MP2/6-31G(d,p) (second-order Møller−Plesset perturbation theory) single point energy calculations at the HF optimized geometries. For each pair we have also calculated the Boys−Bernardi counterpoise (CP) correction to the basis set superposition error (BSSE),[27] at the MP2/6-31G(d,p) level. For these basis sets, the CP correction reduces the BSSE (see Sponer and Hobza[28] and references therein), and has been shown to increase the accuracy of predicted hydrogen-bonding energies.[29−31]

The gas-phase association energies $\Delta E^{(g)}$ are calculated by subtracting the energies of the isolated bases (always the major tautomer forms) from the base-pair energy, as in $\Delta E^{(g)} = E^{(g)}_{A-B} - (E^{(g)}_{A} + E^{(g)}_{B})$. The association energies reported here are calculated directly from the quantum chemical electronic energies. To obtain true association enthalpies, the association energies would have to be summed with the relative zero point nuclear vibration energies and other thermal corrections. The zero point corrections to the enthalpies are calculated from the ab initio-derived harmonic vibrational frequencies, and typically have little effect on relative association enthalpies. For the C−G and T−A pairs at a comparable level of theory but smaller basis sets than used here [MP2/6-31G(d)//HF/3-21G], inclusion of zero point corrections changes the association energy by 2.6 and 1.5 kcal/mol, respectively, but the relative association energy differs by only 1 kcal/mol.[32]

Solvent Phase Calculations. The gas-phase calculations do not take into account solvent effects which reduce the base-pair electrostatic interactions. To remedy this, we have done solvent-phase (aqueous-phase) HF/6-31G(d,p) and B3LYP/6-31G(d,p) calculations of the bases and base pairs, using a conductor solvent model (COSMO)[33] as implemented in Gaussian 98 [keyword: scrf=(cpcm,solvent=water)].[23] The conductor solvent model accounts for electrostatic water−solute interactions, and the Gaussian 98 implementation includes approximate nonelectrostatic terms for the solute−solvent interaction, including a cavitation term based on a scaling of the molecular surface area[34−36] and dispersion and repulsion terms based on hard-sphere models. The solvent calculations yield an estimate of the solvation free energy of the bases, as well as association energies of the solvated base pairs. The solvent-phase association energies $\Delta E^{(aq)}$ are calculated by subtracting the energies of the isolated bases (the major tautomers) from the base-pair energy, as in $\Delta E^{(aq)} = E^{(aq)}_{A-B} - (E^{(aq)}_{A} + E^{(aq)}_{B})$. Since it is not possible in the current implementation of the solvent model to apply the counterpoise correction to the solvent-phase calculations, we have recalculated the B3LYP/6-31G(d,p) geometries with a larger basis set [6-31G++(d,p)], including diffuse functions to reduce the basis set superposition error. A recent study of C−G and T−A base pairs compared optimizations using Gaussian-type functions with BSSE-free optimizations using plane-wave functions, and found close agreement between CP-corrected 6-31G(d,p) energies and the uncorrected 6-31G++(d,p) energies—the energies were within 1 kcal/mol of each other and also within 1 kcal/mol of the BSSE-free energies.[31]

3. Results

The quantum chemically optimized structures of the various base pairings are depicted in Figure 1, where hydrogen-bonding distances are indicated. Each geometry/tautomer is identified by a Roman numeral. The base pairs, gas-phase optimized in the absence of constraints (using C_1 symmetry), are essentially flat except for configuration V as discussed below. The Watson−Crick base pairs, G−C (I), mG−mC, mT−mA, as well as Cimino-*Z*−Genol-*E* (II) and reverse mT−mA (not in Figure 1), are all planar within several thousandths of an angstrom. Except for a very slightly pyramidal amino group on guanine, the bases of the Cimino-Genol (III) and the reverse Cimino−Genol (VI) configurations are individually very flat but are propeller twisted by a few degrees. The unconstrained reverse wobble mC−mG pair (V) is buckled and twisted and contains pyramidal amino groups, but the energy of the planar constrained (via C_S symmetry) optimization of V ("rev wobble CS mC−mG") is only 0.56 kcal/mol higher in interaction energy. Since this very small energetic penalty would be easily overcome by the helical structure which tends toward planar base pairs, we consider only the planar optimization (CS) to be relevant in what follows, and for comparisons with other configurations and for solvent-phase calculations, the planar-constrained geometry is used. As mentioned, the twisted, reverse C−G pair does not constitute an energetic minimum, and therefore configuration IV could not be optimized. The N1−N9 distances are 8.9 and 9.0 Å for mT−mA and reverse mT−mA base pairs, and those distances were 9.0, 9.1, 9.0, 9.1, and 8.9 Å for I, II, III, V, and VI, respectively. The coordinates of the B3LYP/6-31G(d,p) optimized structures are available in the Supporting Information. For visually comparing the six configurations and T−A base pairs, a figure (Figure S1), based on the optimized coordinates, is also available.

The absolute electronic gas-phase energies of the bases and base pairs optimized at the HF/6-31G(d) and B3LYP/6-31G(d,p) levels of theory, as well as energies using Møller−Plesset theory with the HF geometries [MP2/6-31G(d,p)//HF/6-31G(d)] are listed in Table S1 of the Supporting Information. In Table 1 we present the derived gas-phase base-pair association energies. In Table 2 we present the free energies of solvation for the bases and base pairs. Finally, in Table 3 we present the base-pair association energies in the aqueous phase. In Figure 1 we also present the gas-phase and aqueous-phase energies from Table 1, column 4, and Table 3, column 4, respectively—values that were obtained with the same level of theory and basis set [B3LYP/6-31++G(d,p)//B3LYP/6-31G(d,p)].

Which base pairs are most stable depends on whether they are in the gas phase or the aqueous phase. Inspection of Table 1, column 4, shows that in the gas phase the reverse Cimino−Genol conformation VI is slightly favored over the reverse wobble C−G pair V, by about 1 kcal/mol. The aqueous-phase results, however, indicate that the wobble pair V is favored by about 9 kcal/mol over the minor tautomer pair VI (Table 3, column 4).

The aqueous phase results for the individual bases can be compared with earlier computational studies, as done previously.[37,38] Our results produce relative energies similar to previous results, obtained by a variety of methods. The absolute solvation free energies obtained from our best methodology [CPCM-B3LYP/6-31++G(d,p)//B3LYP/6-31G(d,p)] (Table 2, column 4) agree within 1 kcal/mol of the experimental values

TABLE 1: Gas-Phase Association Energies $\Delta E^{(g)}$ [kcal/mol] Derived from Optimizations of Single Bases, Tautomers, and Base pairs[a]

	optimized HF/6-31G(d)	optimized B3LYP/6-31G(d,p)	B3LYP/6-31++G(d,p)// B3LYP/6-31G(d,p)	MP2/6-31G(d,p)// HF/6-31G(d)
C & G (I)	−25.53 (0.0)	−30.32/−25.54	−26.17 (0.0)	−30.38/−24.39 (0.0)
mC & mG (Ia)	−26.77 (0.0)[b]	−31.45	−27.04 (0.0)[b]	−32.58/−26.42 (0.0)[b]
Cimino-*Z* & Genol-*E* (II)	−7.89 (17.6)	−11.87	−8.11 (18.1)	−13.75/−9.11 (15.3)
Cimino & Genol (III)	−14.39 (11.1)	−21.06	−16.36 (9.8)	−21.51/−15.67 (8.7)
rev wobble CS mC & mG[c] (V)	not wobble[d]	−15.95	−13.59 (13.5)[b]	
rev Cimino & Genol (VI)	−13.16 (12.4)	−19.03/−14.45	−14.31 (12.1)	−19.81/−14.06 (10.3)
mT & mA	−11.78 (15.0)[b]	−16.38/−12.21	−12.75 (14.3)[b]	−17.40/−12.48 (13.9)[b]
rev mT & mA	−11.56 (15.2)[b]	−15.78	−12.25 (14.8)[b]	−17.02/−12.15 (14.3)[b]

[a] In this notation "C & G" means $E_{C-G} - E_C - E_G$, where the isolated monomer energies (E_C, E_G, etc.) are consistently those of the major tautomer forms (C, G, mC, mG, mT, mA), and thus the association energies include the energy to form the minor tautomers. Columns 2 and 3 are unconstrained optimizations. Columns 4 and 5 are single point energy calculations using the B3LYP/6-31G(d,p) and HF/6-31G(d) optimized geometries, respectively. In the column headings, we use the usual double slash (//) notation where the single point method appears left of the optimization method. Bases with an "m" prefix are methylated at N9/N1 (purines/pyrimidines). Relative energies, by column, are given in parentheses, where unmethylated bases are compared to I. [b] Methylated bases are compared to Ia. For some calculations, CP-corrected energies are given after a slash (/). [c] A planar constrained (C_S symmetry) optimization. [d] The HF/6-31G(d) optimization did not yield a wobble geometry, even when starting from the B3LYP/6-31G(d,p) optimized geometry.

TABLE 2: Solvation Free Energies $\Delta\Delta G^{(aq)}$ [kcal/mol] Calculated by COSMO[33] at HF/6-31G(d,p), B3LYP/6-31G(d,p), and B3LYP/6-31++G(d,p) Levels Using HF/6-31G(d) and B3LYP/6-31G(d,p) Optimizations[a]

	CPCM-HF/6-31G(d,p)// HF/6-31G(d)	CPCM-B3LYP/6-31G(d,p)// B3LYP/6-31G(d,p)	CPCM-B3LYP/6-31++G(d,p)// B3LYP/6-31G(d,p)
C	−19.73	−16.81	−19.14
G	−23.92	−21.38	−23.81
Cimino	−15.78	−13.49	−15.77
Genol	−19.12	−17.24	−18.81
mC	−16.96	−14.14	−16.22
mG	−22.12	−19.60	−21.97
mT	−10.24	−8.09	−10.05
mA	−12.20	−11.31	−12.67
C−G (I)	−24.06	−20.55	−23.21
Cimino-*Z*−Genol-*E* (II)	−21.58	−18.49	−20.92
Cimino−Genol (III)	−21.75	−19.76	−22.17
rev wobble CS mC−mG (V)	not wobble[a]	−23.91	−26.78
rev Cimino−Genol (VI)	−21.70	−19.48	−21.88
mT−mA	−11.87	−8.98	−11.16
rev mT−mA	−12.05	−9.24	−11.44

[a] See Table 1.

TABLE 3: Solvent-Phase Association Energies $\Delta E^{(aq)}$ [kcal/mol], with Methods as in Table 2[a]

	CPCM-HF/6-31G(d,p)// HF/6-31G(d)	CPCM-B3LYP/6-31G(d,p)// B3LYP/6-31G(d,p)	CPCM-B3LYP/6-31++G(d,p)// B3LYP/6-31G(d,p)
C & G (I)	−6.14 (0.0)	−12.71 (0.0)	−6.49 (0.0)
mC & mG	−6.20 (−0.1)	−12.90 (−0.2)	−6.47 (0.0)
Cimino-*Z* & Genol-*E* (II)	12.40 (18.5)	7.84 (20.5)	13.85 (20.3)
Cimino & Genol (III)	5.75 (11.9)	−2.61 (10.1)	4.40 (10.9)
rev wobble mC−mG (V)	not wobble	−6.15 (6.6)	−2.18 (4.3)
rev Cimino & Genol (VI)	7.09 (13.2)	−0.32 (12.4)	6.70 (13.2)
mT & mA	−1.26 (4.8)	−6.02 (6.7)	−1.19 (5.3)
rev mT & mA	−1.23 (4.9)	−5.67 (7.0)	−0.97 (5.5)

[a] Relative energies for each column are given in parentheses.

of −13.6 kcal/mol and −9.1 to −12.7 kcal/mol for mA and mT, respectively, values originally reported in a logarithmic plot[39] and later converted to standard temperature.[40] Interestingly, the CPCM model with HF theory at the 6-31G(d,p) basis set agrees with the CPCM model with B3LYP theory at the much larger basis set 6-31++G(d,p). It is also interesting to compare the solvation free energies for the base pairs with those recently obtained for base pairs using the Langevin dipole method.[41] By the latter methodology, solvation free energies of −13.3, −4.0, and −4.1 kcal/mol for (unmethylated) C−G, T−A, and reverse T−A base pairs are considerably reduced in magnitude from our respective values of −23.2, −11.2, and −11.4 kcal/mol (the last two values for mT−mA and reverse mT−mA), yielding relative solvation energies that differ from ours by about 3 kcal/mol. Compared to the experimental solvation free energy (see above), the Langevin dipole method (−10.8 and −12.6 for adenine and thymine, respectively) agrees only to within about 3 kcal/mol.

In considering the issue of basis set superposition error (BSSE), we note that there is fairly close agreement between the gas-phase B3LYP/6-31++G(d,p) and the CP-corrected B3LYP/6-31G(d,p) results (the numbers appearing in column 3 of Table 1 to the right of the slash), indicating that B3LYP/6-31++G(d,p) results contain less BSSE than the B3LYP/6-31G(d,p) results. We infer from this that the CPCM-B3LYP/6-31++G(d,p) results will have less BSSE than the CPCM-B3LYP/6-31G(d,p) results. Coincidently, the CP-corrected MP2/6-31G(d,p)//HF/6-31G(d) energies, the uncorrected HF/6-31G(d)

energies, and the B3LYP/6-31++G(d,p) results all agree within 1 kcal/mol.

We employed methylation at the glycosyl nitrogens to assist in some of the optimizations. An unmethylated reverse-wobble C−G pair optimized [HF/6-31G(d)] to a completely different structure involving (C)N1−H1···O6(G) and (C)O2···H1−N1(G) hydrogen bonds, a structure impossible for full nucleotides since the cytosine H1 would be replaced by C1′. A reverse-wobble-pair optimized geometry V was also not reached for methylated bases by Hartree−Fock theory [HF/6-31G(d)], but was achieved by density functional theory [B3LYP/6-31G(d,p)]. Replacing the hydrogen by a methyl at the glycosyl site has only a small effect on the base-pairing energies; for the standard Watson−Crick pairings of G and C, the methyl increases the pairing stability (i.e., gives more negative association energies) by 1−2 kcal/mol (See "G & C" and "mG & mC" rows of Table 1). The methyl groups change the solvation energies by 1−3 kcal/mol for the isolated bases (see Table 2).

As noted in the Introduction, the symmetry of the thymine−adenine hydrogen bonds provides an obvious route to parallel DNA formation. The optimizations carried out reveal that the Watson−Crick and reverse Watson−Crick mT−mA configurations are very close in energy. The difference between the Watson−Crick pairing and the reverse-Watson−Crick pairing was well under 1 kcal/mol by all methods (cf. "mT & mA" and "rev mT & mA" rows of Table 1), with the Watson−Crick pairing predicted to be slightly more stable by all methods.

4. Discussion

Energies of Association. Although earlier quantum chemical studies predicted that the favorable amino−amino contacts in the reverse WC C−G base pair yield an interaction energy of about −5 kcal/mol (in the gas phase),[8] we have found that the pairing of the minor C−G tautomers VI (i.e., G to enolguanine and C to iminocytosine) and a wobble pair geometry V each yield much stronger interaction energies of about −14 kcal/mol each—binding energies even stronger than the gas-phase stability of a T−A base pair (ca. −13 kcal/mol). This energy, however, is about 12 kcal/mol higher in energy (less favorable) than the conventional WC C−G pairing energy, a result qualitatively similar to a 9 kcal/mol reduction obtained by Zhanpeisov et al. for an enol-isocytosine−keto-cytosine base pair.[21] When considering the minor tautomer pair VI relative to WC C−G and the wobble pair V relative to WC mC-mG, we find the minor tautomer pair VI is slightly favored by 1.4 kcal/mol.

Generally, when hydrogen-bonded paired molecules are solvated (in an aqueous environment), their interaction energy drops considerably due to competition by the solvent (water) for the hydrogen bonds involved in the pairing. This happens for all DNA base pairs, as observed in Table 3: C−G stability drops by a factor of more than 4 and mT−mA stability drops more than 10-fold. The wobble pair V, however, has a lower (more favorable) solvation energy than any of the other pairs considered (see Table 2), likely because the cytosine amino group and the guanine carbonyl group remain exposed to the solvent in the base pair. This yields a reduced destabilization due to solvation of the wobble pair V; it remains mildly favorable in water, more so than mT−mA.

The remarkable stability of the minor tautomer pair VI in the gas phase is partly due to the small cost in energy (ca. 1 kcal/mol each, cf. Table S1) to form the Genol and Cimino tautomers. In the aqueous phase, however, the cost for such a tautomerization is around 4 kcal/mol each (cf. Table 2). Such solvent effects are well-known and even qualitative ordering of tautomer stability can be very different in the gas phase compared to that in the aqueous phase.[42,43] The effect of the solvent, however, is strongly dependent on the compound; in some cases the solvent makes little difference, as observed in the unchanged relative populations of the lactam/lactim tautomers of hydroxyquinolone in water versus benzene.[44] Our results here are comparable to earlier determinations of the tautomeric constants as we discuss below.

An important question is whether a base pair within a DNA helix is better described by gas-phase or aqueous-phase calculations. Both crystallographic and model studies of DNA show that mainly the edges of the bases are solvent exposed while the flat surfaces are mainly stacked against the flanking bases. From a purely electrostatics point of view, this is consistent with the use of the polarizable continuum solvent for just the edges of the bases. Analysis of the solvation free energies of pairs V and VI revealed that they differ mainly in the electrostatic component (10 kcal/mol) and little in the nonelectrostatic component (1 kcal/mol), suggesting that the difference is mainly due to exposed hydrogen-bonding edges, and not water−solute dispersion interactions at the unexposed nonpolar surfaces. This suggests, therefore, that the continuum solvent calculations more closely describes the true DNA environment than the gas phase. The experimental evidence, however, also reveals a very slow exchange of "outside" amino hydrogens.[45−47] The most plausible explanation, then, is that the water structure is changed in the vicinity of major and minor grooves of DNA, relative to the bulk. By correlating the density fluctuations in a 5 ns molecular dynamics simulation with the highly ordered water seen in crystal structures, it was recently shown that the water in the major and minor grooves of B-DNA is twice as ordered as the bulk.[48,49] A worry then is that a PCM solvent alone does not account for non-bulk-solvent effects such as solute−solvent hydrogen bonds. While this remains a concern, it has been shown that at least for weak hydrogen bonds XH−NH3 (X = F, Cl, and Br) the continuum solvent reproduces the solute−solvent interactions quite well.[50] Furthermore, for a few C and G tautomers (including Cimino and Genol) Colominas et al. found the same results to within a few tenths of kcal/mol for a similar PCM method[51] and by Monte Carlo free energy perturbation calculations employing discrete water molecules.[43] The supermolecule approach mentioned in the Introduction, where water molecules are explicitly present in the ab initio optimization, may better account for the solute−solvent interactions, but is limited to a small number of water molecules for which there are multiple minima,[52] and it does not include the dynamical effects of the solvent.

As mentioned above, all base pairs interact less favorably in the aqueous phase due to competition by the solvent for the interbase hydrogen bonds. Comparing Table 1, column 4, with Table 3, column 4 (which have same level of theory and basis set—see also Figure 1), we see that, except for the reverse wobble pair (V), all forms of the C−G base pairs are destabilized by 20−22 kcal/mol in the solvent, relative to the gas phase. The high favorability of the minor tautomer pair VI in the gas phase is greatly reduced in the aqueous phase. Our finding, that in the solvent the WC C−G pair I is favored by 14 kcal/mol, is similar to the 15 kcal/mol aqueous-phase result found by Zhanpeisov et al.[21] who examined the analogous the enol-isocytosine−keto-cytosine base pair, relative to an isocytosine−cytosine base pair by a supermolecular approach. The agreement is remarkable considering that we obtain very different values for the solvation energies of the base pairs.

Importantly, this comparison of the gas-phase and aqueous-phase results shows that the solvent destabilizes the wobble pair by only 11 kcal/mol. This favoring of the wobble pair V over the minor tautomer base pair VI can be understood in terms of the solvent access to two otherwise partially buried amino groups, as in a Watson−Crick base pair I. While we have not considered the effects of pH on the tautomerism[52] and base-pair association energies, we can expect that under acidic conditions the association of VI would be even less favorable due to protonation of N3 in C (major, or amino, tautomer) and loss of the central hydrogen bond.[43]

Energies of Tautomers. Although the focus of this paper is on the nature and geometry of C−G base pairs in parallel-stranded DNA, the accurate determination of tautomeric constants is an important aim by itself and sheds light on the likely accuracy of the proposed minor tautomer base pairs. There have been many theoretical studies of DNA base tautomers, most recently a study by Colominas et al. of the G and C tautomers in the gas phase at a higher level of theory and with larger basis sets than here [MP4/6-311++G(d,p)//MP2/6-31G(d)][43]—see also some 26 references therein. Of course, it would be prohibitive to use such methods for the complete base pairs considered here which contain many more atoms than cytosine or guanine alone. Nevertheless, our B3LYP/6-31G(d,p) optimized tautomerization energies $\Delta E^{(g)}_t$, 0.8 and 1.3 kcal/mol for G → Genol and C → Cimino, respectively, correspond fairly well to the Colominas et al. gas-phase energies of 1.8 and 0.9 kcal/mol, respectively. The paper by Colominas et al. also includes aqueous-phase calculations by the continuum model of Miertus, Scrocco, and Tomasim (MST-HF/6-31G(d)),[51] and by free energy perturbation. In the aqueous phase, the differences $\Delta\Delta G^{(aq)}$ in our free energies of solvation (cf. Table 2, column 4) of 5.0 and 3.4 kcal/mol for Genol and Cimino (relative to G and C), respectively, are similar to the values of 6.2 and 4.5 kcal/mol, respectively, by Colominas et al. Finally, the aqueous-phase tautomerization energies $\Delta G^{(aq)}_t$, calculated from the gas-phase tautomerization energies and the solvation energies $\Delta E_t + \Delta\Delta G^{(aq)}$, yields here 5.8 and 4.7 kcal/mol for G → Genol and C → Cimino, respectively, compared with 8.0 and 6.1 kcal/mol by Colominas et al.

These last quantities determine the tautomeric constant K_T, the equilibrium constant between major and minor tautomers. There have been very few experimental numbers published, owing to the difficulty of detecting scarcely populated species. Early experimental numbers for cytosine were obtained by Kenner et al. based on protonation of 3-methylcytosine versus 1,3-dimethylcytosine, yielding a ratio of amino to imino populations (K_T) of 5×10^4, which at 300 K corresponds to a free energy of 6.4 kcal/mol,[53] in close agreement with 6.1 kcal/mol calculated by Colominas et al. Note that the experimental results are based on the acid constants for *N*-methyl-substituted cytosines and therefore are not identical to the free energy of actually converting the major (amino) tautomer to the minor (imino) tautomer of neutral cytosine.

Recent argon matrix isolation FTIR studies determined an approximate tautomeric constant $K_T = 7.2$ (i.e., $\Delta G = 1.4$ kcal/mol for $T = 348$ K)[54] for the C/Cimino population ratio which corresponds very closely to our gas-phase result of 1.3 kcal/mol. Nevertheless, there is some question whether the experimental procedure yielded an equilibrium distribution of tautomers, so that their relative concentrations may depend on kinetic factors that will complicate their estimation of $\Delta G°$. In fact, we have not adjusted the ab initio energies ΔE(0 K) that we computed by zero point vibrational energies, thermal corrections, and an entropy term to obtain free energies ΔG(298 K) as in Colominas et al., but the values differ from ΔE(0 K) by only a few tenths of kcal/mol. Note that in this study the Cimino tautomer was detected in water, but the tautomeric constant was not determined.[54]

Very recently, an imino tautomer of a related compound, 5-hydroxy-2′-deoxycytidine, was detected by UV resonance Raman spectroscopy, and it was estimated that the imino form is only 3×10^{-3} to 7×10^{-3} as populated as the amino form, and thus is 10^2 to 10^3-fold more prevalent than the imino form (Cimino) of the natural base (C). The highly mutagenic character of OH^5Cyt has been ascribed to the increased prevalence of the imino form,[55] although other explanations have been proposed (see the discussion in Suen et al.[55]).

Wobble Pair Energies. We have argued that the base pairs are best described by the aqueous-phase calculations, and we have seen that in the aqueous phase the wobble pair V is highly favored over the minor tautomer pair VI by about 9 kcal/mol (Table 3, column 4). In these calculations we did not calculate the CP correction, but we expect the B3LYP/6-31++G(d,p) calculations to involve rather little basis set superposition error, based on our results in the gas phase, showing that CP-corrected B3LYP/6-31G(d,p) energies were very close to B3LYP/6-31++G(d,p) energies.

It has been suggested elsewhere, however, that strain induced by the wobble configuration on the DNA backbone may destroy the energetic advantage of the wobble pair.[21] The favorable energetics of base−base stacking,[41,56,57] which can significantly increase the helical stability, may also be reduced by wobble pairing. Simple molecular mechanical minimizations have been carried out to investigate such factors. For a parallel-stranded DNA dodecamer containing only C−G base pairs, Lui et al.[10] found less than a 1 kcal/mol increase for backbone strain, but they found a 4 kcal/mol penalty (increase) in base-stacking energy and 4 kcal/mol in penalty in base-pair electrostatic interactions, per base pair, relative to antiparallel-stranded DNA. The rise in base-pair electrostatic energy should be mainly due to the lack of the third hydrogen bond. Note that these results were obtained from gas-phase minimizations and depend sensitively on the empirically parametrized molecular dynamics force field, which is parametrized for use in the aqueous phase. Nevertheless, if we ignore the base-pair electrostatic term—because it is already included in our ab initio calculation—and sum the other two terms, we arrive at a "helix penalty" on the order of 5 kcal/mol which falls well short of the 9 kcal/mol advantage that the wobble pair has over the minor tautomer pair in the aqueous phase. Based upon these results, we would predict that the wobble pair V should be the predominant form in parallel DNA. Should this prediction not hold true, it would imply that the combined forces of stacking interactions and backbone strain must contribute 9 kcal/mol to the overall energy.

The wobble pair prediction V agrees with that of Mohammadi et al.,[11] who employed FTIR and UV spectroscopy to look at base pairing in antiparallel and parallel DNA. The observed stretching frequency of 1664 cm^{-1} was assigned to both the C6═O6 group of guanine and the C4═O4 group of thymine. From the spectra, it appeared that the guanine O6 is not involved in hydrogen bonding, which is evidence against the minor tautomer model VI but which is consistent with both models IV and V. Furthermore, substitution of guanine by hypoxanthine (essentially guanine without the 2-amino group) disallowed the formation of parallel-stranded DNA, which was interpreted to mean that the amino group is required to hydrogen bond with cytosine in parallel-stranded DNA. While this may suggest that

two hydrogen bonds are required for C−G base pairs, this fact alone rules out neither the minor tautomer stabilization theory VI nor the amino−amino stabilization theory IV.

Two recent studies, although dealing with antiparallel helices, may shed further light on the problem. Robinson et al. have studied by NMR and X-ray crystal structure analysis the related problem (lacking, however, a potential amino−amino interaction), of iG−T base pairs in antiparallel DNA.[14] Both a minor (enol) tautomer form of iG and a wobble pair conformation were observed (at different sites) in the crystal structure. These forms were also seen by NMR, with the wobble pair predominant at 2 °C, exchanging readily at room temperature with the enol form of iG which is populated to about 40% at 40 °C. In a separate X-ray crystallography study, RNA helix G−U base pairs were observed in wobble form.[15] Both studies suggest that the DNA backbone is sufficiently flexible to allow wobble base pairing in antiparallel helices.

In summary, we have argued that the weight of experimental and theoretical data indicates that the C−G wobble pair is formed in the parallel-stranded DNA helix. More generally, this problem is a good illustration of the fine balance between inter- and intramolecular interactions that determine macromolecular structure. The ultimate disentangling of these subtly balanced energies requires both experimentally structures and equilibrium constants and accurate calculations of the individual energy terms.

Acknowledgment. This work has been done at Lawrence Livermore National Laboratory under the auspices of the U.S. Department of Energy, Contract No. W-7405-ENG-48.

Supporting Information Available: Coordinates of all B3LYP/6-31(d,p) optimized base pairs, absolute energies of the bases and base pairs (Table S1), and visual comparison of G−C and T−A base pairs (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Pattabiraman, N. *Biopolymers* **1986**, *25*, 1603.

(2) Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984.

(3) van de Sande, J. H.; Ramsing, N. B.; Germann, M. W.; Elhorst, W.; Kalisch, B. W.; Kitzing, E. v.; Pon, R. T.; Clegg, R. C., Jovin, T. M. *Science* **1988**, *241*, 551.

(4) Ramsing, N. B.; Jovin, T. M. *Nucleic Acids Res.* **1988**, *16*, 6659.

(5) Otto, C.; Thomas, G. A.; Rippe, K.; Jovin, T. M.; Peticolas, W. L. *Biochemistry* **1991**, *30*, 3062.

(6) Rippe, K.; Ramsing, N. B.; Klement, R.; Jovin, T. M. *J. Biomol. Struct. Dyn.* **1990**, *7*, 1199.

(7) Allawi, H. T.; SantaLucia, Jr., J. *Biochemistry* **1998**, *37*, 2170.

(8) Šponer, J.; Hobza, P. *J. Biomol. Struct. Dyn.* **1994**, *12*, 671.

(9) Rippe, K.; Kuryavyi, V. V.; Westhof, E.; Jovin, T. M. Polymorphism and possible biological functions of parallel-stranded DNA. In *Structural Tools for the Analysis of Protein-Nucleic Acid Complexes*; Lilley, D. M. J., Heumann, H., Suck, D., Eds.; Birkhäuser Verlag: Berlin, 1992; Chapter 6, p 81.

(10) Liu, C.-Q.; Shi, X.-F.; Bai, C.; Zhao, J., Wang, Y. *J. Theor. Biol.* **1997**, *184*, 319.

(11) Mohammadi, S.; Klement, R.; Shchyolkina, A. K.; Liquier, J.; Jovin, T. M.; Taillandier, E. *Biochemistry* **1998**, *37*, 16529.

(12) Watson, J. D.; Crick, F. H. *Nature* **1953**, *171*, 964.

(13) Topal, M. D.; Fresco, J. R. *Nature* **1976**, *263*, 285.

(14) Robinson, H.; Gao, Y.-G.; Bauer, C.; Roberts, C.; Switzer, C.; Wang, A. H.-J. *Biochemistry* **1998**, *37*, 10897.

(15) Shi, K.; Wahl, M.; Sundaralingam, M. *Nucleic Acids Res.* **1999**, *27*, 2196.

(16) Fresco, J. R.; Alberts, B. M. *Proteins: Struct. Funct. Genet.* **1960**, *46*, 311.

(17) Aboul-ela, F.; Koh, D.; Tinoco, Jr., I. *Nucleic Acids Res.* **1985**, *13*, 4811.

(18) Peyret, N.; Seneviratne, P. A.; Allawi, H. T.; SantaLucia. *Biochemistry* **1999**, *38*, 3468.

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

(20) Gao, X.; Huang, X.; Smith, G. K.; Zheng, M.; Liu, H. *J. Am. Chem. Soc.* **1995**, *117*, 8883.

(21) Zhanpeisov, N. U.; Šponer, J.; Leszczynski, J. *J. Phys. Chem. A* **1998**, *102*, 10374.

(22) Zhanpeisov, N. U.; Leszczynski, J. *J. Phys. Chem. B* **1998**, *102*, 9109.

(23) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, Jr., J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, Revision A.4; Gaussian, Inc.: Pittsburgh, PA, 1998.

(24) Hehre, W. J.; Radom, L.; Schleyer, P. v.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; John Wiley and Sons: New York, 1986.

(25) Becke, A. D. *J. Chem. Phys.* **1993**, *90*, 5648.

(26) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(27) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

(28) Hobza, P.; Šponer, J. *Chem. Rev.* **1999**, *99*, 3247.

(29) Liedl, K. R. *J. Chem. Phys.* **1997**, *108*, 3199.

(30) Rablen, P. R.; Lockman, J. W.; Jorgensen, W. L. *J. Phys. Chem. A* **1998**, *102*, 3782.

(31) Fellers, R.; Barsky, D.; Gigi, F.; Colvin, M. E. *Chem. Phys. Lett.* **1999**, *312*, 548.

(32) Gould, I. R.; Kollman, P. A. *J. Am. Chem. Soc.* **1994**, *116*, 2493.

(33) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995.

(34) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717.

(35) Birnstock, F.; Hofmann, H. J.; Kohler, H. J. *Theor. Chem. Acta* **1976**, *42*, 311.

(36) Mennucci, B.; Cossi, M.; Tomasi, J. *J. Phys. Chem.* **1996**, *100*, 1807.

(37) Miller, J. L.; Kollman, P. A. *J. Phys. Chem.* **1996**, *100*, 8587.

(38) Barsky, D.; Kool, E. T.; Colvin, M. E. *J. Biomol. Struct. Dyn.* **1999**, *16*, 1119.

(39) Clark, L. B.; Peschel, G. G.; Tinoco, Jr., I. *J. Phys. Chem.* **1965**, *69*, 3615.

(40) Ferguson, D. M.; Pearlman, D. A.; Swope, W. C.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 362.

(41) Florian, J.; Šponer, J.; Warshel, A. *J. Phys. Chem. B* **1999**, *103*, 884.

(42) Kobayashi, R. *J. Phys. Chem. A* **1998**, *102*, 10813.

(43) Colominas, C.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **1996**, *118*, 6811.

(44) Tucker, Jr., G. F.; Irvin, J. L. *J. Am. Chem. Soc.* **1951**, *73*, 1923.

(45) Englander, J. J.; v. Hippel, P. H. *J. Mol. Biol.* **1972**, *63*, 171.

(46) Teitelbaum, H.; Englander, S. W. *J. Mol. Biol.* **1975**, *92*, 55.

(47) Michalczyk, R.; Russu, I. M. *Biophys. J.* **1999**, *76*, 2679.

(48) Young, M. A.; Jayaram, B.; Beveridge, D. L. *J. Am. Chem. Soc.* **1997**, *119*, 59.

(49) Young, M. A.; Ravishanker, G.; Beveridge, D. L. *Biophys. J.* **1997**, *73*, 2313.

(50) Abkowicz-Bienko, A.; Biczysko, M.; Latajka, Z. *Comput. Chem.* **2000**, *24*, 303.

(51) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.

(52) Montero, L. A.; Esteva, A. M.; Molina, J.; Zapardiel, A.; Hernández, L.; Márquez, H.; Acosta, A. *J. Am. Chem. Soc.* **1998**, *120*, 12023.

(53) Kenner, G. W.; Reese, C. B.; Todd, A. R. *J. Chem. Soc.* **1955**, 855.

(54) Smets, J.; Adamowicz, L.; Maes, G. *J. Phys. Chem.* **1996**, *100*, 6434.

(55) Suen, W.; Spiro, T. G.; Sowers, L. C.; Fresco, J. R. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4500.

(56) Šponer, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem. A* **1997**, *101*, 9489.

(57) Alhambra, C.; Luque, F. J.; Gago, F.; Orozco, M. *J. Phys. Chem. B* **1997**, *101*, 3846.