Tautomerism and Protonation of Guanine and Cytosine. Implications in the Formation of Hydrogen-Bonded Complexes

Carles Colominas,[†] Francisco J. Luque,[*,‡] and Modesto Orozco[*,†]

Contribution from the Departament de Bioquímica i Biologia Molecular, Facultat de Química, Universitat de Barcelona, Martí i Franquès 1, Barcelona 08028, Spain, and Departament de Farmàcia, Unitat Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Avgda, Diagonal s/n, Barcelona 08028, Spain

Received December 28, 1995[⊗]

Abstract: Tautomerism and protonation of guanine and cytosine in the gas phase and in aqueous solution have been examined by theoretical methods. High level *ab initio* calculations with inclusion of correlation effects at the Møller−Plesset level have been used to study these processes in the gas phase. The influence of solvent has been examined using self-consistent reaction field and Monte Carlo free energy perturbation simulations. The results provide a complete and accurate picture of tautomerism and protonation of these nucleic acid bases. Comparison with the available experimental data gives confidence in the quality of the results derived from theoretical computations. Inspection of the most stable tautomeric forms for the neutral and protonated nucleic acid bases allows rationalization of the formation of unusual DNA structures like the triple helix.

Introduction

The nucleic acid bases have tremendous versatility in the formation of hydrogen-bond complexes because of the presence of numerous hydrogen-bond donor and acceptor groups. These interactions determine the specificity of recognition between nucleic acid bases in DNA and ultimately are responsible for maintaining the genetic code. According to the Watson−Crick model[1] (Figure 1) the adenine(A)−thymine(T) pair is stabilized by two hydrogen bonds between the atoms $N_1(A) \leftarrow N_3(T)$ and $N_6(A) \rightarrow O_4(T)$, whereas the recognition between guanine (G) and cytosine (C) is determined by three hydrogen-bonds, which involve the atoms $N_1(G) \rightarrow N_3(C)$, $N_2(G) \rightarrow O_2(C)$, and $O_6(G) \leftarrow N_4(C)$. It is important to remember that formation of hydrogen-bond interactions between A−T and G−C base pairs in DNA does not exhaust the possibilities for establishing additional hydrogen bonds, since several hydrogen-bond forming groups still remain available. Inspection of Figure 1 shows that the atoms N_3(A,G), N_2(G), and O_2(C,T) are pointing toward the minor groove, while the atoms N_7(A,G), O_6(G), N_6(A), O_4(T), and N_4(C) are oriented towards the major groove. These groups are able to make specific hydrogen-bond interactions with other molecules, ranging from small drugs recognized in the minor groove of DNA (minor-groove binders) to macromolecules interacting with the major groove. It is worth noting that the pattern of hydrogen bonds in the minor and major grooves of DNA is specific for each base pair (see Figure 1). This point is extremely important, since it enables the reading of DNA sequences without opening base pairs, allowing specificity for recognition and binding of other molecules to DNA.

The specific reading of DNA along the major groove is crucial in the control of replication and transcription.[2] Furthermore, sequence-specific recognition through the minor groove has been exploited for the design of new antibiotic and antitumoral drugs.[3] Recently, promising therapeutic strategies have been devised based on the triple helix, that is formed from the interaction between a DNA duplex and a third polynucleotide chain which binds to the major groove of the duplex.[4] The structure of the

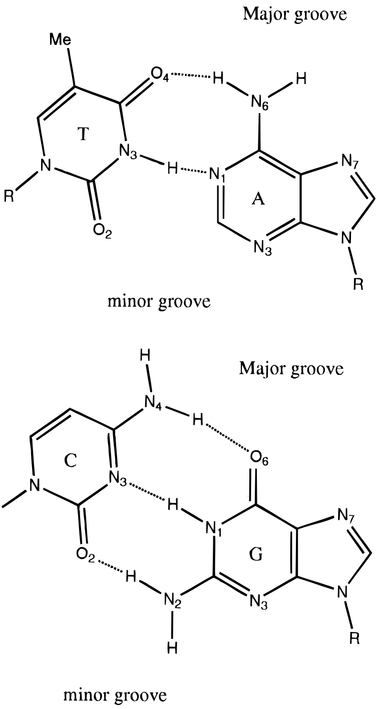


Figure 1. Watson−Crick pairings between adenine(A)−thymine(T) and guanine(G)−cytosine(C).

[*] To whom correspondence should be sent.

[†] Departament de Bioquímica i Biologia Molecular.

[‡] Departament de Farmàcia.

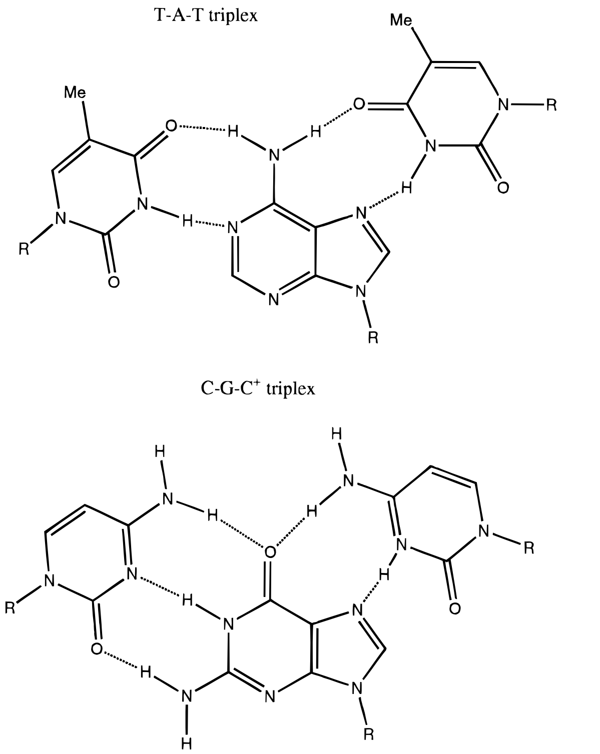
[⊗] Abstract published in *Advance ACS Abstracts,* June 15, 1996.

[1] Watson, J. D.; Crick, F. H. C. *Nature* **1953**, *171*, 737.

[2] (a) Sinden, R. R. *DNA Structure and Function*; Academic Press: San Diego, CA, 1994; Chapter 8, and references therein. (b) Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, Chapters 4−5 and references therein.

[3] (a) Dickerson, R. E.; Kopka, M. L.; Pjura, P. E. In *DNA−Ligand Interactions*; Guschlbauer, W., Saenger, W., Eds.; Plenum Publishing Corporation: 1987; pp 45−62. (b) Wang, A. H.-J. In *Nucleic Acids and Molecular Biology, Vol 1*; Eckstein, F., Lilley, D. M. J., Eds.; Springer-Verlag: Berlin, Heidelberg, 1987; pp 54−69. (c) Neidle, S.; Pearl, L. H.; Skelly, J. V. *Biochem. J.* **1987**, *243*, 1. (d) Gago, F.; Reynolds, C. A.; Richards, W. G. *Mol. Pharmacol.* **1989**, *35*, 232. (e) Geirstanger, B. H.; Wemmer, D. E. *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 463.

S0002-7863(95)04293-4 CCC: $12.00 © 1996 American Chemical Society


T-A-T triplex

C-G-C$^+$ triplex

Figure 2. Hydrogen-bond interactions in TAT and CGC triplexes.

triple helix is modulated by hydrogen bonds between the Watson−Crick base pair of the DNA duplex and the nucleic acid base of the third strand, which can be composed either of pyrimidines (pyrimidine motif) or purines (purine motif). Two types of pyrimidine motifs of particular interest are the dT·dA·dT and dC·dG·dC triplexes (Figure 2). Historically, poly(dT·dA·dT) triple helices[4a,b] were first discovered. Subsequent studies showed that poly(dC·dG·dC) also formed very stable triple helices.[4c,d] A relevant point is that the formation of the triple helix seems to be quite sequence-specific. This property is very interesting, since it opens the possibility of designing oligomeric sequences that block a given DNA duplex *via* triple helix formation. Research on the structure and properties of triplex DNA is an exciting new area of pharmacology.

The complex network of hydrogen-bond interactions that modulate recognition of DNA bases is based on the assumption of specific tautomeric and ionic states for the nucleic acid bases.[1] The importance of tautomeric equilibria has been widely recognized since the early work of Watson and Crick. Several models of spontaneous mutation in DNA are based on the existence of minor tautomeric forms of the bases.[5] This explains the great experimental and theoretical effort focused on the study of tautomerism of nucleic acid bases (for recent studies on tautomerism of guanine and cytosine see refs 5−7). Nevertheless, several aspects still remain unclear because of experimental problems in studying scarcely populated species, and theoretical difficulties in obtaining accurate results from quantum mechanical calculations in the aqueous phase. For instance, there is scarce information on the role of tautomerism in the reading of base pair sequence through the major and minor grooves. In particular, knowledge of the role of tautomeric equilibria in the stabilization of triple helix structures has not been systematically analyzed. This is surprising considering that minor tautomeric forms, i.e., the imino form of cytosine in the third strand, might be important in the stabilization of poly(dC·dG·dC), as is clear from a detailed inspection of the triple base structure (Figure 2).

The influence of acid−base equilibria on the stabilization of DNA structure has also received scant attention. Undoubtedly, this is because nucleic acid bases are neutral under physiological conditions, implying a negligible role for ionization in the physiological structure of DNA. However, recent data suggest that ionization may be relevant in determining mutagenic properties of analogs of nucleic acid bases.[8] Moreover, it is known that DNA polymerase can incorporate ionized base pairs into DNA.[9] There is overwhelming evidence that the triple helix of poly(dC·dG·dC) is greatly stabilized at acidic pH,[10] which suggests that protonation of the bases might contribute significantly to the stabilization of the triplex structure (Figure 2). Because of the pK_a values in aqueous solution of guanine and cytosine,[11] the latter base is predicted to be protonated. Nevertheless, there is no direct evidence on this point, and recent experimental data in the gas phase seems to argue against this idea.[12] Knowledge of the attachment of the proton to either guanine or cytosine is essential for the design of new intercalating drugs that stabilize the triple helix.[13]

In this paper a systematic study of the tautomerism and protonation of guanine and cytosine in the gas phase and in

[4] (a) Felsenfeld, G.; Davies, D. R.; Rich, A. *J. Am. Chem. Soc.* **1957**, *79*, 2023. (b) Arnott, S.; Selsing, E. **1974**, *88*, 509. (c) Mirkin, S. M.; Lyamichev, V. I.; Drushlyak, K. N.; Dobrynin, V. N.; Filippov, S. A.; Frank-Kamenetskii, M. D. *Nature* **1987**, *330*, 495. (d) Letai, A. G.; Palladino, M. A.; Fromm, E., Rizzo, V.; Fresco, J. R. *Biochemistry* **1988**, *27*, 9108. (e) Radhakrishnan, I.; Patel, D. J. *Biochemistry* **1994**, *38*, 11405. (f) Frank-Kamenetskii, M. D.; Mirkin, S. M. *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 319. (g) Plum, G. E.; Pilch, D. S.; Singleton, S. F.; Breslauer, K. J. *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 319.

[5] Kwiatkowski, J. S.; Pullman, B. *Adv. Het. Chem.* **1975**, *18*, 199. (b) Topal, M. D.; Fresco, J. R. *Nature* **1976**, *260*, 285.

[6] (a) Dreyfus, M.; Bensaude, O.; Dodin, G.; Dubois, J. E. *J. Am. Chem. Soc.* **1976**, *93*, 6338. (b) Kaito, A.; Hatano, M.; Ueda, T.; Shibuya, S. *Bull. Chem. Soc. Jpn.* **1980**, *53*, 3073. (c) Lin, J.; Yu, M.; Peng, S.; Akiyama, I.; Li, K.; Lee, L. K.; LeBreton, P. R. *J. Phys. Chem.* **1980**, *84*, 1006. (d) Szczepaniak, K.; Szczesniak, M. *J. Mol. Struct.* **1987**, *156*, 29. (e) Szczepaniak, K.; Szczesniak, M.; Person, W. B. *Chem. Phys. Lett.* **1988**, *153*, 39. (f) Szczesniak, M.; Szczepaniak, K.; Kwiatkowski, J. S.; KuBulat, K.; Person, W. B. *J. Am. Chem. Soc.* **1988**, *110*, 8319. (g) Nowak, M. J.; Lapinski, L.; Fulra, J. *Spectrochimica Acta* **1989**, *45A*, 229. (h) LeBreton, P. R.; Yang, X.; Urano, S.; Fetzer, S.; Yu, M.; Leonard, N. J.; Kumar, S. *J. Am. Chem. Soc.* **1990**, *112*, 2138. (i) Szczepaniak, K.; Szczesniak, M.; Szajda, W.; Person, W. B.; Leszczynski, J. *Can. J. Chem.* **1991**, *69*, 1705. (j) Thewalt, U.; Bugg, C. E.; Marsh, R. E. *Acta Crystallogr., Sect B.* **1971**, *27*, 2358.

[7] (a) Cieplak, P.; Bash, P.; Singh, U. C.; Kollman, P. A. *J. Am. Chem. Soc.* **1987**, *109*, 6283. (b) Lés, A.; Adamowicz, L.; Bartlett, R. J. *J. Phys. Chem.*, **1989**, *93*, 4001. (c) Kwiatkowski, J. S.; Person, W. B.*Theoretical Biochemistry and Molecular Biophysics*; Beveridge, D. L., Lavery, R., Eds.; Adenine Press: 1990; pp 153−171. (d) Kwiatkowski, J. S.; Leszczynski, J. *J. Mol. Struct. (THEOCHEM)* **1990**, *208*, 35. (e) Sabio, M.; Topiol, S.; Lumma, W. C. *J. Phys. Chem.* **1990**, *94*, 1366. (f) Leszcynski, J. *Chem. Phys. Lett.* **1990**, *174*, 347. (g) Katritzky, A. R.; Karelson, M. *J. Am. Chem. Soc.* **1991**, *113*, 1561. (h) Gould, I. R.; Green, D. V. S.; Young, P.; Hillier, I. H. *J. Org. Chem.* **1992**, *57*, 4434. (i) Young, P.; Green, V. S.; Hillier, I. H.; Burton, N. A. *Mol. Phys.* **1993**, *80*, 503. (j) Ford, G. P.; Wang, B. *J. Mol. Struct. (THEOCHEM)* **1993**, *283*, 49. (k) Stewart, E. L.; Foley, C. K.; Allinger, N. L.; Bowen, J. P. *J. Am. Chem. Soc.* **1994**, *116*, 7282. (l) Hall, R. J.; Burton, N. A.; Hillier, I. H.; Young, P. E. *Chem. Phys. Lett.* **1994**, *220*, 129. (m) Estrin, D. A.; Paglieri, L.; Coringiu, G. *J. Phys. Chem.* **1994**, *98*, 5653. (n) Leszcynski, J. *J. Mol. Struct. (THEOCHEM)* **1994**, *311*, 37. (o) Gould, I. R.; Burton, N. A.; Hall, R. J.; Hillier, I. H. *J. Mol. Struct. (THEOCHEM)* **1995**, *331*, 147. (p) Kwiatkowski, J. F.; Barlett, R. J.; Person, W. B. *J. Am. Chem. Soc.* **1988**, *116*, 7282.

[8] Sowers, L. C.; Goodman, M. F.; Eritja, R.; Kaplan, B. E.; Fazakerley, G. V. *J. Mol. Biol.* **1989**, *205*, 437.

[9] Yu, H.; Eritja, R.; Bloom, L. B.; Goodman, M. F. *J. Biol. Chem.* **1993**, *268*, 15935.

[10] Völker, J.; Klump, H. H. *Biochemistry* **1994**, *33*, 13502

[11] (a) Christensen, J. J.; Rytting, J. H.; Izatt, R. M. *J. Phys. Chem.* **1967**, *71*, 2700, (b) Taylor, H. F. W. *J. Chem. Soc.* **1948**, 765.

[12] Greco, F.; Liguori, A.; Sindona, G.; Uccella, N. *J. Am. Chem. Soc.* **1990**, *112*, 9092.

[13] (a) Wilson, W. D.; Tanious, F. A.; Mizan, S.; Yao, S.; Kiselyov, A. S.; Zon, G.; Strekowski, L. *Biochemistry* **1993**, *32*, 10614. (b) Fox, K.; Polucci, P.; Jenkins, T. C.; Neidle, S. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 7887.

aqueous solution is presented. High level *ab initio* methods have been combined with self-consistent reaction field (SCRF) and Monte Carlo free energy perturbation (MC-FEP) techniques. The results provide a detailed picture of tautomerism and protonation of these bases and may give insight into the chemistry of unusual DNA structures such as the poly(dC•dG•dC) triple helix.

Methods

The study of all possible tautomers of neutral and monoprotonated guanine and cytosine at a high *ab initio* computational level becomes exceedingly expensive, making the use of a stepwise elimination scheme advisable. Accordingly, the stability of all tautomers in the gas phase was determined at the AM1[14] semiempirical level, and the influence of hydration on tautomerism was estimated from SCRF-AM1 calculations (see below). Those tautomers whose stability relative to the most stable tautomeric form was less than 10 kcal/mol (either in gas phase or in aqueous solution) were considered for further analysis at the *ab initio* level. Initially, *ab initio* calculations in the gas phase were performed at the HF/6-31G(d)//HF/6-31G(d) level,[15] and solvent was introduced using SCRF-6-31G(d) methods. This allowed us to further limit the number of tautomers included in the final part of the study. Typically, those tautomers having a free energy difference (either in gas phase or in aqueous solution) within 3−4 kcal/mol from that of the most stable tautomer were considered. The final part of the study included geometry optimization at the MP2/6-31G(d) level followed by single point calculations at the MP2/6-311++G(d,p)[16] level. In addition, corrections for electron correlation using up to fourth-order Møller−Plesset level[17] were also considered (see below). Finally, those tautomers that presumably have a relevant biological role were always included, irrespective of whether or not the relative stability fulfilled the cutoff criteria mentioned in the stepwise scheme.

The *ab initio* calculations in the final part of the study were performed at different levels with a twofold purpose: to examine the accuracy of the results and to determine the suitability of inexpensive methods for further studies. Single-point calculations were carried out with the 6-311++G(d,p) basis set at the SCF and MP2 levels using the HF/6-31G(d) and MP2/6-31G(d) optimized geometries. Correlation effects up to fourth order were introduced assuming the transferability of the correction between MP4 (MP3) and MP2 levels determined with the 6-31G(d) basis to the MP2/6-311++G(d,p) results (the correction was determined using the MP2/6-31G(d) geometry). Single, double, triple, and quadruple excitations were considered in MP4 calculations for cytosine, but triple excitations were neglected for guanine. All the MPx calculations were performed with the frozen-core approximation. The highest level calculation determined in this way is denoted in the text as MP4/6-311++G(d,p)//MP2/6-31G(d). Finally, density functional calculations (DFT) were performed using the Becke3−Lee−Yang−Parr (B3LYP) functional.[18] The MP2/6-31G(d) geometry was used in DFT calculations with the 6-31G(d) and 6-311++G(d,p) basis sets. In all cases thermal and entropic corrections were computed from the HF/6-31G(d) geometries using standard procedures in Gaussian 92-DFT.

The free energy of tautomerization or protonation in solution was determined according to eq 1. The relative free energies of hydration ($\Delta\Delta G_{A\to B}{}^{hyd}$) were computed from the absolute free energies of hydration ($\Delta G_A{}^{hyd}$, $\Delta G_B{}^{hyd}$) as determined from SCRF calculations using the AM1[19] and *ab initio* 6-31G(d)[20] optimized versions of the continuum model developed by Miertus, Scrocco, and Tomasi (MST).[21] MST calculations were performed following the standard protocol for neutral and protonated species.[19−22] Calculations were carried out using gas phase geometries, since small geometrical changes were expected upon solvation of rigid molecules like those considered here.

$$\Delta G^{aq}_{A\to B} = \Delta G^{gas}_{A\to B} + \Delta G^{hyd}_{B} - \Delta G^{hyd}_{A} = \Delta G^{gas}_{A\to B} + \Delta\Delta G^{hyd}_{A\to B} \quad (1)$$

In particularly relevant cases, MC-FEP calculations were also performed to examine the adequacy of MST estimates. The values of $\Delta\Delta G_{A\to B}{}^{hyd}$ were estimated from the mutation between the species A and B according to Zwanzig's theory.[23] The solute was placed in a cubic box (~8000 Å^3) containing approximately 260 TIP4P water molecules.[24] Periodic boundary conditions were used in conjunction with a residue-based 9 Å cutoff for nonbonded interactions. Simulations were performed in the isothermal-isobaric ensemble (NPT, 1 atm, 298 K). Solute rotations and translations were adjusted to obtain around 40% acceptance. The mutation was carried out in 21 double-wide sampling windows, which allowed determination of the hysteresis error in the calculation of $\Delta\Delta G_{A\to B}{}^{hyd}$. In each window 2×10^6 configurations were used for equilibration and 3×10^6 configurations for averaging. The molecular geometries, which were taken from the MP2/6-31G(d) geometry optimizations, were not sampled during the simulations, making the SCRF and MC-FEP fully comparable. Standard electrostatic atomic charges[25] were determined from the 6-31G(d) wave functions. The van der Waals force field parameters were taken from the OPLS force-field.[26]

Gas phase calculations were carried out using MOPAC93-Rel. A[27] and Gaussian 92-DFT[28] computer programs. MST calculations were performed using locally modified versions of MOPAC93-Rel A and MonsterGauss.[29] Electrostatic charges were determined using the MOPETE/MOPFIT programs.[30] Monte Carlo simulations were carried out using the BOSS-3.4 computer program.[31] All simulations were performed on the

[14] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

[15] Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1978**, *28*, 213.

[16] McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639.

[17] Möller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

[18] (a) Becke, A. D., *J. Chem. Phys.* **1993**, *98*, 5648. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B.* **1988**, *37*, 785.

[19] (a) Luque, F. J.; Bachs, M.; Orozco, M. *J. Comput. Chem.* **1994**, *15*, 847. (b) Orozco, M.; Bachs, M.; Luque, F. J. *J. Comput. Chem.* **1995**, *16*, 563.

[20] Bachs, M.; Luque, F. J.; Orozco, M. *J. Comput. Chem.* **1994**, *15*, 446

[21] (a) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117. (b) Miertus, S.; Tomasi, J. *Chem. Phys.* **1982**, *65*, 239.

[22] Orozco, M.; Luque, F. J. *Chem. Phys.* **1994**, *182*, 237.

[23] Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.

[24] Jorgensen,W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

[25] (a) Momany, F. A. *J. Phys. Chem.* **1978**, *82*, 592. (b) Bonaccorsi, R.; Scrocco, E.; Petrongolo, C.; Tomasi, J. *Theor. Chim. Acta* **1971**, *20*, 331. (c) Orozco, M.; Luque, F. J. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 31.

[26] Pranata, J.; Wierschke, S. G.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1991**, *113*, 2810.

[27] Stewart, J. J. P. MOPAC93, Rev 2; Fujitsu Limited: 1993.

[28] Frisch, M. J.; Trucks, G. W.; Head-Gordon, M.; Gill, P. M. W.; Wong, M. W.; Foresman, J. B.; Johnson, B. G.; Schlegel, H. B.; Robb, M. A.; Replogle, E. S.; Gomperts, R.; Andres, J. L.; Raghavachari, K.; Binkley, J. S.; Gonzalez, C.; Martin, R. L.; Fox, D. J.; Defrees, D. J.; Baker, J.; Stewart, J. J. P.; Pople, J. A. Gaussian 92; Gaussian Inc.: Pittsburgh, PA, 1992.

[29] Peterson, M.; Poirier. R. MONSTERGAUSS; Department of Biochemistry, University of Toronto, Canada; version modified by Cammi, R.; Bonaccorsi, R.; Tomasi, J. 1987, and by Luque, F. J.; Orozco, M. 1994.

[30] Luque, F. J.; Orozco, M. MOPETE/MOPFIT Computer programs; University of Barcelona, 1995.

[31] Jorgensen, W. L. BOSS 3.4 Computer Program; Yale University, 1993.

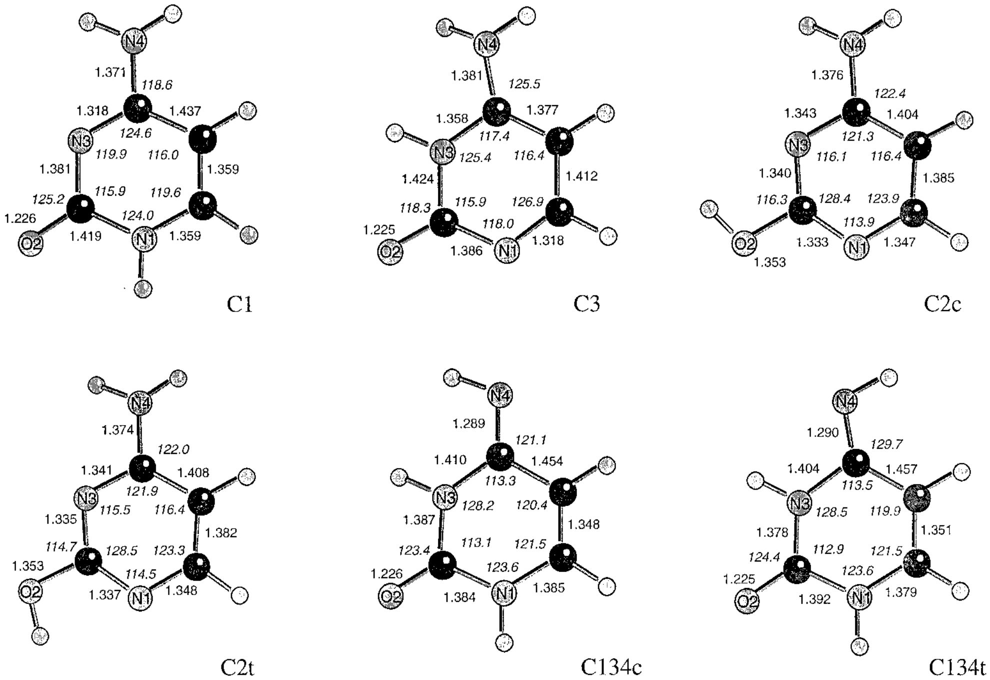


Figure 3. MP2/6-31G(d) structural parameters of the six tautomers of neutral cytosine included in the final study after the stepwise elimination process (see text for details).

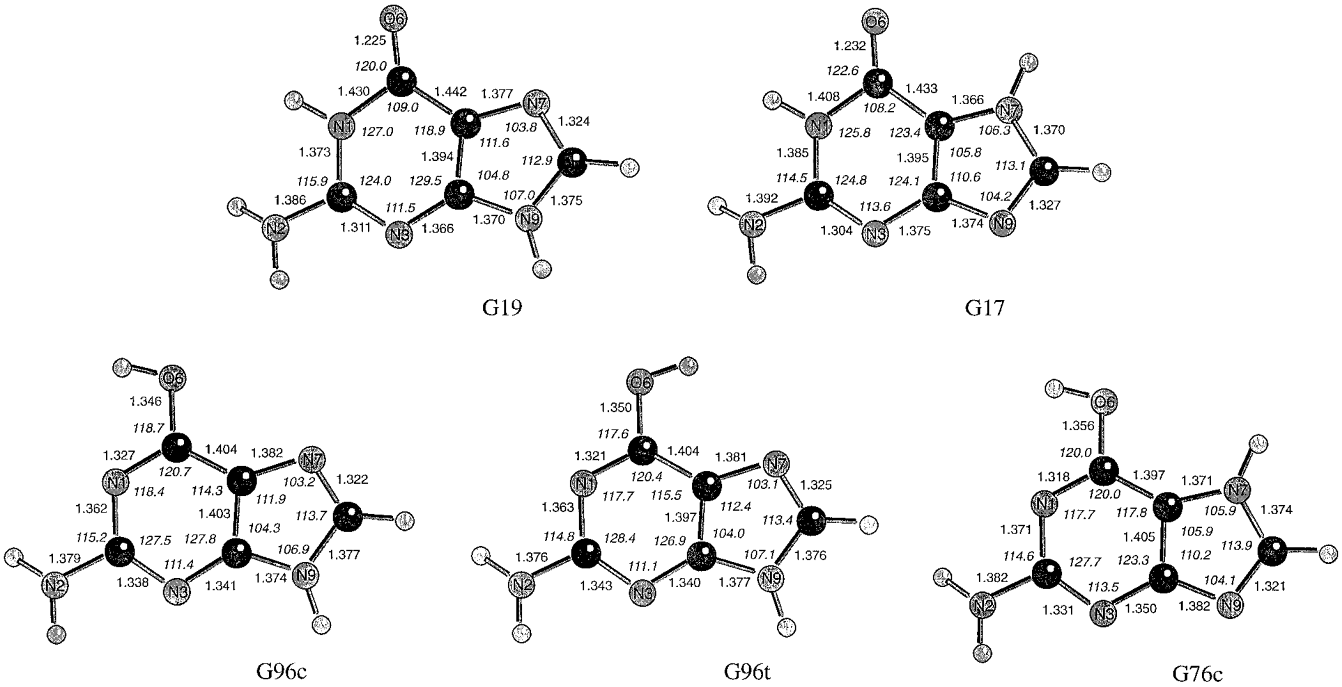


Figure 4. MP2/6-31G(d) structural parameters of the five tautomeric forms of neutral guanine included in the final study after the stepwise elimination process (see text for details).

Cray-YMP of the Centre de Supercomputació de Catalunya, and on HP and SGI workstations in our laboratory.

Results

Tautomerism of Neutral Cytosine and Guanine in Gas Phase. Inspection of the semiempirical and *ab initio* HF/6-31G(d) results allowed the exclusion of a large number of tautomers of neutral cytosine from further consideration. The enol-imino forms are extremely unstable and were excluded. Only six tautomers fulfilled the cutoff criteria (see Methods) and were considered for further analysis (Figure 3). At the highest level of theory (method D in Table 1), the enol form with the hydroxyl hydrogen *trans* to N3 (tautomer C2t) is the most stable tautomer of cytosine, followed by the N1-H keto-amino (C1) and the *cis* enol (C2c) tautomers, which are around 0.8 kcal/mol less stable. The imino forms are always less favored than the amino tautomers, but the difference is slight. Thus, the *trans* imino form (C134t) is around 1.6 kcal/mol less stable than the keto-amino C1 tautomer. Accordingly, the existence of a small fraction of imino forms cannot be ruled out. Finally, the N3-H keto-amino (C3) tautomer is clearly less stable than the C1 form.

Only five tautomers of guanine (Figure 4) were considered after screening in the stepwise elimination protocol: two keto-amino (G19 and G17) and three enol-amino (G96c, G96t, and G76c) forms. Semiempirical calculations indicated that the

Table 1. Differences[a] (kcal/mol) in Energy, Enthalpy, and Free Energy in the Gas Phase for Selected Tautomers of Neutral and Protonated Cytosine and Guanine

tautomer[b]	method[c]	*E*	Δ*H*	Δ*G*	tautomer[b]	method[c]	*E*	Δ*H*	Δ*G*
				Neutral Cytosine					
C3	A	6.9	6.8	6.9	C134c	A	2.2	2.6	2.9
	B	7.3	7.1	7.3		B	2.5	2.8	3.1
	C	7.0	6.9	7.0		C	3.5	3.9	4.1
	D	*7.1*	*6.9*	*7.0*		D	*2.4*	*2.8*	*3.0*
	E	6.4	6.3	6.4		E	3.0	3.4	3.6
	F	7.2	7.0	7.1		F	3.1	3.5	3.7
C2c	A	0.6	0.5	0.9	C134t	A	0.5	1.0	1.3
	B	−0.5	−0.5	−0.1		B	0.7	1.2	1.4
	C	−1.8	−1.9	−1.5		C	1.7	2.2	2.5
	D	*−0.4*	*−0.4*	*0.0*		D	*0.9*	*1.3*	*1.6*
	E	1.8	1.7	2.1		E	1.3	1.8	2.0
	F	1.7	1.6	2.0		F	1.4	1.9	2.2
C2t	A	−0.1	−0.2	0.2					
	B	−1.2	−1.3	−0.9					
	C	−2.6	−2.6	−2.2					
	D	*−1.1*	*−1.1*	*−0.8*					
	E	1.2	1.1	1.5					
	F	1.0	0.9	1.3					
				Neutral Guanine					
G17	A	0.7	0.8	0.8	G96t	A	2.7	2.5	2.6
	B	0.4	0.5	0.5		B	1.2	1.1	1.2
	C	−0.5	−0.4	−0.5		C	0.5	0.3	0.4
	D	*0.2*	*0.2*	*0.2*		D	*1.8*	*1.7*	*1.8*
	E	−0.1	0.0	−0.1		E	2.8	2.6	2.8
	F	−0.5	−0.4	−0.4		F	1.9	1.7	1.8
G96c	A	1.5	1.4	1.5	G76c	A	6.1	5.8	5.8
	B	0.2	0.0	0.1		B	4.4	4.1	4.1
	C	−0.1	−0.3	−0.2		C	3.2	2.9	2.9
	D	*1.1*	*0.9*	*1.1*		D	*4.7*	*4.4*	*4.4*
	E	1.9	1.8	1.9		E	5.3	4.9	5.0
	F	1.1	0.9	1.1		F	4.0	3.7	3.7
				Protonated Cytosine					
pC12c	A	0.4	0.3	0.5		D	*−0.5*	*−0.6*	−0.4
	B	−1.4	−1.4	−1.3		E	1.4	1.3	1.4
	C	−1.5	−1.6	−1.4		F	0.2	0.2	0.3
				Protonated Guanine					
pG137	A	4.0	3.8	3.7	pG196t	A	5.4	4.9	4.9
	B	4.0	3.8	3.7		B	3.8	3.3	3.3
	C	4.0	3.8	3.7		C	5.5	4.9	5.0
	D	*3.8*	*3.6*	*3.5*		D	*5.6*	*5.1*	*5.1*
	E	5.1	4.9	4.8		E	6.9	6.4	6.5
	F	5.1	4.9	4.8		F	6.4	5.8	5.9
pG796c	A	4.6	4.5	4.7	pG376c	A	1.1	0.8	0.9
	B	3.2	3.0	3.3		B	−0.3	−0.6	−0.4
	C	2.2	2.0	2.3		C	1.1	0.8	0.9
	D	*3.5*	*3.3*	*3.6*		D	*1.9*	*1.6*	*1.8*
	E	4.3	4.1	4.3		E	2.6	2.4	2.5
	F	3.4	3.2	3.6		F	2.1	1.8	1.9

[a] Relative to tautomers C1 and G19 of neutral cytosine and guanine and tautomers pC13 and pG179 of protonated cytosine and guanine. [b] See Figures 3 and 4 and 6 and 7 for nomenclature of neutral and protonated species, respectively. [c] Computations were performed at the following levels: A: HF/6-31G(d)//HF/6-31G(d); B: HF/6-311++G(d,p)//HF/6-31G(d); C: MP2/6-311++G(d,p)//HF/6-31G(d); D: MP4/6-311++G(d,p)//MP2/6-31G(d); E: B3LYP(6-31G(d))//MP2/6-31G(d); F: B3LYP(6-311++G(d,p))//MP2/6-31G(d). Results at the highest level of calculation are shown in italics (see text for details).

enol-imino forms were very unstable, and were not included in the final part of the study. The keto-imino tautomers were at least 7 kcal/mol less stable than the reference G19 form at the HF/6-31(d) level and were excluded. Despite the possible numerical uncertainties in the results (see below), a few general trends are clear from the free energy differences reported in Table 1. The two keto-amino tautomers (G19 and G17) have a similar stability. The slight preference of the G19 form (0.2 kcal/mol) is clearly within the expected error of our best calculations. The enol tautomers are always less favored than the keto ones. However, the difference is small when the imidazole hydrogen is attached to the N9 atom (tautomers G96c and G96t). Thus, at the highest calculated level the N9-H enol tautomers are 1.1−1.8 kcal/mol less stable than the keto structure, while the N7-H enol (G76c) form differs by more than 4 kcal/mol.

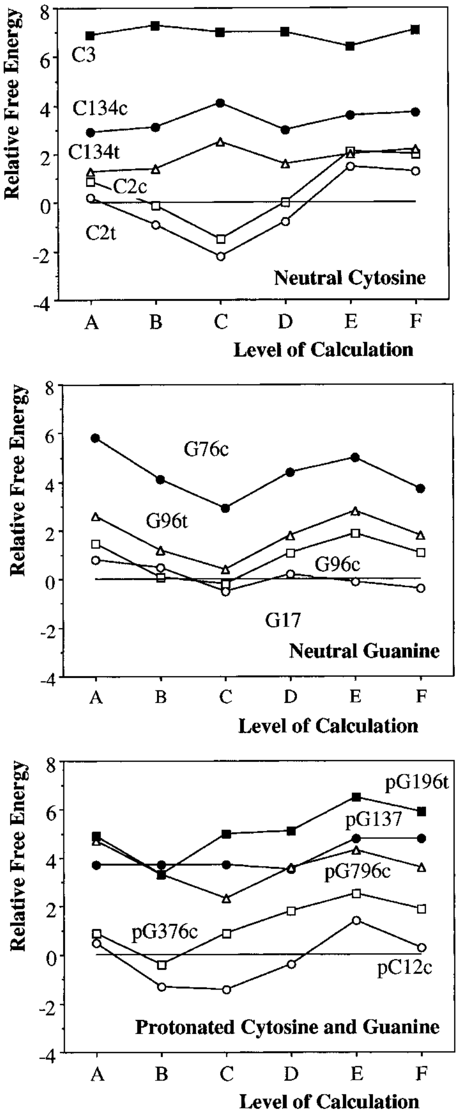


Figure 5. Variation of the relative free energy differences determined at the different levels of theory. The values are relative to the stability of the tautomers C1 and G19 for neutral cytosine and guanine and pC13 and pG179 for protonated cytosine and guanine (see Figures 3, 4, 6, and 7 for nomenclature and footnote c in Table 1 for computational methods).

It is interesting to compare the different theoretical estimates reported in Table 1. This permits examination of convergence of results as the computational level increases and to identify the least expensive computational strategy suitable for further studies. Figure 5 shows the shifts in relative stability in the gas phase at the different levels of theory. In spite of some quantitative differences that will be discussed below, the relative free energy differences between tautomers exhibit a rough parallelism irrespective of the level of computation. Indeed, it is worth noting that the changes in stability at different levels of theory are notably smaller than the cut-offs used in the stepwise elimination process. This gives confidence in the stepwise selection process performed to select the most stable tautomers.

Considering the size of these molecules, the 6-311++G(d,p) basis set is expected to be sufficiently extended and flexible to properly represent molecular properties. Extension of the basis set from 6-31G(d) to 6-311++G(d,p) (sets A and B in Figure 5) has a small effect on the relative stability determined at the HF level, the influence being somewhat larger for the tautomers of guanine. When the 6-311++G(d,p) basis is used, the stability of enol forms increases between 1.0 (C2c) and 1.7 (G76c) kcal/mol, but the relative free energies of the keto (C3 and G17) and imino (C134c and C134t) tautomers remains nearly unaffected. The net effect is that the relative stability of tautomers G17 and G96c of guanine is reversed, but more importantly the enol tautomers C2c and C2t of cytosine become more stable than the reference keto form (C1) by −0.1 and −0.9 kcal/mol at the HF level, respectively. It is interesting to note that the magnitude of the changes in relative stability induced by the basis extension are lower at the DFT level (sets E and F).

Inclusion of correlation effects at the MP2 level (compare sets B and C in Figure 5) increases the stability of enol forms (C2t and C2c) of cytosine by around 1.4 kcal/mol, while the imino species (C134c and C134t) are disfavored by 1.0 kcal/mol. For guanine the relative stability between forms G17 and G96c is reversed again, and these species become slightly more stable than the reference keto tautomer (G19). No relevant changes in the relative stability are found when the MP2/6-31G(d) optimized geometry is considered (data not shown). The most notable differences between HF and MP2 structural parameters concern the length of the carbonyl group, which is enlarged ~0.03 Å, and the geometry of the amino group. Comparison of results at the MP2 and MP4 levels (sets C and D) reveals significant changes from inclusion of correlation effects at the highest levels of theory. The enol forms of cytosine are severely destabilized, even though they are still preferred over the keto tautomer C1. The gain in stability achieved at the MP2 level is counterbalanced by inclusion of correlation effects up to fourth-order, and the net result is that the ordering of stability determined at the HF/6-311++G(d,p) level is nearly recovered. A similar effect is observed for tautomers of guanine. Some quantitative discrepancies are found for the DFT results, at least for the nonlocal B3LYP functional used here. In fact, the SCF values are generally closer to the best estimates than the DFT results. Recent studies[7o] confirm this latest point for other basis sets and molecules.

Confidence in the estimates determined at the highest level of theory can be gained from comparison of the results at the MP2, MP3, MP4(SDQ) and MP4(SDTQ) levels (data not shown). The difference in the free energy of tautomerization (relative to the tautomer C1) for cytosine between MP4(SDQ) and MP4(SDTQ) results is, on average, 0.5 kcal/mol, while it amounts to 1.3 kcal/mol between MP2 and MP4(SDQ) levels, and to 0.8 kcal/mol between MP2 and MP3 levels. This suggests that the MP4/6-311++G(d,p)//MP2/6-31G(d) results are reasonably converged, but they also stress that caution is required for quantitative analysis. Similar results are found for guanine. Thus, the differences between MP2 and MP4 simulations are, on average, less than 1 kcal/mol. However, neglect of triple excitations for guanine introduces additional uncertainties, since the stability of enol forms appear to be underestimated by 0.5 kcal/mol, according to the cytosine results.

Protonation of Cytosine and Guanine. Guided by the stepwise selection scheme only two forms of protonated cytosine were selected for further analysis in the gas phase: the keto-amino pC13 and the enol-amino pC12c forms (see Figure 6). All other enol forms were at least 9 kcal/mol less stable at the *ab initio* HF/6-31G(d) level. The imino species were extremely unstable (the free energy difference relative to the keto-amino form was larger than 20 kcal/mol). At the highest computational level (Table 2) the tautomers pC13 and pC12c have similar stability, the enol form being slightly preferred (−0.4 kcal/mol). Keeping in mind the results for neutral cytosine in the gas phase (Table 1), the most stable enol (C2t and C2c) tautomers would be mainly protonated at N1, even though protonation of C2t requires a change in the orientation of the hydroxyl oxygen. The small fraction of keto (C1) tautomer may be protonated at N3, which generates the pC13 form, or at the oxygen atom, leading to the pC12c tautomer. Therefore three atoms in neutral cytosine (N1, N3, and O2) are susceptible to be protonated all within a range of 1 kcal/mol.

Five tautomers of protonated guanine (see Figure 6) were found to lie within 5 kcal/mol: two keto-amino (pG179 and pG137) and three enol-amino (pG196t, pG376c, and pG796c) forms. MP4/6-311++G(d,p)//MP2/6-31G(d) calculations (Table 1) demonstrate that the preferred form in gas phase is pG179. The most stable enol tautomer (pG376c) is 1.8 kcal/mol less favored, while the rest of the tautomers differ by more than 3.5 kcal/mol. Therefore, results in Table 1 suggest that in the gas phase the G19 tautomer is protonated at N7, while the G17 form protonates at N9, leading to the same species (pG179).

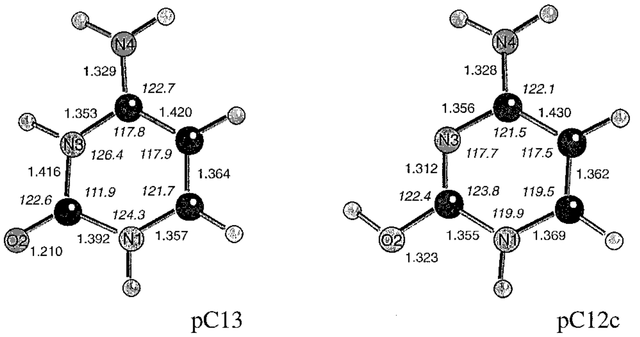
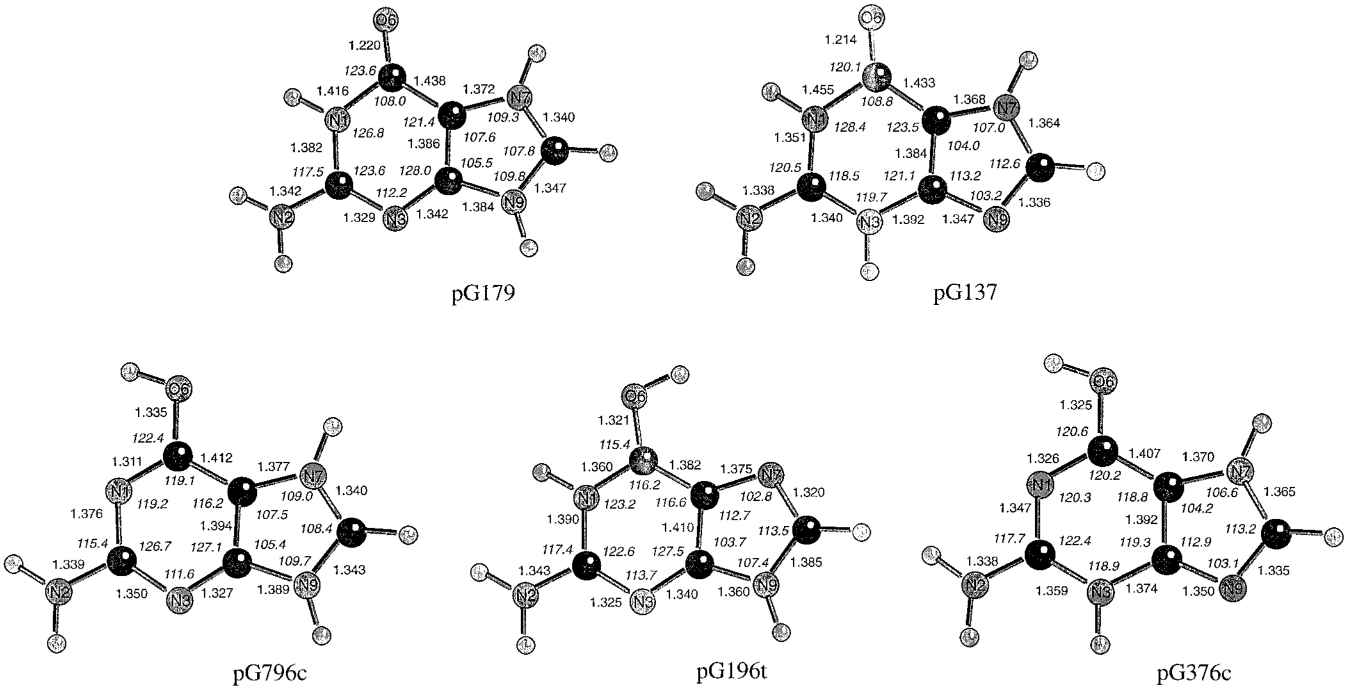


Figure 6. MP2/6-31G(d) structural parameters of the two tautomeric forms of protonated cytosine included in the final study after the stepwise elimination process (see text for details).

Table 2. MST and FEP Free Energies of Hydration ($\Delta\Delta G_{\mathrm{hyd}}$; kcal/mol) and Free Energy Differences[a] ($\Delta G_{\mathrm{t}}^{\mathrm{aq}}$; kcal/mol) in Aqueous Solution for Selected Tautomers of Neutral and Protonated Cytosine and Guanine[b]

tautomer	$\Delta\Delta G^{\mathrm{hyd-MST}}$	$\Delta\Delta G^{\mathrm{hyd-FEP}}$	$\Delta G_{\mathrm{t}}^{\mathrm{aq-MST}}$	$G_{\mathrm{t}}^{\mathrm{aq-FEP}}$
		Neutral Cytosine		
C3	−1.4	−1.2 ± 0.3	5.6	5.8
C2c	7.1		7.1	
C2t	7.6		6.8	
C134c	4.2		7.2	
C134t	4.5	4.5 ± 0.3	6.1	6.1
		Neutral Guanine		
G17	0.8	1.7 ± 0.2	1.0	1.9
G96c	6.1	6.4 ± 0.3	7.2	7.5
G96t	6.2		8.0	
G76c	4.4		8.8	
		Protonated Cytosine		
pC12c	11.1		10.7	
		Protonated Guanine		
pG137	−2.5		1.1	
pG196t	7.5		12.7	
pG376c	8.0		9.8	
pG796c	7.1		10.7	

[a] See footnotes *a* and *b* in Table 1. [b] The tautomerization free energy in the gas phase determined at the MP4/6-311++G(d,p)//MP2/6-31G(d) level was used to compute the tautomerization free energy in aqueous solution (eq 1).



pG179 pG137

pG796c pG196t pG376c

Figure 7. MP2/6-31G(d) structural parameters of the five tautomeric forms of protonated guanine included in the final study after the stepwise elimination process (see text for details).

Depending on the specific tautomer of neutral guanine, the two imidazole centers are susceptible to protonation. Poorer proton affinity is expected for the pyrimidine nitrogens.

As noted for the neutral species, comparison of the different theoretical estimates reveals that extension of the basis set stabilizes the enol forms (pC12c, pG376c, pG796c, and pG196t), while the stability of the keto tautomer pG137 is unaffected. This preferential stabilization is less pronounced at the DFT level. Again, the differences between HF/6-31G(d) and MP2/6-31G(d) optimized geometries are very small and have little effect on the relative stability. However, the counterbalancing effects observed for neutral species upon inclusion of electron correlation are not found for the protonated tautomers. Inspection of the MP2, MP3, and MP4 values (data not shown) indicates a reasonable convergence. Comparison of the SCF and DFT results with the values determined at the highest calculational level reveals some discrepancies at these computational levels for protonated cytosine and guanine.

Solvent Effects on Neutral Cytosine and Guanine. The solvent effect on tautomerism was introduced by means of *ab initio* HF/6-31G(d) MST calculations. However, in some cases it was also determined from MC-FEP simulations (see Methods). The differences in the free energy of hydration ($\Delta\Delta G_{\text{hyd}}$) for tautomers of neutral cytosine and guanine (relative to tautomers C1 and G19 respectively) are given in Table 2. The free energy of tautomerization in aqueous solution (ΔG_{t}^{aq}), as determined by addition of $\Delta\Delta G_{\text{hyd}}$ to the free energy of tautomerization in the gas phase, is also given. The free energy of tautomerization in the gas phase was taken from the result computed at the MP4/6-311++G(d,p)//MP2/6-31G(d) level (method D in Table 1).

The relevance of hydration on tautomerism of cytosine is apparent from the results in Table 2 and Figure 8. The C3 tautomer is the best solvated form, 1.2−1.4 kcal/mol more stabilized by water than the C1 tautomer. The keto-amino (C1 and C3) tautomers are better solvated than the enol (by 7.1−7.6 kcal/mol) and imino (by 4.2−4.5 kcal/mol) forms. As a result, the C1 tautomer is preferred in aqueous solution by more than 5 kcal/mol relative to any other tautomer. There is a large solvent effect for the C2t tautomer, the preferred form in the gas phase. Its population in aqueous solution is expected to be negligible (ΔG_{t}^{aq} > 6 kcal/mol). It is also interesting to note the agreement between the FEP and MST estimates of the relative free energy of hydration, which gives confidence in the results.

Results in Table 2 and Figure 8 show a large destabilization (around 4.4−6.2 kcal/mol) of the enol tautomers (G96c, G96t and G76c) of guanine upon solvation, which favors the keto-amino forms (G17 and G19). The G19 tautomer is more stable than the G17 form by 1.0−1.9 kcal/mol. The population of the other tautomers is expected to be negligible. Again, a close agreement is found between the FEP and MST estimates of the relative free energy of hydration.

Solvent Effects on the Protonation of Cytosine and Guanine. Solvent has a great effect on protonation, due to better hydration of charged species. Thus, while free energies of hydration of neutral tautomers range (in absolute values) between 10−30 kcal/mol, the range is 70−90 kcal/mol for protonated species. For cytosine (Table 2) the pC13 tautomer is the best hydrated species if highly unstable tautomers like pC23c (not shown) or the imino-enol forms are excluded. All the enol forms are disfavored in aqueous solution typically by more than 12 kcal/mol at the 6-31G(d) level. In particular, the pC12c tautomer, which is the most stable in the gas phase, is 10.7 kcal/mol less favored than the pC13 form.

The keto-amino tautomers of guanine are preferentially stabilized upon solvation (Table 2 and Figure 8). The pG137 form is better hydrated than pG179 by 2.5 kcal/mol, but this does not revert the relative stability in the gas phase. As noted before for cytosine the enol forms are largely destabilized. In fact, the pG376c tautomer, which is less stable (1.8 kcal/mol) than pG179 in the gas phase, is disfavored by 9.8 kcal/mol in aqueous solution. Therefore, the protonated guanine is predicted to exist mainly as the pG179 tautomer in aqueous solution with a small fraction (around 15%) of the pG137 form also expected.

Discussion

A detailed picture of tautomerism and protonation of cytosine and guanine in the gas phase and in aqueous solution may help

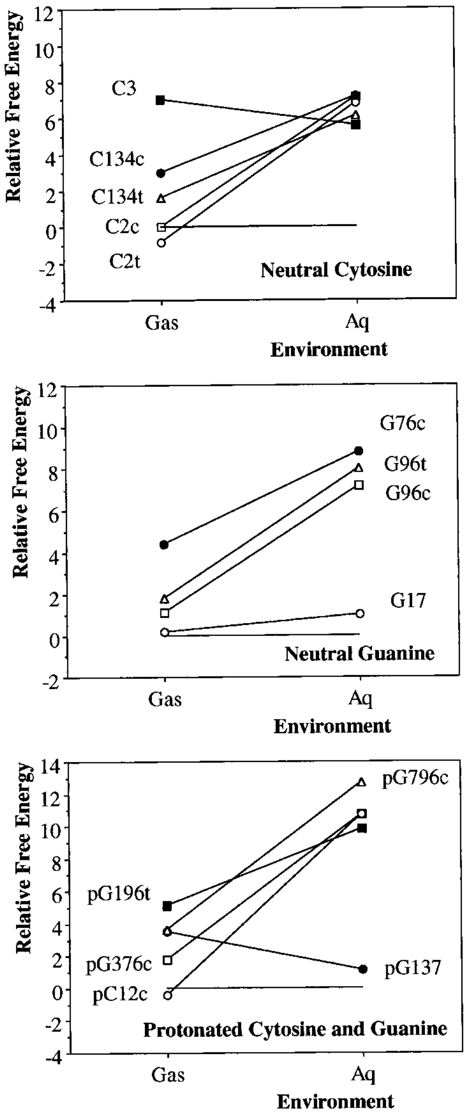


Figure 8. Variation of the relative stability upon solvation. The values in gas phase correspond to the MP4/6-311++G(d,p)//MP2/6-31G(d) estimates. Free energy differences in aqueous solution are determined upon addition of the MST free energies of hydration (see Table 2). The values are relative to the stability of the tautomers C1 and G19 for neutral cytosine and guanine and pC13 and pG179 for protonated cytosine and guanine (see Figures 3, 4, 6, and 7 for nomenclature).

explain the stability of anomalous DNA structures. This has been obtained from high level *ab initio* calculations combined with MST/SCRF and MC-FEP simulations. Furthermore, examination of the results at different levels of theory may be useful in the design of efficient theoretical approaches for the study of more complex systems. In this respect, results indicate that SCF calculations can give reasonable estimates provided that a large basis set is used. In contrast, the nonlocal (B3LYP) DFT formalism seems less adequate for the study of tautomerism in heterocycles. This agrees with findings reported by other authors for similar systems.[9o] Finally, it is worth noting that only small uncertainties, typically less than 1 kcal/mol, are found at the highest level of theory reported here. To our knowledge this is the most accurate computational study on these molecules to date. The similarity between MST and FEP estimates is encouraging considering the fundamental differences of these methodologies. This agreement reinforces confidence in the free energies of solvation estimated in this study. Furthermore, it suggests that the use of effective potentials in MC-FEP simulations is quite efficient for description of polarization effects in tautomerism, and that the parametrization of the 6-31G(d) MST method corrected most of the intrinsic shortcomings of this continuum method.

Tautomerism of Neutral Cytosine and Guanine. In the gas phase cytosine exists as a mixture of at least three main tautomers: the keto-amino C1 and two enol-amino (C2c, C2t) forms, whose relative stabilities lie within 1 kcal/mol of each other. At the highest level of theory the C2t tautomer is the most stable, while the C1 form is destabilized by 0.8 kcal/mol. This estimate is close to the experimental value of 0.4 kcal/mol.[6f] It also agrees with recent theoretical estimates determined at the CCSD/DZP//HF/3-21G (restricted optimization) level of theory by Les et al.[7b] and at the QCSID(T)/6-31G(d)//MP2/6-31G(d) level by Hillier and co-workers,[7o] which also predicts the C2t tautomer to be the most stable (the energy difference is estimated to be around 1.3 kcal/mol). In contrast, previous studies at the MP2/6-31G(d)//HF/6-31G(d) level by Kwiatkowski et al.[7p] and at the MP4/DZP//HF/6-31G(d,p) level by Young et al.[7i] suggested the C1 tautomer to be slightly more stable (by a few tenths of a kcal/mol) than the C2t species. The use of smaller basis sets, the treatment of correlation effects, and the molecular geometry may be responsible for the discrepancy between theoretical results.

Even though the imino forms are minor species, they are still quite stable, as shown by the free energy difference with respect to the C1 tautomer, which ranges between 1.6 (C134t) and 3.0 (C134c) kcal/mol. Accordingly, significant amounts of imino tautomers are expected in the gas phase. On this point, the experimental data are not very precise, but it seems that keto-amino tautomers are preferred over keto-imino forms by a free energy difference similar to or greater than 1.4 kcal/mol,[6f] which agrees with our estimates. Recently, high level simulations[7b,i,o] found the imino form C134t to be a minor, but significant species, whose energy difference relative to the C1 form ranges from 0.5 to 1.8 kcal/mol.

Guanine exists in the gas phase as a mixture of keto-amino (G17 and G19) tautomers with a very small fraction of enol-amino forms. Our best estimates suggest that the G19 tautomer is slightly more stable than G17, but the difference (0.2 kcal/mol) probably lies beyond the accuracy of these calculations. Previous estimates at the MP2/6-31G(d,p) level also found a slight preference (an energy difference of 0.1−0.3 kcal/mol) for G19. The experimental evidence is not clear, since the photoelectron spectra of guanine resemble more closely that of 7-methylguanine,[6c] which suggests that the G17 tautomer is the most stable form in the gas phase. The G19 tautomer is found in the crystal structure of guanine monohydrate,[6j] but in this latter case crystal lattice effects may play a decisive influence. Overall, both experimental and theoretical data indicate a similar stability of G17 and G19 tautomers.

The two keto-amino (G17 and G19) tautomers of guanine are more stable in the gas phase than the enol-amino forms by around 1 kcal/mol. This agrees with previous theoretical calculations,[7f,k,n,o] which provided energy differences ranging from 0.3 to 1.8 kcal/mol. In contrast, experimental data collected in an argon matrix at low temperature suggests a similar population of keto-amino and enol-amino tautomers.[6i] Comparison of SCF, MP2, and MP4 results suggests that the MP4(SDQ) calculations likely overestimates the stability of the keto-amino form due to the neglect of triple excitations and to the small basis set used to compute the MP4−MP2 correction. However, whether or not these factors can justify a difference of stability of approximately 1 kcal/mol with regard to experiment is unclear.

Solvation has a dramatic influence on the tautomerism of neutral cytosine. All the tautomers are strongly destabilized with respect to the keto-amino C1 form, even the enol tautomers C2t and C2c, which are most stable in the gas phase. This finding agrees well with the available theoretical data[7h,i] and with experimental evidence, which precludes the existence of enol tautomers in aqueous solution.[2b] In addition, present calculations reveal that the keto-imino forms are disfavored with regard to the C1 tautomer by around 4.5 kcal/mol. This value matches previous MD-FEP estimates reported by Kollman et al.[7a] and also agrees with SCRF results of Young et al.[7i] and Gould et al.[7h] The solvent-induced destabilization for the imino forms is smaller than for the enol tautomers, which changes the relative population of these tautomers with respect to the situation in the gas phase. Thus, our best estimate suggests that the most stable imino form is preferred by about 1 kcal/mol over the most stable enol tautomer in aqueous solution. The relative stability of keto-imino *versus* keto-amino tautomers ($\Delta G_t^{aq} = 6.1-7.2$ kcal/mol) agrees with the experimental values, which range from 5.5 to 6.8 kcal/mol.[6a]

Tautomerism of guanine is notably changed upon solvation. The enol forms are largely destabilized. The G17 tautomer is disfavored with regard to the G19 form. In fact, this latter tautomer is more stable by 1−2 kcal/mol. Experimentally only keto-amino forms are detected in aqueous solution, but to our knowledge there is no information available concerning which is the major tautomeric form. The influence of solvation on guanine tautomerism has been examined by a few theoretical studies, but only at the semiempirical level,[7g,j] due probably to the size of the molecule. The results are in general agreement with the large solvent-induced stabilization of keto-amino tautomers reported here.

Protonation of Cytosine and Guanine. Protonation of nucleic acid bases plays an essential role in numerous enzymatic reactions, might contribute to the stabilization of irregular DNA structures, and may also be relevant in mutagenic processes.[2,4] Present results suggest that in the gas phase the neutral cytosine exists in the enol form. Protonation occurs mainly at the N1 atom, generating the pC12c form, the major tautomer of protonated cytosine in the gas phase. However, the protonated keto-amino pC13 form is also very stable, as shown by the relative free energy difference (0.4 kcal/mol), and should coexist with pC12c. The gas phase proton affinity of cytosine at the MP4/6-311++G(d,p)//MP2/6-31G(d) level is 227 kcal/mol.[32] This estimate is close to experimental values, which range from 223.8 kcal/mol[33] to 225.9 kcal/mol.[12]

Protonation of guanine in the gas phase occurs mainly at the N7 position, leading to the keto-amino pG179 tautomer as the major species. A minor, but non-negligible form is the enol-amino pG376c tautomer. Considering the most stable tautomers in neutral and protonated states, our best estimate of the proton affinity of guanine is 225.8 kcal/mol, which compares well with the experimental range of proton affinities ranging from 223.0[33] to 227.4[12] kcal/mol.

Comparison of gas phase proton affinities of cytosine and guanine determined at the highest level of theory suggests that protonation of cytosine is easier by 1.2 kcal/mol when the most stable species are considered (nearly the same value is obtained for the standard C1, G19, pC13, and pG179 forms). Unfortunately, the experimental results are controversial. Thus, the proton affinity of cytosine has been reported to be 0.8 kcal/mol larger than that of guanine,[33] but another study found the proton affinity of guanine to be greater by 1.5 kcal/mol.[12] Undoubtedly, the difference in proton affinities is too small to fully guarantee the reliability of our theoretical estimate. However, the similarity between estimated proton affinities and experimental data, the similar difference found between relative proton affinities of cytosine and guanine estimated from all levels of theory, and the excellent agreement found in calculation of relative pK_a's between guanine and cytosine (see below) gives some confidence in the results.

In aqueous solution the keto-amino C1 tautomer of cytosine is mainly protonated at N3, which generates the pC13 species. Similarly, the keto-amino G19 tautomer of guanine is mainly protonated at N7, which leads to the pG179 form. These results agree with all available experimental data[2,11] for protonation of these nucleic acid bases. The relative free energy of protonation between guanine and cytosine in aqueous solution can be estimated from the relative free energy of protonation in the gas phase (determined at the MP4/6-311++G(d,p)//MP2/6-31G-(d) level) and the corresponding free energies of hydration. The results predict that protonation of cytosine is easier by a free energy difference of 1.8 kcal/mol, which closely agrees with the experimental value (1.74 kcal/mol from data in ref 11a,b). The solvent effect accounts for only 0.6 kcal/mol of the difference in the relative pK_a's of cytosine and guanine.

High level *ab initio* calculations reported here state that cytosine is easier to protonate than guanine upon solvation, but they also indicate the preferential protonation of cytosine even in the gas phase. These results suggest that in a GC$^+$ Hoogsteen pair the proton is expected to be provided by the cytosine N3 atom rather than by the guanine N7, in agreement with the generally accepted picture of this interaction. Nevertheless, caution is still required because of the approximations underlying the theoretical models used to compute relative proton affinities in the gas phase and the solvent effects, and also due to the larger complexity of the molecular environment in biochemical systems, particularly in the triple helix.

Biological Implications in the Formation of (dC·dG·dC) Triplex. The CGC pyrimidine motif is found in stable DNA triple helices,[2a,4c−g,10] the maximum stability being achieved at pH ∼4.5.[10] Indeed, a much larger dependence of binding on ionic strength is observed for the TAT triplex than for CGC.[10] These results support the accepted idea that the poly(dC) third strand is protonated in CGC triplexes. On this point, it is interesting that a triple helix containing around 30% GC content is slightly more stable than a poly(dT·dA·dT) triplex at pH 7.0 and 0.1 M of ionic strength,[10] which reveals that the poly(dC·dG·dC) structure is very stable 2.5 pH units beyond the pK_a of cytosine. Since thermal studies have determined that the pK_a of cytosine in the triple helix is nearly identical to that found for the free cytosine in solution,[10,34] other noncovalent interactions may contribute to a local stabilization of the Hoogsteen pairing.

A precise understanding of the stability of triple helices is not feasible without a complete description of the complex interactions in DNA, such as base stacking, sugar-phosphate backbone structure, and solvent-counterion environment. This limits the suitability of present results on the tautomerism and protonation of cytosine and guanine to gain deeper insight into the structure of triple helices, but they do provide a basis for discussing the ability of these nucleic acid bases to establish hydrogen-bonded pairings. In this context, the results reported here support the assumption that the proton required for Hoogsteen pairing of the GC$^+$ dimer is mainly provided by cytosine, and that most of the positive charge in the CGC$^+$ motif

[32] Molecular energies for neutral and protonated species are available upon request to the authors.

[33] Lias, S. G.; Liebmann, J. F.; Levin, R. D. *J. Phys. Chem. Ref. Data* **1984**, *13*, 695.

[34] Record, M. T.; Anderson, C. F.; Lohman,T. M. *Q. Rev. Biophys.* **1978**, *11*, 103.

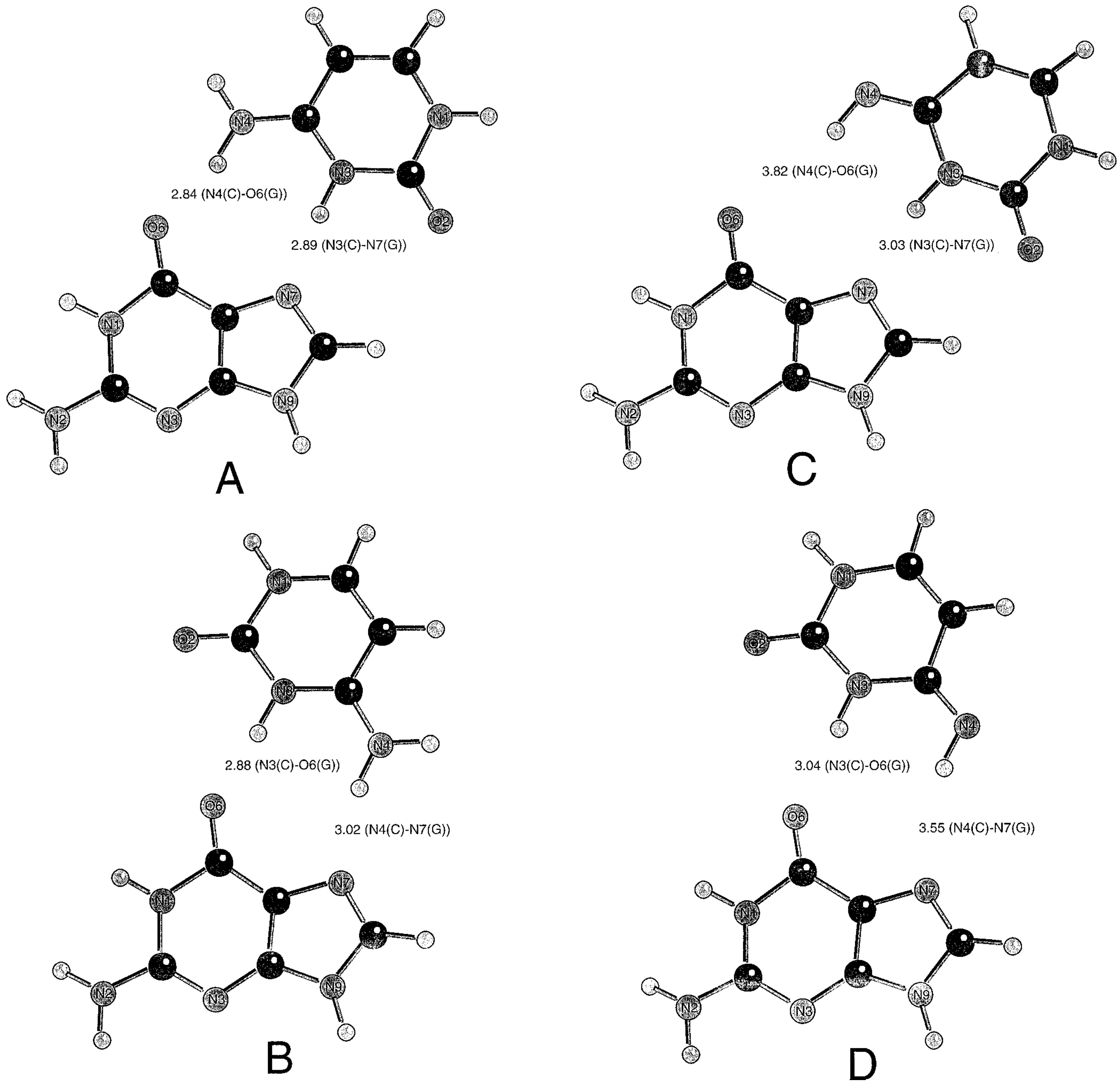


Figure 9. HF/6-31G(d) structural parameters for the Hoogsteen (A, C) and reverse Hoogsteen (B, D) pairings between guanine and protonated cytosine (A, B) and between guanine and the imino species of neutral cytosine (C, D).

is concentrated in the poly(dC) third strand. This finding has potentially relevant implications for the design of new intercalating drugs able to specifically stabilize DNA triplexes.[13]

The poly(dC·dG·dC$^+$) triplex is very stable in an anhydrous environment. HF/6-31G(d) geometry optimizations of the GC$^+$ dimer shows that Hoogsteen pairing (Figure 9) is favored by around 37.5 kcal/mol, and reverse Hoogsteen by around 35 kcal/mol (values determined after correction of the BSSE error by the counterpoise method[35]). These results, which will be influenced by the local environment in the DNA, reveal the intrinsic stability of the CGC$^+$ trimer at acidic pH, which is experimentally known to be greater than that of the TAT triplex.[10] The great stability of the complex likely compensates for the existence of protonated cytosine at neutral or even slightly basic environments.[36]

The formation of the GC pairing at high pH can also be explained assuming the existence of cytosine as the imino tautomer (C134c). In this case both Hoogsteen and reverse Hoogsteen GC(imino) pairings (Figure 9) are stable by -9.5 and -7.1 kcal/mol at the HF/6-31G(d) level after correction of the BSSE error, even though the hydrogen-bond distances are slightly larger than the length typically found in nucleic acid structures (Figure 9). Since the C134c tautomer is disfavored by 3.0 kcal/mol in the gas phase, the interaction energy of the GC(imino) complex seems to be enough to guarantee its stability. Although the simplicity of the theoretical models precludes a rigorous comparison, it is worth noting that the net stabilization energy would be less favorable than that for the AT pairing, which amounts to around 12 kcal/mol according to experimental measures[37] and to 10−13 kcal/mol from recent theoretical calculations.[38] Accordingly, the CGC(imino) triplex

[35] Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

[36] Betts, L.; Joset, J. A.; Veal, J. M.; Jordan, S. R. *Science* **1995**, *270*, 1838.

[37] Yanson, I.; Teplitsky, A.; Sukhodub, L. *Biopolymers* **1979**, *18*, 1149.

is stable, and can contribute to triplex stability, but it is less stable than the TAT triplex, as it is experimentally found in triplex structures at basic pH.

The entirety of these results suggests that protonated cytosine is the species responsible for the formation of the CGC triplex under acidic or neutral conditions. However, at basic pH present results do not allow us to preclude the role of imino tautomers of cytosine. A particular case where the GC(imino) pairing may be important is the triple helix formed by polydC·polydG·polydC fragments. In this case the protonated cytosine of the central CGC trimer would be placed in an extremely positive field resulting from the positive charges of the flanking CGC trimers located around 3.4 Å. Accordingly, it may be advisable the presence of protonated and imino forms of cytosine, even under pH conditions where the protonated cytosine is more abundant than the imino forms.

Acknowledgment. We express our gratitude to Dr. G. C. Shields for helpful suggestions in the preparation of this manuscript. We thank Prof. J. Tomasi for a copy of the MST routines, which were modified by us to perform the present study. We also thank Prof. W. L. Jorgensen for a copy of his BOSS computer program. This work was supported by the Centre de Supercomputació de Catalunya (CESCA, Mol. Recog. Project) and the Dirección General de Investigación Científica y Técnica (DGICYT; PB93-0779 and PB94-0940).

JA954293L

[38] Gould, I. E.; Kollman, P. A. *J. Am. Chem. Soc.* **1994**, *116*, 2493.