J|A|C|S
ARTICLES
Published on Web 01/05/2007

Does the A·T or G·C Base-Pair Possess Enhanced Stability? Quantifying the Effects of CH···O Interactions and Secondary Interactions on Base-Pair Stability Using a Phenomenological Analysis and ab Initio Calculations

Jordan R. Quinn,[†] Steven C. Zimmerman,*[,†] Janet E. Del Bene,[‡] and Isaiah Shavitt[†]

Contribution from the Department of Chemistry, Roger Adams Laboratory, University of Illinois, 600 South Mathews Avenue, Urbana, Illinois 61801, and the Department of Chemistry, Youngstown State University, Youngstown, Ohio 44555

Received August 31, 2006; E-mail: sczimmer@uiuc.edu

Abstract: An empirically based relationship between overall complex stability ($-\Delta G°$) and various possible component interactions is developed to probe the question of whether the A·T/U and G·C base-pairs exhibit enhanced stability relative to similarly hydrogen-bonded complexes. This phenomenological approach suggests ca. 2−2.5 kcal mol^{-1} in additional stability for A·T owing to a group interaction containing a CH···O contact. Pairing geometry and the role of the CH···O interaction in the A·T base-pair were also probed using MP2/6-31+G(d,p) calculations and a double mutant cycle. The ab initio studies indicated that Hoogsteen geometry is preferred over Watson−Crick geometry in A·T by ca. 1 kcal mol^{-1}. Factors that might contribute to the preference for Hoogsteen geometry are a shorter CH···O contact, a favorable alignment of dipoles, and greater distances between secondary repulsive sites. The CH···O interaction was also investigated in model complexes of adenine with ketene and isocyanic acid. The ab initio calculations support the result of the phenomenological approach that the A·T base-pair does have enhanced stability relative to hydrogen-bonded complexes with just N−H···N and N−H···O hydrogen bonds.

Introduction

The molecular-level genetic storage units, A·T and G·C base-pairs, are clearly special. For example, it is well documented that a single tautomeric form predominates in each of the four bases, disfavoring mismatches. What is less well-known is whether the strength of base pairing is enhanced by interactions in addition to the strong N−H···N and N−H···O hydrogen bonds that are present. Early measurements of association constants (K_{assoc}) of base-pair analogues in chloroform by Rich and co-workers showed that A·T/U and G·C base-pairs are more stable than mismatched complexes or base dimers.[1,2] The concept of a special *electronic complementarity* was most clearly supported by the observation that the K_{assoc} for the A·U base-pair was significantly higher than the K_{dimer} for A·A and U·U, despite all three pairs containing two hydrogen bonds. Jorgensen attributed the greater stability of the A·T base-pair to differences in the strength of primary hydrogen bonds (acidity/basicity) and to secondary electrostatic interactions between proximal hydrogen-bonding sites.[3] Rebek,[4a] Hunter,[4b] and others have suggested a CH···O hydrogen bond between H-8 of adenine and O-2 of thymine.

In addition to the studies noted above, many theoretical studies in the literature have reported the results of electronic structure calculations on the A·T base-pair. The majority of these studies employed density-functional methods[5] or low-level ab initio calculations,[6] neither of which is generally reliable for the treatment of hydrogen bonding.[7,8] In particular, including diffuse basis functions on non-hydrogen atoms and polarization functions on hydrogen atoms and accounting for electron correlation are essential for the proper description of hydrogen bonds.

A theoretical study of DNA base-pairs by Sponer, Leszczynski, and Hobza[9] reported optimized HF/6-31G(d,p) geom-

[†] University of Illinois.
[‡] Youngstown State University.

(1) Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984; Chapter 6.
(2) (a) Kyogoku, Y.; Lord, R. C.; Rich, A. *Proc. Natl. Acad. Sci. U.S.A.* **1967**, *57*, 250−257. (b) Kyogoku, Y.; Lord, R. C.; Rich, A. *Biochim. Biophys. Acta* **1969**, *179*, 10−17.
(3) (a) Jorgensen, W. L.; Pranata, J. *J. Am. Chem. Soc.* **1990**, *112*, 2008−2010. (b) Pranata, J.; Wierschke, S. G.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1991**, *113*, 2810−2819. (c) Jorgensen, W. L. *Chemtracts Org. Chem.* **1991**, *4*, 91−119.
(4) (a) Jeong, K. S.; Tjivikua, T.; Muehldorf, A.; Deslongchamps, G.; Famulok, M.; Rebek, J., Jr. *J. Am. Chem. Soc.* **1991**, *113*, 201−209. (b) Leonard, G. A.; McAuley-Hecht, K.; Brown, T.; Hunter, W. N. *Acta Crystallogr., Sect. D* **1995**, *51*, 136−139.
(5) (a) Fonseca Guerra, C.; Bickelhaupt, F. M.; Snijders, J. G.; Baerends, E. J. *Chem.—Eur. J.* **1999**, *5*, 3581−3594. (b) Fonseca Guerra, C.; Bickelhaupt, F. M.; Snijders, J. G.; Baerends, E. J. *J. Am. Chem. Soc.* **2000**, *122*, 4117−4128. (c) Richardson, N. A.; Wesolowski, S. S.; Schaefer, H. F., III. *J. Phys. Chem. B* **2003**, *107*, 848−853. (d) Kumar, A.; Knapp-Mohammady, M.; Mishra, P. C.; Suhai, S. *J. Comput. Chem.* **2004**, *25*, 1047−1059.
(6) Ohta, Y.; Tanaka, H.; Baba, Y.; Kagemoto, A.; Nishimoto, K. *J. Phys. Chem.* **1986**, *90*, 4438−4442. Gould, I. R.; Kollman, P. A. *J. Am. Chem. Soc.* **1994**, *116*, 2493−2499. Spirko, V.; Sponer, J.; Hobza, P. *J. Chem. Phys.* **1997**, *106*, 1472−1479.
(7) Del Bene, J. E. Hydrogen Bonding 1. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; Vol. 2, pp 1263−1271.
(8) Del Bene, J. E.; Jordan, M. J. T. *THEOCHEM* **2001**, *573*, 11−23.

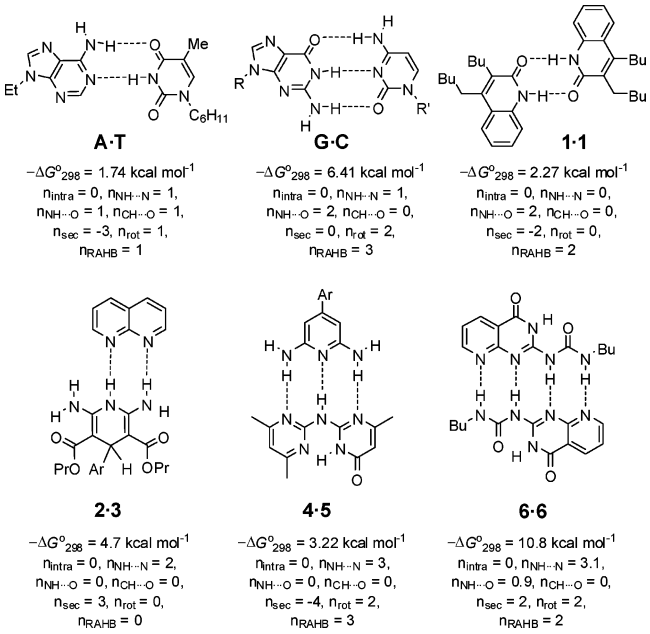
10.1021/ja066341f CCC: $37.00

etries and single-point MP2/6-31G(d,p) energies for the optimized structures and also provided optimized structures and energies obtained using the B3LYP density-functional model with the same basis set. Similar calculations by Kabeláč and Hobza[10] that included diffuse d polarization functions found unusual energy orderings between different structures, with structures other than WC and H having the lowest energies. Another ab initio study of the A·T base-pair was reported by Kryachko and Sabin,[11] who explored the energy surface of A·T for various stationary points, including transition states and local minima. Their structure optimizations were carried out using the B3LYP density-functional model and Hartree−Fock calculations with the 6-31+G(d) basis set, followed by single-point energy calculations for the optimized structures at the correlated MP2 level with the same basis set. However, their claim to have calculated the energy difference between the Watson−Crick and Hoogsteen conformers of the complex is incorrect, because their reported Hoogsteen structure actually is reverse-Watson−Crick. Later calculations of interaction energies of base-pairs and their methylated derivatives were carried out by Jurečka and Hobza[12] and by Šponer, Jurečka, and Hobza,[13] using an approximate MP2 (RI-MP2) method with a cc-pVTZ basis plus CCSD(T) corrections based on a smaller basis set. Their results were summarized recently by Jurečka et al.[14]

Herein, we probe for evidence of enhanced stability in A·T and G·C base-pairs using two different approaches. The first involves developing an empirically based relationship between overall complex stability and various possible component interactions. This phenomenological approach using fixed energetic increments for each identifiable contributor to complex stability is an extension of the approach first used by Jorgensen[3] and subsequently tested and used by others.[15] (It should be noted that despite its intuitive appeal and striking correlation with the experimental data, as seen in the current study, the Jorgensen secondary electrostatic interactions model has been criticized.[16]) The objective of the phenomenological analysis is to establish a correlation between assignable increments and complexation free energy.

The second approach uses ab initio calculations at the MP2/6-31+G(d,p) level to obtain optimized structures (constrained to C_s symmetry) and energies for the A·T pair in the Watson−Crick (WC), reverse-Watson−Crick (r-WC), Hoogsteen (H), and reverse-Hoogsteen (r-H) conformations, to perform theoretical double-mutant cycles for WC and H, and to obtain interaction

Chart 1



energies for model ketene·adenine and HCNO·adenine complexes in WC and H conformations. Unconstrained (C_1) optimized MP2 structures, binding energies, enthalpies, and free energies are reported for the smaller model A·U pair, using a mixed 6-31+G(d,p) and 6-31G(d) basis set. The objectives of the ab initio studies are (1) to compare the binding energies of the WC, H, r-WC, and r-H forms of the A·T base-pair; (2) to use a theoretical double-mutant cycle for WC and H in an attempt to quantify the effect of an interaction between a thymine carbonyl oxygen and an adenine CH; (3) to obtain interaction energies for model ketene·adenine and HCNO·adenine complexes in an attempt to elucidate the role of the CH···O interaction in the binding; and (4) to obtain structural data for these complexes.

In comparing the ab initio and empirical data it should be noted that the ab initio calculations pertain to gas-phase complexes, while the empirical analysis uses experimental solution data.

Methods

Empirical Model. The phenomenological analysis uses a dataset containing 256 hydrogen-bonded complexes with reported association constants (K_{assoc}) in chloroform of at least 1 M^{-1}. Some closely resemble natural DNA (RNA) base-pairs, others less so, but all compounds contain an approximately linear array of hydrogen-bond donor and acceptor groups. Clefts and macrocycles and related host−guest systems were not included, nor were complexes measured in other solvents such as several substituted pyridone dimers. With the exception of a handful of complexes excluded for specific reasons that are justified in the Supporting Information, the dataset is believed to be comprehensive. The complexes contain between two and six primary hydrogen bonds and the free energies of complexation ($-\Delta G°_{298}$) range from ca. −0.8 to 10.9 kcal mol^{-1}. Representative examples are shown in Chart 1. For a more detailed discussion of the criteria used in selecting complexes and a complete listing of their structures see the Supporting Information. A combinatorial, multivariate linear regression analysis[17,18]

(9) Šponer, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem.* **1996**, *100*, 1965−1974.
(10) Kabeláč, M.; Hobza, P. *J. Phys. Chem. B* **2001**, *105*, 5804−5817.
(11) Kryachko, E. S.; Sabin, J. R. *Int. J. Quantum Chem.* **2003**, *91*, 695−710.
(12) Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608−15613.
(13) Šponer, J.; Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142−10151.
(14) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985−1993.
(15) Selected examples: (a) Murray, T. J.; Zimmerman, S. C. *J. Am. Chem. Soc.* **1992**, *114*, 4010−4011. (b) Sartorius, J.; Schneider, H.-J. *Chem.−Eur. J.* **1996**, *2*, 1446−1452. (c) Beijer, F. H.; Sijbesma, R. P.; Kooijman, H.; Spek, A. L.; Meijer, E. W. *J. Am. Chem. Soc.* **1998**, *120*, 6761−6769. (d) Lüning, U.; Kühl, C. *Tetrahedron Lett.* **1998**, *39*, 5735−5738. (e) Zimmerman, S. C.; Corbin, P. S. *Struct. Bonding* **2000**, *96*, 63−94. (f) Alvarez-Rua, C.; García-Granda, S.; Goswami, S.; Mukherjee, R.; Dey, S.; Claramunt, R. M.; Santa María, M. D.; Rozas, I.; Jagerovic, N.; Alkorta, I.; Elguero, J. *New J. Chem.* **2004**, *28*, 700−707.
(16) (a) Lukin, O.; Leszczynski, J. *J. Phys. Chem. A* **2002**, *106*, 6775−6782. (b) Popelier, P. L. A.; Joubert, L. *J. Am. Chem. Soc.* **2002**, *124*, 8725−8729. (c) See however: Uchimaru, T.; Korchowiec, J.; Tsuzuki, S.; Matsumura, K.; Kawahara, S. *Chem. Phys. Lett.* **2000**, *318*, 203−209.
(17) Hair, J. F.; Anderson, R. E.; Tatham, R. L.; Black, W. C. *Multivariate Data Analysis*; Prentice Hall: New York, 1998.
(18) See Supporting Information for details.

Table 1. Variables Used in Multivariate Regression Analysis

symbol	variable
n_{intra}	intramolecular H-bonds broken before complexation
$n_{NH\cdots O}$	intermolecular H-bonds formed between NH and C═O groups
$n_{NH\cdots N}$	intermolecular H-bonds formed between NH and −N═ groups
$n_{CH\cdots O}$	intermolecular interactions between CH and C═O groups
n_{sec}	net secondary electrostatic interactions (attractive minus repulsive)
n_{RAHB}	resonance assisted H-bonds (cooperative H-bonds)[1,5a,19]
n_{rot}	number of unconstrained single bonds frozen in complex

was carried out using all 256 experimental $-\Delta G°_{298}$ values, the components (n_x) listed in Table 1, and equations of the general form

$$\Delta G°_{298} = \text{constant} + an_a + bn_b + ... \quad (1)$$

The n_{intra} variable is used when an intramolecular hydrogen bond must be broken in one of the components prior to complexation (e.g., pyridylureas). The all-possible-subsets regression was carried out using Microsoft Excel. Collinearity data and statistics for the best fits are given in the Supporting Information.

Ab Initio Methods. Ab initio calculations on the A·T pair were carried out at the correlated MP2 level, with the 6-31+G(d,p) basis set,[20] using the Gaussian 03 program.[21] Because of the size of the calculations and the difficulties in optimizing geometrical parameters that involve very shallow energy profiles, the optimized structures of W, H, r-W, and r-H were constrained to have C_s symmetry. The corresponding A and T monomers were also optimized with C_s symmetry, but additional optimizations of the monomers with no symmetry were carried out to assess the effect of the symmetry constraints. In an attempt to assess the magnitude of the CH···O interaction in WC and H, structures for a double-mutant cycle were calculated at the same level. In addition, single-point MP2/6-31+G-(d,p) calculations were performed on complexes of adenine with ketene (H_2C═CO) and with isocyanic acid (HN═CO). In these cases, the ketene or HCNO molecule was placed so that the distance from the H of the ring to the oxygen of the C═O group and the corresponding C−H···O and H···O═C angles were the same as they are in the WC and H conformers of A·T.

To obtain optimized structures of the A·T complexes without symmetry constraints and to assess the energy effects of these constraints, as well as to determine vibrational and thermal contributions to the free energy of binding, optimized C_1 structures were obtained for model WC and H adenine−uracil (A·U) pairs and for the corresponding U and A monomers. For these calculations, the 6-31+G-(d,p) basis set was placed on those atoms involved in the formation of hydrogen bonds in the complexes, as well as the O and C−H atoms involved in the CH···O interaction. The 6-31G(d) basis was used on all other atoms. This mixed 6-31+G(d,p)/6-31G(d) basis will be referred to as the reduced basis set. All of the calculations reported in this paper were performed on the Cray X1 at the Ohio Supercomputer Center.

Results and Discussion

Empirical Model. In a preliminary analysis that followed Jorgensen's method,[3] better fits were obtained by using different variables for the two types of primary hydrogen bonds (NH···O and NH···N), whereas using a single variable for all the secondary electrostatic components worked as well as the three-component method (repulsive H···H, repulsive N/O···N/O, attractive H···N/O).[3c] Thus the single-variable method was chosen for the secondary interactions in the full analysis, as indicated in Table 1. The inclusion of a constant in eq 1 was necessary for good fits. In addition to collecting statistical errors,[17] the intercept represents the loss in translational and rotational entropy when the two components are held together ($\Delta G°_{trans,rot}$). Although its magnitude (5.6 kcal mol^{-1}) is lower than the value of 7−11 kcal mol^{-1} generally agreed upon as representing the loss in energy when two components are held rigidly fixed,[22] residual motion in complexes will likely lower the value.[23] In any event, this magnitude is certainly within the range of related constants determined in regression analyses of ligand/drug−receptor and enzyme−inhibitor complexes,[24] even if its significance, beyond providing predictive value, is unclear.

Of the 127 regression analyses performed with different choices of variables, four gave $R^2 > 0.9$, each minimally containing n_{intra}, $n_{NH\cdots O}$, $n_{NH\cdots N}$, $n_{CH\cdots O}$, and n_{sec}. Equation 2 fit the data well with an adjusted $R^2 = 0.905$ and P values for the intercept and variables that were all $<10^{-6}$, indicating a high degree of statistical significance. (In this equation n_{sec} is the number of attractive secondary interactions minus the number of repulsive interactions.)

$$\Delta G°_{298} = 5.6 + 3.2n_{intra} - 3.5n_{NH\cdots N} - 4.1n_{NH\cdots O} - 0.7n_{sec} - 2.2n_{CH\cdots O} \quad (2)$$

Addition of the terms n_{RAHB} and n_{rot} alone or together minimally increased the adjusted R^2.[18] That a term for RAHB[19] is not needed is consistent with the conclusions of previous studies of RAHBs that suggest that it is the σ-skeleton of an unsaturated system that is responsible for the increased stability of intramolecular hydrogen bonds.[25] By contrast the best fit of the data to a simple two-increment model,[3,15] using just the total numbers of primary and secondary hydrogen bonds, gave an adjusted $R^2 = 0.554$. Even allowing for a best-fit, nonzero intercept (2.8 kcal mol^{-1}) and excluding those complexes that contain intramolecular hydrogen bonds, the adjusted R^2 rose to only 0.750.[18]

There are several assumptions made in this analysis. One is that the heterocyclic bases adopt the tautomeric forms shown in their complexes.[18] Furthermore, when nondegenerate complexes containing the same hydrogen-bonding motif are possible, the K_{assoc} values were statistically corrected, although direct evidence for multiple-complex modes is unavailable. It is also clear that structurally similar complexes can exhibit significant differences in their K_{assoc} values because of differences in the pK_a values of the donor and acceptor sites as a result of substituent effects. Even identical complexes measured in

(19) Gilli, G.; Bellucci, F.; Ferretti, V.; Bertolasi, V. *J. Am. Chem. Soc.* **1989**, *111*, 1023.
(20) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys*. **1982**, *56*, 2257−2261. Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213−222. Spitznagel, G. W.; Clark, T.; Chandrasekhar, J.; Schleyer, P. v. R. *J. Comput. Chem.* **1982**, *3*, 363−371. Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem*. **1983**, *4*, 294−301.
(21) Frisch, M. J.; et al. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
(22) (a) Page, M. I; Jencks, W. P. *Proc. Natl. Acad. Sci. U.S.A.* **1971**, *68*, 1678−1683. (b) Ajay; Murcko, M. A. *J. Med. Chem.* **1995**, *38*, 4953−4967.
(23) (a) Searle, M. S.; Williams, D. H. *J. Am. Chem. Soc.* **1992**, *114*, 10690−10697. (b) Searle, M. S.; Williams, D. H.; Gerhard, U. *J. Am. Chem. Soc.* **1992**, *114*, 10697−10704.
(24) (a) Horton, N.; Lewis, M. *Protein Sci.* **1992**, *1*, 169−181. (b) Krystek, S.; Stouch, T.; Novotny, J. *J. Mol. Biol.* **1993**, *234*, 661−679. (c) Böhm, H.-J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256. (d) Head, R. D.; Smythe, M. L; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.
(25) (a) Alkorta, I.; Elguero, J.; Mó, O.; Yáñez, M.; Del Bene, J. E. *Mol. Phys.* **2004**, *402*, 2563. (b) Alkorta, I.; Elguero, J.; Mó, O.; Yáñez, M.; Del Bene, J. E. *Chem. Phys. Lett.* **2005**, *411*, 411.

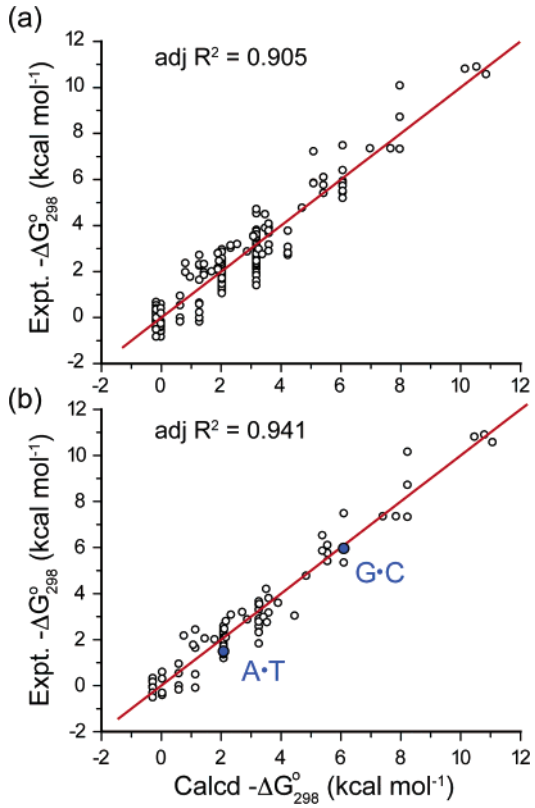


Figure 1. Exptl vs calcd $-\Delta G°_{298}$ using eq 2: (a) 256 separate and (b) 86 grouped complexes.

Table 2. Comparison of the Experimental and Predicted Values of $-\Delta G°_{298}$ for the A·T/U and G·C Base-Pairs (kcal mol^{-1})[a]

	predicted	expt (average)	expt (range)
A·U	2.2	1.83	1.666−1.962
A·T	2.2	1.63	1.085−2.061
G·C	6.1	5.95	5.454−6.407

[a] See the Supporting Information for the experimental data.

different laboratories may give different K_{assoc} values. For this reason, the 256 complexes were grouped into 86 structural classes. In addition to averaging out some of the variabilities discussed above, this approach avoids over-weighting complex types with multiple entries. Shown in Figure 1a is a plot of $-\Delta G°_{298}$(exptl) vs $-\Delta G°_{298}$(calcd) using eq 2. The grouping of complexes gives, within experimental error, the same equation, but an improved adjusted R^2 value (Figure 1b).

As seen in Table 2, the A·T/U and G·C experimental $-\Delta G°_{298}$ values fit the predicted values from eq 2 within experimental error (standard error 0.65 kcal mol^{-1}). But it should be noted that in the case of A·T/U the K_{assoc} was corrected by a factor of 4, assuming equal portions of both normal and reverse Watson−Crick and Hoogsteen complexes. Early work on the A·U pair in chloroform-*d* indicated ca. 70:30 ratio of Hoogsteen to Watson−Crick forms,[26a] whereas more recent studies in $CDClF_2/CDF_3$ suggest stronger bonding in the Watson−Crick arrangement.[26b] A distinction between normal and reverse forms was not evident in either case.

Ab Initio Results. The electronic energies and binding energies obtained in this work for the symmetry-constrained optimizations at the MP2 level with the full 6-31+G(d,p) basis set are shown in Table 3. This table lists the total electronic energies (in hartree atomic units) of the monomers and the four conformers of the A·T complex, the binding energies (in kcal mol^{-1}) of the complexes, as well as the energy differences between the WC and H structures. Additionally, estimates of zero-point vibrational contributions to the binding energies and thermal contributions to the enthalpy and free energy of binding, obtained from the unconstrained optimizations of the model adenine−uracil WC and H complexes with the reduced basis set, are included in Table 3.

WC versus H. The results in Table 3 show the Hoogsteen structure to be favored over WC by 1.14 kcal mol^{-1}. It is interesting to note that the zero-point and thermal energies obtained for the model A·U pairs that were used to estimate binding enthalpies and free energies do not affect this difference significantly. From Table 3 it can be seen that the reverse forms of WC and H (r-WC and r-H) have slightly higher electronic energies than WC and H, respectively. The reverse structures are not considered further in this work.

One factor that may be responsible for the greater binding energy of H compared to WC is the relative alignment of the dipole moment vectors of the monomers in these two forms of the complex, shown in Figure 2. It is seen that this alignment in H is not far from the favorable head-to-tail arrangement, compared to opposing dipoles in WC. Using the calculated dipole moments of thymine, 4.507 D, and adenine, 2.813 D, and their orientations, the dipole−dipole interaction energy is computed approximately as −1.0 kcal mol^{-1} in H and +1.6 kcal mol^{-1} in WC. However, the point-dipole approximation breaks down when the monomers are in close contact and an accurate assessment of the electrostatic interaction energy would require consideration of the detailed charge distributions of the two monomers.

Another factor may involve the secondary interactions[3] between the two primary hydrogen bonds. In adenine, the greater distance of the amine group from N-7 than from N-1 results in a greater distance between the NH···O and NH···N hydrogen bonds in H compared to WC, and thus reduces the magnitude of the unfavorable secondary interactions in the former relative to the latter.

C−H···O Interaction. There have been various assessments of the role of CH···O hydrogen bonds in base-pairs and other complexes.[4b,5a,27,28] In some computational studies the contribution of the CH···O interaction to the overall A·T base-pair stability has ranged from negligible[5a,29a] to ca. 6% of the total interpair bond energy.[29b] For example, Scheiner et al. [27e] carried out MP2/6-31+G(d,p) calculations of CH···O hydrogen-bond energies between water and a number of amino acids and found these energies to be in the range 1.9−2.5 kcal mol^{-1}, with a preferred C−O distance of 3.32 Å. However, the CH···O

(26) (a) Iwahashi, H.; Sugeta, H.; Kyogoku, Y. *Biochemistry* **1982**, *21*, 631−638. (b) Dunger, A.; Limbach, H.-H.; Weisz, K. *J. Am. Chem. Soc.* **2000**, *122*, 10109−10114.

(27) (a) Gu, Y.; Kar, T.; Scheiner, S. *J. Am. Chem. Soc.* **1999**, *121*, 9411−9422. (b) Scheiner, S. *Adv. Mol. Struc. Res.* **2000**, *6*, 159−207. (c) Scheiner, S.; Gu, Y.; Kar, T. *THEOCHEM* **2000**, *500*, 441−452. (d) Gu, Y.; Kar, T.; Scheiner, S. *J. Mol. Struc.* **2000**, *552*, 17−31. (e) Scheiner, S.; Kar, T.; Gu, Y. *J. Biol. Chem.* **2001**, *276*, 9832−9837. (f) Scheiner, S.; Grabowski, S. T.; Kar, T. *J. Phys Chem. A* **2001**, *105*, 10607−10612.

(28) Brandl, M.; Meyer M.; Suhnel, J. *J. Biomol. Struct. Dyn.* **2001**, *18*, 545−555. Hartmann, M.; Wetmore, S. D.; Radom, L. *J. Phys. Chem. A* **2001**, *105*, 4470−4479.

(29) (a) Asensio, A.; Kobko, N.; Dannenberg, J. J. *J. Phys. Chem. A* **2003**, *107*, 6441−6443. (b) Starikov, E. B.; Steiner, T. *Acta Crystallogr., Sect. D* **1997**, *53*, 345−347.

Table 3. MP2/6-31+G(d,p) Energies for the A·T Monomers and Complexes in C_s Symmetry[a]

	au (E_h)	kcal mol^{-1}				
	E	ΔE	ΔZPE	ΔH (0 K)	ΔH (298.15 K)	ΔG (298.15 K)
thymine	−452.88465					
adenine	−466.02321					
WC	−918.93337	−16.01	(0.59)	(−15.42)	(−14.89)	(−4.82)
H	−918.93519	−17.15	(0.57)	(−16.58)	(−16.02)	(−6.01)
H−WC	−0.00182	−1.14	(−0.02)	(−1.16)	(−1.13)	(−1.19)
r-WC	−918.93303	−15.79				
r-H	−918.93489	−16.96				

[a] The data in parentheses for zero-point energy and thermal contributions are approximate, taken from the reduced-basis calculations for the A·U complex.

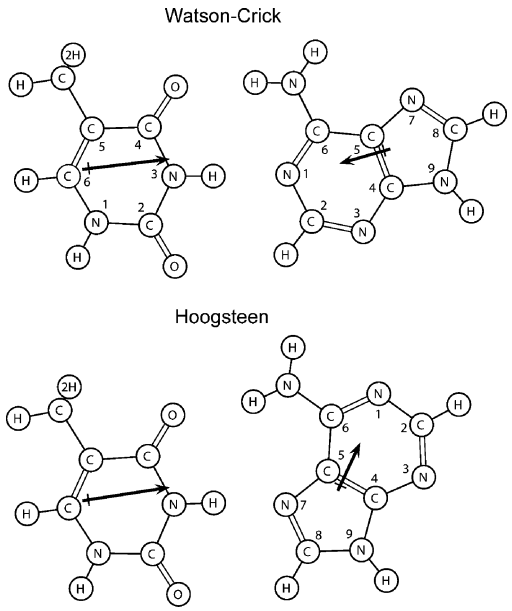
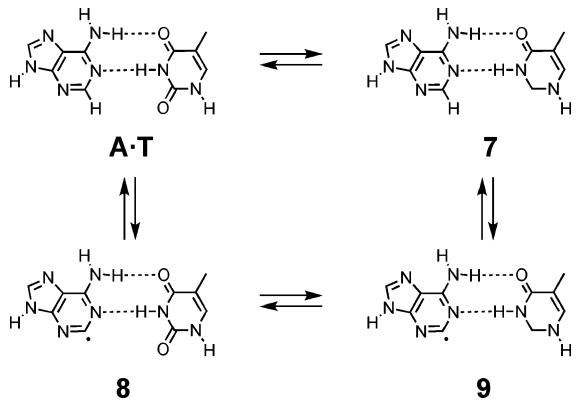


Figure 2. Dipole moment vectors of the monomers positioned as in the Watson−Crick (top) and Hoogsteen (bottom) A·T complexes.

Scheme 1



configuration in A·T is highly nonlinear and involves a long C−O distance, signaling a weak interaction that is not likely to represent a true hydrogen bond. The optimized structural parameters of the three hydrogen bonds in each of the two complexes are shown in Table 4. Although the overall length (the distance between C and O) of the distorted CH···O "hydrogen bond" in H is not very much longer than the optimal length found by Scheiner et al.,[27e] the H···O distance in it is much greater than their optimal value of ∼2.24 Å, owing to the nonlinearity. The C−O distance in CH···O is significantly shorter in H than in WC, which may be a contributing factor to the greater binding energy of the former. (The shortening of the CH···O bond in H relative to WC is accomplished by a slight tilting of the two monomers relative to each other, within the plane of the rings, slightly lengthening the NH···O hydrogen bond and shortening the NH···N bond.)

We have attempted to estimate the contribution of the CH···O interaction in base-pair binding energies by computing a theoretical double-mutant cycle[30] for each of the WC and H structures of A·T (shown for WC in Scheme 1).[31] This cycle compares the computed binding energy of the original A·T complex with those of three other structures in which the CH···O interaction is eliminated: (a) a complex in which the C═O group in this interaction is replaced by CH_2 (**7**), (b) a structure in which the adenine hydrogen atom involved in the interaction is removed (**8**), and (c) a structure in which both of these changes are made (**9**). The contribution of the CH···O interaction is then obtained from the four binding energies as

$$\Delta E = E(\text{A·T}) - E(\mathbf{7}) - E(\mathbf{8}) + E(\mathbf{9}) \qquad (3)$$

This cycle is designed to cancel the contributions of any new interactions that may arise in the modified structures.

For optimal cancellation we took advantage of the fact that electronic-structure calculations, unlike experimental studies, are not limited to stable structures that can be studied in the laboratory, but can be carried out for arbitrary arrangements of the atoms. We therefore adopted a "minimal change" strategy, in which all atoms not changed or replaced maintain the positions they had in the unmodified A·T. For **7** we replaced the relevant thymine carbonyl by CH_2 (with the CH_2 perpendicular to the plane of the rings and the HCH angle bisector along the line of the original C═O bond), while for **8** we simply removed the relevant H atom of adenine, leaving a radical behind. Structure **9** combined both changes. The calculations for the radicals (**8** and **9**) were carried out using ROMP2 (restricted open-shell MP2). The calculated energies of the double-mutant cycle and the resultant interaction energies are listed in Table 5. They show a CH···O interaction energy of −2.3 to −2.5 kcal mol^{-1}, consistent with the value found in the empirical analysis, although the former are gas phase ΔE values and the latter solution phase ΔG data.

(30) Selected examples: (a) Carter, P. J.; Winter, G.; Wilkinson, A. J.; Fersht, A. R. *Cell* **1984**, *38*, 835−840. (b) Faiman, G. A.; Horovitz, A. *Protein Eng.* **1996**, *9*, 315−316. (c) Carver, F. J.; Hunter, C. A.; Seward, E. M. *Chem. Commun.* **1998**, 775−776. (d) Hof, F.; Scofield, D. M.; Schweizer, W. B.; Diederich, F. *Angew. Chem., Int. Ed.* **2004**, *43*, 5056−5059.

(31) Performing a double mutant cycle experimentally is complicated by the potential presence of differing amounts of normal and reverse Watson−Crick and Hoogsteen forms. In DNA the deletion of the thymine carbonyl group significantly destabilizes the double helix but the deletion alters metal ion binding sites, the spine of hydration, duplex curvature, etc.: (a) Woods, K.; Lan, T.; McLaughlin, L. W.; Williams, L. D. *Nucleic Acids Res.* **2003**, *31*, 1536−1540. (b) Meena; Sun, Z.; Mulligan, C.; McLaughlin, L. W. *J. Am. Chem. Soc.* **2006**, *128*, 11756−11757.

Table 4. Calculated Hydrogen-Bond Structural Parameters for A·T[a]

	NH···O			NH···N			CH···O			
	N···O	N–H	∠HNO	N···N	N–H	∠HNH	C···O	C–H	H···O	∠HCO
WC	2.957	1.018	4	2.854	1.044	0	3.616	1.084	2.786	34
H	2.988	1.016	7	2.822	1.042	3	3.430	1.079	2.724	42

[a] Distances in Å, angles in deg.

Table 5. Double-Mutant Cycle in C_s Symmetry (ΔE, kcal mol^{-1})

	A·T	**7**	**8**	**9**	A·T − **7** − **8** + **9**
WC	−16.01	−14.59	−13.94	−14.85	−2.33
H	−17.15	−13.08	−14.20	−12.62	−2.49

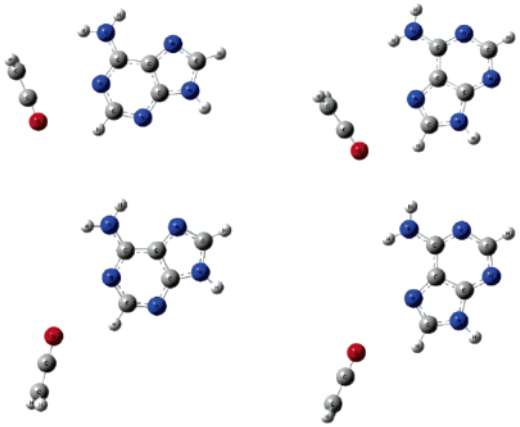


Figure 3. Model adenine·ketene complexes, placed so as to reproduce the O···HC and CO···H geometries of the WC (left) and H (right) A·T complexes in the cis (top) and trans (bottom) configurations.

Table 6. Computed Binding Energies (in kcal mol^{-1}) of the Adenine·Ketene and Adenine·HNCO Models

	cis		trans	
probe	WC	H	WC	H
ketene	−1.5	−2.2	−0.5	−0.6
HNCO	−2.0	−2.6	−0.2	−0.2

Because the CH···O configuration in the A·T complexes and in other complexes listed in the Supporting Information does not conform to the usual requirements for a hydrogen bond, and especially because the strength of the interaction is unusually high, we attempted to characterize this interaction further by computing the binding energies of model adenine·ketene and adenine·isocyanic acid complexes, with the ketene or HCNO placed so as to keep the oxygen lone pairs in the plane of the rings, and reproduce the C−H···O═C geometrical arrangement in the WC and H complexes, respectively. The complexes with ketene are shown in Figure 3, and the energies for both ketene and HNCO model complexes are reported in Table 6 under the heading "cis". The difference in binding energies between H and WC in these models, −0.7 and −0.6 kcal mol^{-1}, accounts for about half of the corresponding difference in A·T, −1.14 kcal mol^{-1} (Table 3). This difference can be partly attributed to the difference in distance between the probe and the adenine, as seen from calculated interaction energies for cis adenine·ketene, using the WC binding site but with the H distance between the monomers (−2.0 kcal mol^{-1}) and using the H binding site but with the WC distance (−1.7 kcal mol^{-1}).

In a further attempt to ascertain the mechanism of the CH···O interaction, additional calculations were carried out for adenine·ketene and adenine·HNCO complexes in which the probes (ketene and HNCO) were reflected in the H···O line, producing a "trans" configuration, with results also shown in Table 6. This trans arrangement (shown for ketene in Figure 3) has the same C═O···H angle as in the cis configuration, and therefore preserves the approximate alignment of an oxygen lone pair along the H···O line. For a normal hydrogen bond one may expect the interaction energies in the cis and trans configurations to be similar, so that the results in Table 6 may be an indication that the interaction does not represent such a bond. Of greater significance may be the fact that the adenine CH bond points toward different regions of the probe carbonyl charge distribution in the two configurations.

It is also significant that all the complexes with $n_{CH\cdots O} > 0$ used in the phenomenological analysis (see Supporting Information) include the same N−C═O···H−C−N configuration, and some part of this configuration may play a role in the interaction. Attempting to attribute the difference in binding energy between the trans and cis configurations to differences in dipole−dipole interactions fails, because the most favorable dipole alignment actually occurs in the Hoogsteen trans configuration (see Figure 2), and other attempts to account for the interaction purely in terms of electrostatic contributions (e.g., an N: to C═O interaction) encounters the problem that many complexes with $n_{CH\cdots O} = 0$ used in the phenomenological analysis appear to have similar electrostatic interactions to the complexes for which $n_{CH\cdots O} > 0$, without requiring an extra contribution in the fitting. Thus, although our results suggest that the CH···O interaction plays an important role, the detailed mechanism of this interaction remains unclear but definitely involves a broader group of atoms.

Effects of Symmetry Constraints. In the calculations for the A·T pairs we have restricted the complexes to C_s symmetry. It is reasonable to ask what effect this restriction has on the binding energies of the complexes. The data in Table 7 provide binding energies, enthalpies, and free energies for fully optimized A and U monomers and A·U complexes (WC and H) using the reduced basis. It is significant that the energy difference between fully optimized A·U WC and H with the reduced basis set is 0.84 kcal mol^{-1} in favor of H, which is similar to the 1.14 kcal mol^{-1} difference for the C_s structures of WC and H in A·T obtained with the full 6-31+G(d,p) basis (Table 3). (It should be noted that the energies of the adenine monomer reported for the reduced basis set are slightly different for WC and H, since the 6-31+G(d,p) basis set was placed on different atoms in the two cases.) The zero-point and thermal contributions to the binding enthalpies and free energies obtained from the reduced-basis calculations for A·U were used as estimates of the corresponding contributions for the full-basis results for A·T in Table 3.

A further evaluation of the energy effects of symmetry constraints can be made from the data in Table 8. The optimization of thymine with the full basis produced a slightly

Table 7. Summary of Results for A·U with the Reduced Basis without Symmetry Restrictions[a]

	au (E_h)	kcal mol^{-1}				
	E	ΔE	ΔZPE	ΔH (0 K)	ΔH (298.15 K)	ΔG (298.15 K)
uracil (planar)	−413.66123					
adenine for WC	−465.98657					
adenine for H	−465.98648					
A·U (WC)	−879.67319	−15.93	0.59	−15.34	−14.81	−4.74
A·U (H)	−879.67444	−16.78	0.57	−16.21	−15.65	−5.64
H−WC	−0.00125	−0.84	−0.02	−0.87	−0.84	−0.90

[a] The adenine amine group is out of the rings plane; the U and A rings are tilted relative to each other, 6.2° for WC, 4.9° for H.

Table 8. Effects of Relaxation of C_s Symmetry (E_h)

	basis	C_s	C_1	change	kcal mol^{-1}
uracil (puckered)	full	−413.69448	−413.69449	−0.00001	−0.01
thymine (puckered)	full	−452.88465	−452.88470	−0.00005	−0.03
adenine (out-of-plane amine)	full	−466.02321	−466.02389	−0.00068	−0.43
adenine for WC	reduced	−465.98600	−465.98657	−0.00057	−0.36
adenine for H	reduced	−465.98587	−465.98648	−0.00061	−0.38
A·U (WC)	reduced	−879.67312	−879.67319	−0.00007	−0.04
A·U (H)	reduced	−879.67439	−879.67444	−0.00005	−0.03

Table 9. Angles for Amine in the Adenine Monomer and in the A·U Complex (Reduced Basis)

	Watson–Crick			Hoogsteen		
	monomer	A·U	difference	monomer	A·U	difference
(A ring plane)−C_6−N	11.8°	12.9°	+1.1°	11.5°	16.0°	+4.5°
sum of amine angles	350.0°	356.7°	+6.7°	349.6°	357.3°	+7.7°

puckered structure, but with a very small energy difference of 0.03 kcal mol^{-1} between the planar and puckered structures, and in view of the zero-point vibrations, the thymine ring may be considered functionally planar. A somewhat greater effect was found in adenine, in which the rings remained planar, but the amine N atom moved to one side of the plane of the rings and the two amine H atoms moved to the opposite side. The angles made by the amine C−N bond with the plane of the rings and the sum of the three angles around the amine N atom for adenine and for the A·U pairs are given in Table 9. The deviation of the sum of the angles from 360° is a measure of the nonplanarity of the amine group. This deviation is significantly less in the complexes than in the adenine monomer. Uracil remains planar in the reduced-basis optimization, although an optimization with the full basis produced slight puckering, with an energy reduction of only 0.000014 E_h, or less than 0.01 kcal mol^{-1}, relative to the planar structure. In the A·U WC and H conformers, the ring planes of uracil and adenine are tilted relative to each other by 6.2° and 4.9°, respectively. The out-of-plane tilting of the rings results in a slightly more favorable geometry for the NH···O hydrogen bond with the out-of-plane amine group. Nevertheless, the energy effect of all of the symmetry constraints is much smaller in the complexes than in the monomers. It can therefore be concluded that the imposition of C_s symmetry does not lead to significant errors in the computed MP2/6-31+G(d,p) binding energies for H and WC.

Conclusions

The success of the phenomenological model in explaining the variance in the experimental K_{assoc} data, together with the statistical significance of the individual predictors (components) is striking. Although the high R^2 across 256 complexes, spanning over 10 kcal mol^{-1} in free energy, does not establish the physical significance of the individual components per se, it is consistent with the importance of primary hydrogen bonding and secondary interactions as well as intramolecular hydrogen bonding in the ureidopyridine-type complexes. Indeed, eq 2 appears to have considerable predictive value.

The goal of this study was to determine if the A·T and G·C base-pairs contained any unusual stability (e.g., special electronic complementarity[1]). Primary hydrogen bonding and secondary interactions are sufficient to account for the stability of the G·C base-pair but the A·T base-pair was found to be at least 2 kcal mol^{-1} more stable than predicted using these two increments alone. This value represents a lower limit because the K_{assoc} was corrected on the assumption that the four possible complexes, normal and reverse Watson−Crick and normal and reverse Hoogsteen, are equally populated. Experiments in solution do not distinguish normal from reverse modes and generally suggest the presence of both Watson−Crick and Hoogsteen forms (vide supra). The ab initio studies described herein indicated that Hoogsteen geometry is preferred over Watson−Crick geometry in A·T by ca. 1 kcal mol^{-1}. Factors that might contribute to the preference for Hoogsteen geometry are a shorter CH···O contact, a favorable alignment of dipoles, and greater distances between secondary repulsive sites. The "reverse" structures are very slightly higher in energy than the respective WC and H structures.

A handful of other complexes similarly showed experimental stabilities in excess of that predicted by primary hydrogen bonding and secondary interactions alone. Each contains a carbonyl group flanking a C−H group. Including an additional parameter in the phenomenological analysis for the putative CH···O interaction led to eq 2 and the excellent fits to the data seen in Figure 1. Although CH···O hydrogen bonds have considerable precedents and have been discussed in the context of DNA base pairing, our analysis requires an unusually high

value ($\geq$2.2 kcal mol^{-1}), which is minimally 60% the strength of an average primary hydrogen bond in the model. The ab initio results for the double-mutant cycle and for the adenine·ketene and adenine·HNCO models certainly suggest that the CH···O interaction plays an important role (though it a much smaller fraction of the ab initio ΔE binding energy of A·T than of the experimental ΔG values), but suggest that it is a group interaction, not a simple CH···O hydrogen bond.

It has been proposed that in the prebiotic world RNA preceded proteins and functioned both as a catalyst and an information storage unit.[32] Given the hydrolytic instability of cytosine, there is further speculation that this "RNA world"[33] initially had a single base-pair (A·U).[30,34,35] But why A·U? The ability to produce adenine from HCN in up to a 0.5% yield[36] is often cited as a key reason, although there is every reason to believe that a broad range of possible compounds would also be available. The results presented herein suggest that the A·U pair may have been selected as a result of its higher stability in comparison to competitors containing the same number of primary and secondary hydrogen bonds.

Acknowledgment. Funding of this work by the NIH (Grant GM65249) is gratefully acknowledged. Peter Beak, Sharon Shavitt, and Howard Zimmerman are thanked for useful suggestions. We also acknowledge the continuing support of the Ohio Supercomputer Center.

Supporting Information Available: Tables showing structures of all compounds used in the analysis and additional criteria for their selection, compound groupings, complete results of multivariate regression analysis, and complete author list for ref 21. This material is available free of charge via the Internet at http://pubs.acs.org.

JA066341F

(32) (a) Woese, C. In *The Genetic Code. The Molecular Basis for Genetic Expression*; Harper-Row: New York, 1967; pp 179−195. (b) Crick, F. H. C. *J. Mol. Biol.* **1968**, *38*, 267−379. (c) Orgel, L. E. *J. Mol. Biol.* **1968**, *38*, 381−393.
(33) Gilbert, W. *Nature (London*) **1986**, *319*, 618.
(34) Reader, J. S.; Joyce, G. F. *Nature (London*) **2002**, *402*, 841−844.
(35) For arguments against an A·U based prebiotic world see: Shapiro, R. *Origins Life Evol. Biosphere* **1995**, *25*, 83−98.
(36) Oró, J. *Biochem. Biophys. Res. Commun.* **1960**, *2*, 407−415.