Molecular Dynamics Simulations in Aqueous Solution of Triple Helices Containing d(G·C·C) Trios

Robert Soliva,[†] Charles A. Laughton,[‡] F. Javier Luque,[§] and Modesto Orozco*,[†]

Contribution from the Departament de Bioquímica i Biologia Molecular, Facultat de Química, Universitat de Barcelona, Marti i Franques 1, Barcelona 08028, Spain, Cancer Research Laboratories, Department of Pharmaceutical Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom, and Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Avgda Diagonal sn, Barcelona 08028, Spain

Received April 3, 1998. Revised Manuscript Received June 19, 1998

Abstract: We report the results of molecular dynamics studies on the stability of different triple helices containing the pyrimidine motif d(G·C·C) in aqueous solution. The stability of triplexes where the Hoogsteen cytosine is protonated is compared with that of triplexes where the same base is present as the neutral imino tautomer. Results support the hypothesis that Hoogsteen cytosines in triplexes where d(G·C·C) trios are contiguous are not fully protonated, but a certain percentage of neutral form exists. The results help in the understanding of the apparently contradictory experimental data on d(G·C·C)-containing triplexes, and encourages new studies in the field.

Introduction

The ability of DNA to form triple helices has been known for decades, and in recent years many studies have focused on the analysis of these structures, which are known to exist not only *in vitro,* but also in vivo. The recent dramatic increase in interest in these structures relates in great part to their possible therapeutic role in antigene strategies.[1−5] Within this pharmacological approach, a polynucleotidic strand (the triplex-forming oligonucleotide, or TFO) is designed to target the major groove of a selected site in duplex genomic DNA, thus selectively interfering with the expression of a selected gene or, less specifically, interfering with DNA replication.

The triple helix can be formed with purine or pyrimidine motifs. In the first case the TFO consists completely or largely of purines, which interact with the Watson−Crick purine·pyrimidine base pairs of the duplex DNA via the major groove. The interaction is specific: G·C base pairs are recognized by G in the TFO and A·T base pairs by A. In the second case the TFO bases are pyrimidines which interact also via major groove with the purine·pyrimidine base pair. In this case, A·T base pairs are recognized by T in the TFO and G·C base pairs by C. The purine motif leads to an antiparallel arrangement of the two purine strands, while in the pyrimidine motif the TFO strand is parallel to the purine Watson−Crick strand.

Triplexes based on the pyrimidine motifs have been studied in more detail than the purine ones, but many aspects of their structure and reactivity are still unclear. At least three different models for their structure have been suggested from experimental data.[6−8] Recent molecular dynamics (MD) simulations[9] suggest that the d(A·T·T) triplex pertains to the B-family, in agreement with recent NMR and IR data in solution,[10−16] and in disagreement with models generated from fiber diffraction data,[6] and from high-resolution X-ray data of the PNA·DNA·PNA triple helix.[8]

As introduced above, two base trios can contribute to the stability of pyrimidine triplexes: the d(A·T·T) trio and the d(G·C·C) trio. In the d(A·T·T) trio the thymine in the third strand is bound to the adenine by means of Hoogsteen hydrogen bonds (see Figure 1). The strength of Hoogsteen dA-dT pairing, and stacking interactions justify the stability of this triplex step at different pH, ionic strength, and temperatures. The situation regarding the d(G·C·C) trio is much less clear-cut. The d(G·C·C) trio is formed spontaneously, and is quite stable not only at acidic,[17] but also at neutral pH.[18,19] The cytosine in the third strand is known to interact with guanine by means of Hoogsteen-

[†] Departament de Bioquímica i Biologia Molecular, Universitat de Barcelona.
[‡] University of Nottingham.
[§] Departament de Fisicoquímica, Universitat de Barcelona.

(1) Cooney, M.; Czernuszewicz, G.; Postel, E. H.; Flint, S. J.; Hogan, M. E. *Science* **1988**, *241*, 456.
(2) Hélène, C.; Toulme, J. J. *Biochem. Biophys. Acta* **1990**, *1049*, 99.
(3) Strobel, S.A.; Dervan, P. *Methods Enzymol.* **1992**, *216*, 309.
(4) Grigoriev, M.; Praseuth, D.; Guieysee, A. L.; Robin, P.; Thuong, N. T.; Hélène, C.; Harel-Bellan, A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3501.
(5) Sun, J. S.; Garestier, T.; Hélène, C. *Curr. Opin. Struct. Biol.* **1996**, *6*, 327.
(6) Arnott, S.; Bond, P. J.; Selsing, E.; Smith, P. J. C. *Nucleic Acids Res.* **1976**, *11*, 4141.
(7) Raghunathan, G.; Miles, H. T.; Sasisekharan, V. *Biochemistry* **1993**, *32*, 455.
(8) Betts, L.; Josey, J. A.; Veal, J. M.; Jordan, S. R. *Science* **1995**, *270*, 1838.
(9) Shields, G.; Laughton, C. A.; Orozco, M. *J. Am. Chem. Soc.* **1997**, *119*, 7463−7469.
(10) Macaya, R. F.; Schultze, P.; Feigon, J. *J. Am. Chem. Soc* **1992**, *114*, 781.
(11) Howard, F. B.; Miles, H. T.; Liu, K.; Frazier, J.; Raghunathan, G.; Sasisekharan, V. *Biochemistry* **1993**, *31*, 10671.
(12) Radhakrishnan, I.; Patel, D. J. *Biochemistry* **1994**, *33*, 1405.
(13) Radhakrishnan, I.; Patel, D. J. *Structure* **1994**, *2*, 395.
(14) Bornet, O.; Lancelot, G. *J. Biomol. Struct. Dyn.* **1995**,*12*, 803.
(15) Wang, E.; Koshlap, K. M.; Gillespie, P.; Dervan, P. B.; Feigon, J. *J. Mol. Biol.* **1996**, *257*, 1052.
(16) Koshlap, K. M.; Schultze, P.; Brunar, H.; Dervan, P. B.; Feigon. J. *Biochemistry* **1997**, *36*, 2659.
(17) Lee, J. S.; Johnson, D. A.; Moogan, A. R. *Nucleic Acids Res.* **1979**, *6*, 3073.
(18) Völker, J.; Klump, H. H. *Biochemistry* **1994**, *33*, 13502.
(19) Wittung, P.; Nielsen, P.; Nordén, B. *Biochemistry* **1997**, *36*, 7973.

10.1021/ja981121q CCC: $15.00
Published on Web 10/27/1998

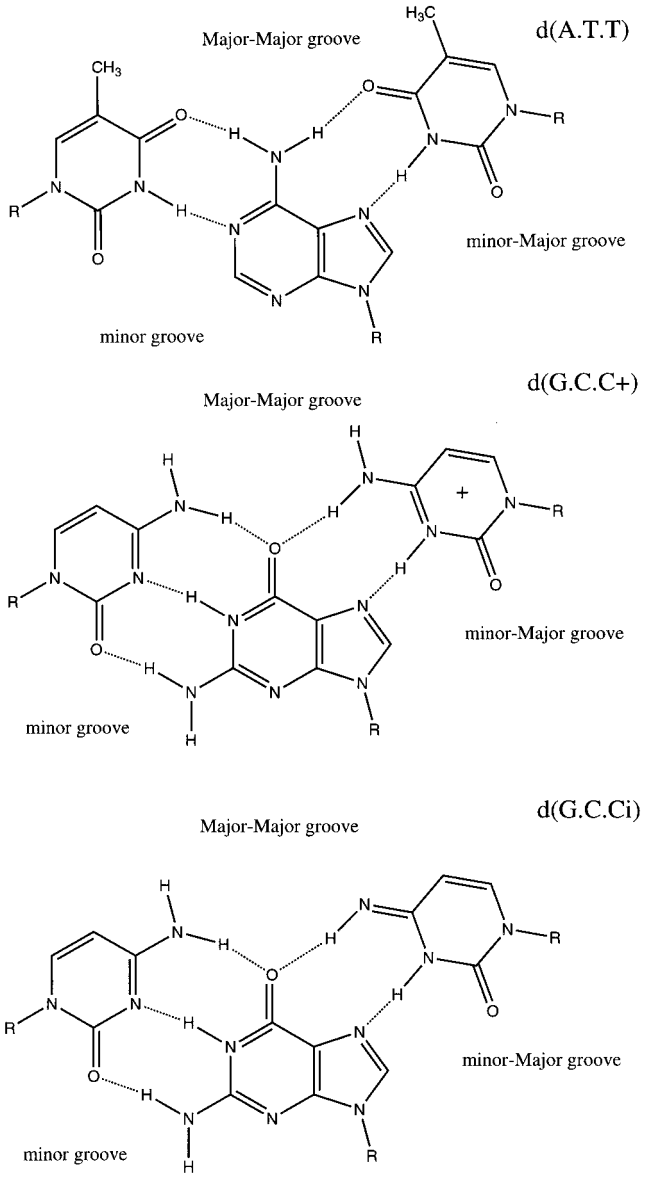

Figure 1. Schematic representation of d(A·T·T), d(G·C·C+) and d(G·C·C(imino)) trios.

like hydrogen bonds.[20,21] Note that the full interaction is not possible (see Figure 1) unless a proton exists at position N3 of cytosine. It is generally assumed that this proton arises from the protonation of the Hoogsteen cytosine. This explains the enhanced stability of d(G·C·C) containing triplexes at low pH, and apparently agrees with NMR data.[12,13,15,16,20,21] However, the protonation hypothesis cannot account adequately for the stability of d(G·C·C) containing triplexes at neutral pH, around 3 units above the pK_a of cytosine. Furthermore, this hypothesis does not explain thermodynamic data showing that during unfolding of d(G·C·C) containing triplexes, the number of protons released to the solvent is much smaller than the number of d(G·C·C) steps.[22] This situation has prompted us to use theoretical methods to investigate d(G·C·C) containing triplexes in some detail.

Theoretical methods have been used for many years in the study of nucleic acid base characteristics. Recently, high level ab initio theory was used to study the protonation and tautomerization characteristics of guanine and cytosine in polar and apolar environments.[23] The results suggested the difference in stability of imino and protonated forms of cytosine (near neutral pH) is not as large as to discard the role of imino forms in the formation of d(G·C·C) trios. Furthermore, ab initio geometry optimization and energy calculations suggested that a possible d(G·C·C(imino)) triplex was stable, even though its stability was slightly lower than that of the d(A·T·T) triplex.[23] Taken together, these results suggest that it is worth considering the possibility of d(G·C·C) containing triplexes where the Hoogsteen cytosine is not protonated, but is in the imino form (see Figure 1), a possibility that is actually explored for the so-called i-form of DNA.[24] What remains to be determined is whether this d(G·C·C(imino)) motif can be expected to lead to stable triplex helices.

MD simulations have been used in recent years to obtain dynamical views of the structure and flexibility of nucleic acid structures.[25] Among these, MD calculations with the AMBER-95 force-field[26] and accurate simulation protocols including treatment of long-range interactions have been shown to be extremely powerful for the determination of conformational characteristics of different nucleic acids,[9,27−39] including triplex structures.[9,27−32] In a recent study,[9] we used MD simulations to analyze the conformation and flexibility of the $d(A)_{10}$·$d(T)_{10}$·$d(T)_{10}$ triple helix. The results demonstrated that the technique was able to define very accurately the conformational characteristics of this triplex, even when the trajectory started from incorrect conformations. This suggests that extended MD simulations can be useful for exploring the conformational space accessible to DNA triplexes, and to detect possible unstable foldings.

In this paper we present a MD study of DNA triplexes containing the d(G·C·C) trio. We have analyzed the stability of different triplexes containing the d(G·C·C) trio in different positions. The Hoogsteen cytosine has been considered in its N3-protonated and imino forms. Simulations have allowed us to gain insight into the role of protonation and tautomerization in the stability of triplexes, as well as on the structure of DNA triplexes containing the d(G·C·C) trio.

(20) Rajagopal, P.; Feigon, J. *Biochemistry* **1989**, *28*, 7859.
(21) Rajagopal, P.; Feigon, J. *Nature* **1989**, *339*, 637.
(22) Plum, G. E.; Breslauer, K. J. *J. Mol. Biol.* **1995**, *248*, 679.
(23) Colominas, C.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **1996**, *188*, 6811.
(24) (a) Sponer, J.; Leszczynski, J.; Vetter, V.; Hobza, P. *J. Biomol. Struct. Dyn.* **1998**, *13*, 695. (b) Spackova, N.; Berger, I.; Egli, M.; Sponer, J. *J. Am. Chem. Soc.* **1998**, *120*, 6147.
(25) Laaksonen, A.; Nilsson, L. G.; Johnsson, B.; Teleman, O. *Biochemistry* **1989**, *129*, 175.
(26) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K.; Fergurson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
(27) Van Vlijmen, H.; Rmae, G. L.; Pettitt, M. *Biopolymers* **1990**, *30*, 517.
(28) Mohan, V.; Smith, P. E.; Pettit, M. *J. Am. Chem. Soc.* **1993**, *115*, 9297.
(29) Mohan, V.; Smith, P. E.; Pettit, M. *J. Phys. Chem.* **1993**, *97*, 12984.
(30) Weerasinghe, S.; Smith, P. E.; Mohan, V.; Cheng, Y.-K.; Pettit, M. *J. Am. Chem. Soc.* **1995**, *117*, 2147.
(31) Weerasinghe, S.; Smith, P. E.; Pettit, M. *Biochemistry* **1995**, *34*, 16269.
(32) Shields, C.; Laughton, C. A.; Orozco, M. *J. Am. Chem. Soc.* **1998**, In Press.
(33) York, D. M.; Yang, W.; Lee, H.; Darden, T.; Pedersen, L. G. *J. Am. Chem. Soc.* **1995**, *117*, 5001.
(34) Cheetham, T. E.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4193.
(35) Cheetham, T. E.; Kollman, P. A. *J. Mol. Biol.* **1996**, *259*, 434.
(36) Yang, L.; Pettit, M. *J. Phys. Chem.* **1996**, *100*, 2564.
(37) Cheatham, T. E.; Kollman, P. A. *J. Am. Chem. Soc.* **1997**, *119*, 4805.
(38) Cieplak, P.; Cheatham, T. E.; Kollman, P. A. *J. Am. Chem. Soc.* **1997**, *119*, 6722.
(39) Spector, T. I.; Cheatham, T. E.; Kollman, P. A. *J. Am. Chem. Soc.* **1997**, *119*, 7095.

Methods

The stability of DNA triplexes containing d(G·C·C) trios was studied by MD simulations. Three triplex sequences were considered: d(GAGAGAGAGA), d(AGGGAAGGGA), and d(GAAGGGAAGA). For d(GAAGGGAAGA) three different structures were created by taking all the Hoogsteen cytosines in their protonated or imino forms, as well as in a mixed situation, where the cytosine bound to guanine 5 was in the imino form, while the others were treated as protonated. The d(GAGAGAGAGA) sequence was studied with all cytosines in their protonated and imino forms. The last sequence (d(AGGGAAGGGA)) was studied considering the Hoogsteen cytosines bound to guanines 3 and 8 in their imino forms, while cytosines at positions 2, 4, 7, and 9 were protonated.

The six structures were generated by molecular modeling from the MD-averaged conformation of the $d(A\cdot T\cdot T)_{10}$ triplex.[9] In all cases the structures were immersed in a box containing 4141 water molecules and the minimum number of sodium counterions necessary for neutralization of the system. The final systems contained around 13400 atoms. Counterions were placed in the regions of most electronegative potential, as described previously.[9] The starting systems were minimized, heated, and equilibrated over 120 ps with use of harmonic restrains to avoid artifactual unfolding of the triplex structures.[39] Once the equilibration process was finished, the six structures were followed through a 1.0 ns unrestrained MD simulation.

Simulations were performed in the isothermic isobaric ensemble ($P = 1$ atm, $T = 300$ K). Periodic boundary conditions and the Particle-Mesh-Ewald algorithm were used.[41−42] All bonds were constrained at their equilibrium distances by using SHAKE,[43] which allowed us to use a 2 fs integration time. The AMBER-95 force-field[26] was used to represent the DNA, and the TIP3P[44] model was used to represent water molecules. Atomic charges for imino and protonated cytosines were obtained by using the standard RESP procedure,[45] and are available from the authors on request. MD simulations were carried out with the AMBER4.1 suite of computer programs.

For trajectories leading to stable conformations, MD-averaged structures were obtained by averaging 1000 snapshots collected over the last 500 ps. Geometry averaged structures were relaxed by restrained-geometry optimization following the procedure described in detail elsewhere.[9]

Poisson−Boltzmann (PB) calculations were performed to predict the relative stability of protonated and imino forms of cytosine in the triplex, and to evaluate the pK_a shifts due to DNA structure. Calculations were carried out by following the procedure developed by Honig and coworkers,[46−48] using an ionic strength of 1 M, a dielectric constant of 2 for the DNA and 80 for the solvent, RESP charges, and all the default values in DELPHI.[49] In all cases the standard B-type structure previously found for the d(A·T·T) triplex[9] was used.

Molecular Interaction Potentials (MIPs) were used to obtain a picture of the recognition characteristics of triplex DNA. The classical probe molecule was a proton with the Lennard-Jones properties of a TIP3P water molecule. The OPLS force-field[50] was used to represent the Lennard-Jones properties of the DNA in MIP calculations. The electrostatic term was determined with the PB method.[46]

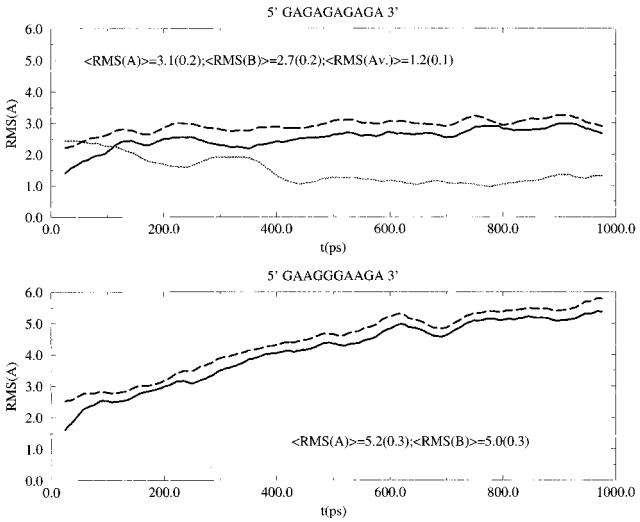

Figure 2. Root-mean-square deviation along the trajectory for triplexes where all the Hoogsteen cytosines are protonated. Reference structures are as follows: A-type triplex (dashed line), B-type triplex (solid line), and the average structure obtained during the last 500 ps of the trajectory (gray line). The root-mean-square deviation from the average structure is computed only when the trajectory has converged.

Results and Discussion

After equilibration all the trajectories look reasonably stable in terms of potential energy, temperature, density, and other macroscopic properties, which suggests a reasonable equilibration of the trajectories for all starting conformations. Analysis of the trajectories shows the existence of two situations: (i) simulations leading to stable triplexes and (ii) simulations leading to unfolding of the triplex, or to very distorted structures. As a first approximation, we can consider that triplexes whose structures are preserved through the 1.1 ns of simulation are stable (for AMBER-95 force-field). On the contrary, we consider that structures which are unfolded at the end of the trajectory are unstable under conditions similar to those considered in the simulation. Note that we discuss here only the intrinsic ability of molecules to form triplex structures at low ionic strength and neutral pH; we do not consider the intrinsic stability of the imino or protonated forms of cytosine (studied in detail in refs 23 and 24). We have to emphasize that caution is therefore necessary when translating our theoretical results to certain experimental conditions, where other factors might lead to stabilization/destabilization of triplexes.

Stability of Triplexes Containing Protonated Cytosines. The first issue that we address is the stability of DNA triplexes containing several protonated cytosines contiguous in the third strand. This can be studied by analysis of the trajectory of the d(GAAGGGAAGA) triplex model in which all the Hoogsteen cytosines are protonated. Results in Figure 2 show that the structure is unstable (see also Figure 3), and leads in 1 ns to unfolded conformations, where the helix is dramatically distorted. Thus, the average root-mean-square deviation from both the B and A triplex conformations is more than 5 Å, and the final root-mean-square deviation is almost 6 Å. Inspection of the final structure (Figure 3) clearly illustrates the large distortion

(40) The equlibration of the systems consists of the followong consecutive steps: (i) generation of the triplex, (ii) minimize water, (iii) 10 ps MD ($T = 100$ K) of water, (iv) 10 ps MD ($T = 100$ K) of system with restrains in DNA ($K = 100$ kcal/(mol Å^2)), (v) 10 ps MD ($T = 300$ K) of system with restrains in DNA ($K = 100$ kcal/(mol Å^2)), (vi) 25 ps MD ($T = 300$ K) of system with restrains in DNA ($K = 50$ kcal/(mol Å^2)), (vii) 25 ps MD ($T = 300$ K) of system with restrains in DNA ($K = 25$ kcal/(mol Å^2)), (viii) 25 ps MD ($T = 300$ K) of system with restrains in DNA ($K = 10$ kcal/(mol Å^2)), (ix) 25 ps MD ($T = 300$ K) of system with restrains in DNA ($K = 5$ kcal/(mol Å^2)). The final system was the starting point of 1.0 ns of unrestrained MD simulation ($T = 300$ K).

(41) Darden, T. E.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(42) Essmann, V.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.

(43) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(44) Jorgensen, W. L.; Chandresekhar, J.; Madura, J.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(45) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(46) Gilson, M. K.; Honig, P. H. *Nature* **1987**, *330*, 84.

(47) Gilson, M. K.; Honig, P. H. *Proteins* **1988**, *4*, 7.

(48) Honig, B.; Sharp, K.; Yang, A. *J. Phys. Chem.* **1993**, *97*, 1101.

(49) Delphi Computer Program; BIOSYM: San Diego, CA, 1994.

(50) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *100*, 1657.

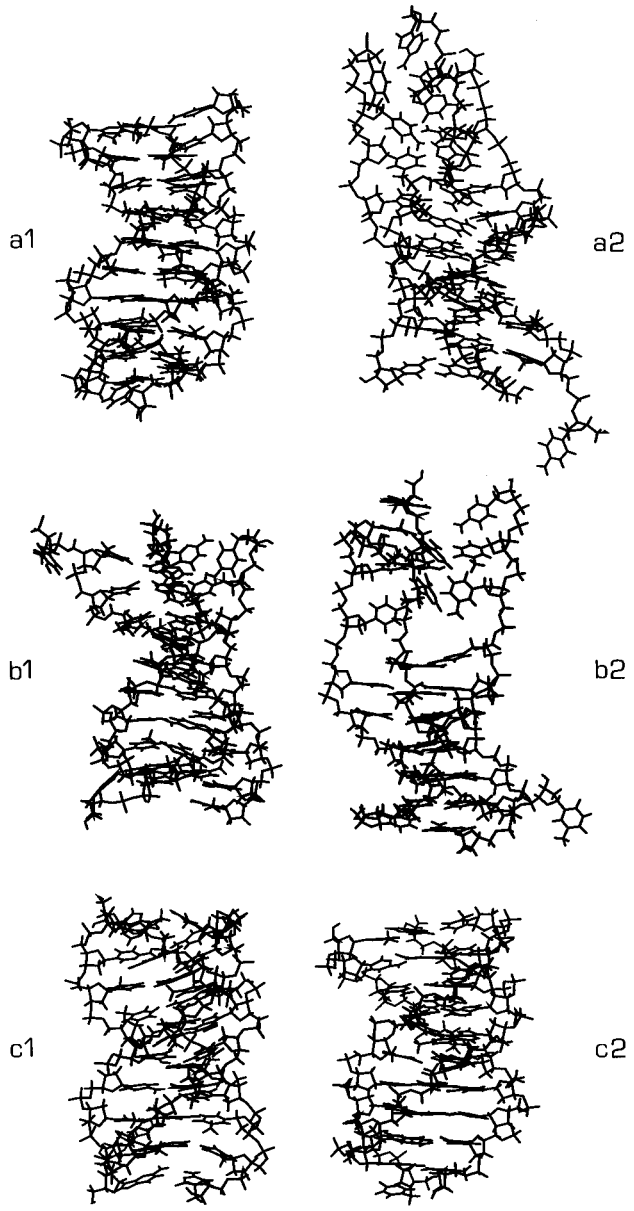

Figure 3. Representation of the structure at the end of the simulation for the different triplexes studied here: **a1**, d(GAGAGAGAGA) with all Hoogsteen cytosines protonated; **a2**, d(GAAGGGAAGA) with all Hoogsteen cytosines protonated; **b1**, d(GAGAGAGAGA) with all Hoogsteen cytosines in the imino form; **b2**, d(GAAGGGAAGA) with all Hoogsteen cytosines in the imino form; **c1**, d(GAAGGGAAGA) with mixing of imino and protonated Hoogsteen cytosines, and **c2**, d(AGGGAAGGGA) with mixing of imino and protonated Hoogsteen cytosines.

in the triplex developed during the trajectory. We hypothesize that strong electrostatic repulsive interactions between protonated cytosines in the central d(G·C·C) steps lead to major structural deformation in this region, which is transmitted to the rest of the helix.

The second issue explored is the stability of triplexes containing protonated cytosines which are separated by at least one neutral base trio. This is illustrated by the d(GAGAGAGAGA) sequence, where all the Hoogsteen cytosines were modeled as protonated. Results in Figure 2 demonstrate that in this case the triplex structure is maintained without dramatic distortions over the full 1 ns of the MD trajectory. The equilibrated triplex is slightly different from that obtained for the d(A·T·T) sequence, since it has some characteristics resembling the A-type triplex model (see Figures 2 and 3 and Table 1). However, it is clear that the resulting structure is stable (see Figure 3), at least over the time scale of the simulation. This suggests that the stability of a d(GAGAGAGAGA) triplex is compatible with the existence of a large proportion of protonated Hoogsteen cytosines. Note however that our results do not exclude the possibility that some of the Hoogsteen cytosines are not protonated.

PB calculations (see Figure 4) suggest that to move a protonated cytosine from water to position 5 of the triplex d(GAGAGAGAGA) is around 11 kcal/mol easier than to perform the same process in the triplex d(GAAGGGAAGA). This free energy value should be taken with caution considering the uncertainties existing in PB calculations, but it clearly supports that a protonated cytosine is more stable when it is surrounded by neutral pyrimidines than when it is contiguous to protonated cytosines. This can be rationalized considering that the C+↔C+ repulsion is clearly smaller for the d(GAGAGAGAGA) sequence than for the d(GAAGGGAAGA) triplex.

In summary, MD and PB simulations suggest that Hoogsteen cytosines in DNA triplexes can all be protonated only when the d(G·C·C) trios are not contiguous. These results are in accord with the experimental observation that triplexes of the type d(GAGAGA) are very stable.[20,51] However, we must do more if we are to explain why triplexes based on polyguanines, e.g. d(GGGGGG), are stable under experimental conditions,[52,53] even when their stability is smaller[51] than that of triplexes of the type d(GAGAGA). This issue is discussed below.

Stability of Triplexes Containing Imino Cytosines. Inspection of Figure 1 shows that to have a stable (fully hydrogen bonded) G·C·C triplex, a reasonable alternative to protonation of cytosines is to assume that Hoogsteen cytosines are in the imino form (see Figure 1). To determine the role of imino tautomers in DNA triplex stability we computed additional MD trajectories for each of the d(GAAGGGAAGA) and d(GAGAGAGAGA) triplexes, where all the Hoogsteen cytosines were considered in their imino form. The average root-mean-square deviations (see Figure 5) from the classical triplex models are between 4 and 5 Å for both sequences. The room-mean-square deviations at the end of the trajectories are greater than 6.0 (d(GAAGGGAAGA)) and 5 (d(GAGAGAGAGA)) Å with respect to any canonical structure of the triplex DNA.

Figure 3 demonstrates how both triplexes unfold through the course of the simulations due to the separation of the Hoogsteen strand. This process clearly has its origins in the high flexibility and moderate stability of the H-bonding between guanine and imino cytosine as predicted from ab initio calculations.[23] In view of the stability of related simulations and the consistency in the MD protocols, we are confident then that the unfolding of the triplexes detected in both trajectories is not an artifact, but relates directly to the reduced stability, and high flexibility, of G−C(imino) pairings.

PB calculations suggest that the work necessary to move an imino cytosine from water to the d(GAGAGAGAGA) triplex is 0.4 kcal/mol larger than the work necessary for the transfer of a protonated cytosine (see Figure 4). This value combined with the experimental difference of stability (pH 7) between protonated and imino forms in water (1.7 kcal/mol from ref 54) leads to a total free energy difference around 2.1 kcal/mol

(51) Roberts, R. W.; Crothers, D. M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4320.
(52) Lyamichev, V. I.; Frank-Kamenetskii, M. D.; Soyfer, V. N. *Nature* **1990**, *344*, 568.
(53) Mirkin, S. M.; Frank-Camenetskii, M. D. *Annu. Rev. Biophys. Biomol. Struct.* **1994**, *23*, 541.

Table 1. Values of the Different Helical Parameters for the Average Conformations and MD-Averaged Values (in parentheses)[a]

parameters	GAGAGAGAGA	AGGGAAGGGA	GAAGGGAAGA
X-displacement, Å	−3.3 Å (−3.0 ± 0.6 Å)	−5.0 Å (−3.7 ± 0.6 Å)	−1.2 Å (−1.1 ± 0.6 Å)
tip, deg	14.9° (12.4 ± 5.2°)	12.8° (9.6 ± 5.9°)	−1.3° (1.3 ± 4.9°)
inclination, deg	12.8° (12.0 ± 5.2°)	5.6° (−2.9 ± 6.1°)	3.3° (6.4 ± 4.3°)
slide, Å	0.0 Å (0.0 ± 0.2°)	−1.2 Å (−1.2 ± 0.1°)	0.1 Å (0.3 ± 0.2°)
shift,	−0.4 Å (−0.4 ± 0.2°)	−0.9 Å (−0.89 ± 0.1°)	−0.3 Å (−0.2 ± 0.2°)
rise, Å	3.4 Å (3.4 ± 0.1 Å)	3.5 Å (3.5 ± 0.1 Å)	3.4 Å (3.3 ± 0.1 Å)
twist, deg	31.7° (32.5 ± 1.6°)	31.1° (32.4 ± 1.5°)	31.4° (33.9 ± 1.8°)
roll, deg	2.8° (2.1 ± 1.8°)	0.3° (0.4 ± 1.3°)	1.5° (2.1 ± 1.8°)
tilt, deg	2.2° (2.0 ± 2.0°)	0.8° (0.9 ± 1.2°)	1.4° (1.6 ± 1.6°)
stagger, Å	0.0 Å (0.0 ± 0.2 Å)	0.0 Å (0.0 ± 0.1 Å)	−0.4 Å (−0.3 ± 0.3 Å)
stretch, Å	0.0 Å (0.1 ± 0.2 Å)	0.2 Å (0.2 ± 0.1 Å)	0.1 Å (0.2 ± 0.1 Å)
shear, Å	0.0 Å (0.0 ± 0.2 Å)	−0.1 Å (−0.2 ± 0.2 Å)	0.1 Å (0.0 ± 0.1 Å)
opening, deg	1.9° (1.9 ± 3.5°)	1.4° (1.6 ± 1.9°)	4.8° (4.3 ± 2.5°)
buckle, deg	6.3° (6.0 ± 5.2°)	−3.8° (−3.9 ± 3.0°)	1.4° (2.1 ± 6.4°)
propeller twist, deg	5.0° (5.2 ± 4.5°)	−7.8° (−6.9 ± 4.4°)	1.8° (0.3 ± 3.8°)
phase, deg	143.7° (148.1 ± 33.9°)	117.6° (116.7 ± 26.3°)	145.70 (149.8 ± 21.4°)
amplitude, deg	41.3° (42.4 ± 6.3°)	39.9° (41.7 ± 5.9°)	42.0° (43.1 ± 6.2°)
MM groove	16.1 (15.8 ± 0.8 Å)	15.9 (16.0 ± 1.1 Å)	14.4 (14.6 ± 1.4 Å)
m groove	12.6 (12.3 ± 0.8 Å)	11.9 (12.0 ± 0.5 Å)	12.1 (12.0 ± 0.8 Å)
mM groove	8.5 (8.5 ± 0.3 Å)	8.5 (8.7 ± 0.3 Å)	8.7 (9.0 ± 0.3 Å)

[a] Standard deviations (SD) of the MD-averaged parameters are given. All the helical values refer to the WC strands.

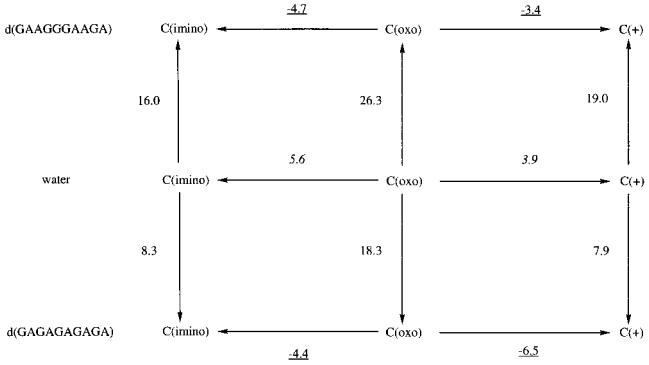

Figure 4. Thermodynamic cycles used to compute free energy differences from Poisson−Boltzmann calculations (see text). Plain numbers are from PB calculations, italic numbers are from experimental estimates (see text), underlined numbers are derived by closing thermodynamic cycles.

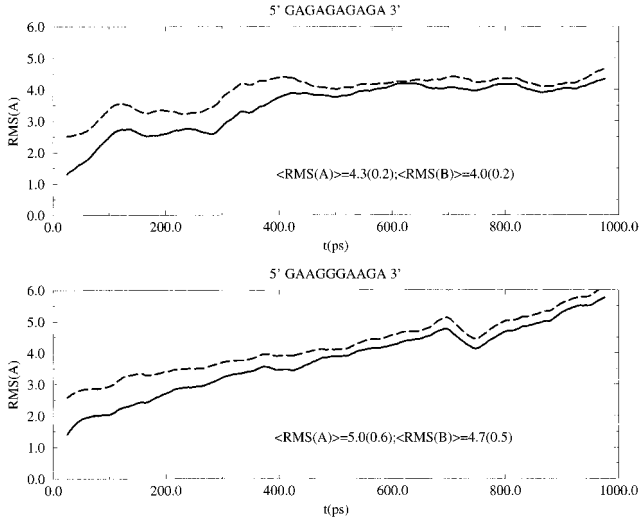

Figure 5. Root-mean-square deviation along the trajectory for triplexes where all the Hoogsteen cytosines are in the imino form. Reference structures are the following: A-type triplex (dashed line) and B-type triplex (solid line).

favoring the presence of protonated cytosine in front of the imino form in the d(GAGAGAGAGA) triplex.

In summary, it is clear that, even aside from the low intrinsic stability of the imino form of cytosine, the stability of triplexes containing the d(G·C·C) base trio cannot be based exclusively on the existence of Hoogsteen cytosines in their imino form. Comparison of the protonated and imino trajectories for d(GAGAGAGAGA) triplex (the former stable, the latter not) strongly suggests that this triplex should have a larger percentage of protonated than imino cytosines, even at pH values which are clearly above the pK_a of cytosine. Comparison of the protonated and imino trajectories for the d(GAAGGGAAGA) triplex (both unstable) strongly suggests that neither protonated nor imino forms can explain by themselves the stability of similar triplexes, which are known experimentally to exist under a wide range of conditions.[17−19] It seems clear that the only possibility is to assume that protonated and imino forms are combined in such a way that the triplex becomes stable.

Stability of Triplexes Based on a Combination of Protonated and Imino Forms of Cytosine. Results presented above suggest that d(G·C·C) containing triplexes can be stable only if (i) protonated cytosines should be separated by at least one neutral base trio and (ii) there are a small number of noncontiguous imino cytosines. That is to say, alternate disposition of imino and protonated forms of C seems the best approach to guarantee the stability of triplexes containing homoguanine tracts. Thus, we generated the d(GAAGGGAAGA) triplex where cytosines attached to guanines 1, 4, 6, and 9 were modeled as protonated, while the central cytosine 5 was modeled in the imino form. To check if the C(+)C(imino)C(+) motif can be repeated several times in a triplex without loss of stability, we also studied the d(AGGGAAGGGA) triplex, where the cytosines 2, 4, 7, and 9 were taken as protonated, while 3 and 8 were assumed in the imino form.

The general triplex structure of d(GAAGGGAAGA) with imino cytosine at position 5 is maintained during the trajectory, as shown in Figures 3 and 6. Inspection of the root-mean-square deviation plots in Figure 6 suggests that the triplex reaches a stable conformation after 300 ps (root-mean-square deviation from average conformation of 1.4 Å). The conformation that is sampled during most of the trajectory exhibits room-mean-square deviations of 3.4 Å from A-type triplex and 3.3 Å from B-type triplex. This suggests that the combination of protonated

(54) Dreyfus, M.; Beusande, O.; Dodin, G.; Dubois, J. E. *J. Am. Chem. Soc.* **1976**, *93*, 6338.

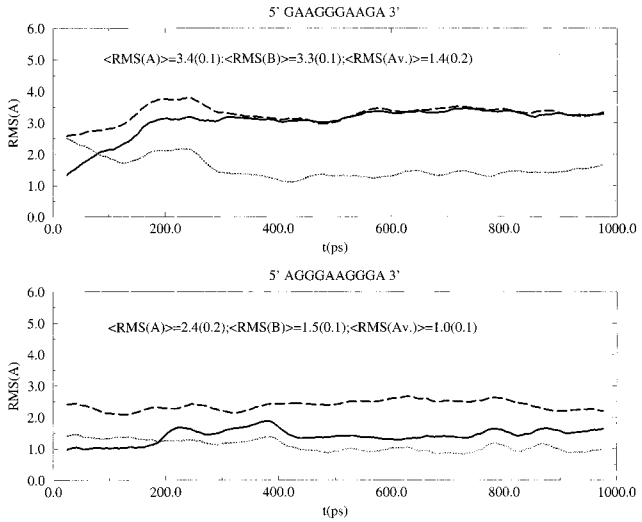

Figure 6. Root-mean-square deviation along the trajectory for triplexes where imino and protonated forms of the Hoogsteen cytosines coexist. Reference structures are the following: A-type triplex (dashed line), B-type triplex (solid line), and the average structure obtained during the last 500 ps of the trajectory (gray line). The root-mean-square deviation from the average structure is computed only when the trajectory has converged.

and imino cytosines leads to a triplex, which is stable during 1 ns of MD, and which has a general geometry intermediate between that of an A-type and B-type triplex (see below for a more detailed description). Note that such stability during 1 ns of MD was not obtained when all the third strand cytosines were modeled in either imino or protonated form, which regardless of the limited duration of the trajectory strongly supports the higher stability of mixed protonated/imino combinations with respect to all-protonated or all-imino combinations.

The repetition of C(+)-C(imino)-C(+) motifs does not cause problems, as the MD trajectory of the d(AGGGAAGGGA) triplex reveals, where this motif is present at positions 2−4 and 7−9. Figure 6 demonstrates the equilibration of the triplex (root-mean-square deviation from average structure of 1.0 Å) around a conformation that is very similar (root-mean-square deviation 1.5 Å) to a standard B-type triplex as that found in d(A·T·T) triplexes. Inspection of the final structure (see Figure 3) confirms that the final structure resembles very closely the canonical B-type triplex model.

PB calculations suggest that the mutation from a protonated cytosine at position 5 of the triplex d(GAAGGGAAGA) to an imino cytosine leads to a stabilization free energy of 1.3 kcal/mol. This value should be taken with caution, but agrees with the hypothesis that a Hoogsteen cytosine surrounded by protonated cytosines slightly prefers to be in the imino form.

PB calculations (Figure 4) allow us to gain additional insight into the changes in acidity induced by the triplex DNA environment. Figure 4 clearly demonstrates that the oxo form of cytosine is the most stable species in solution, but the least stable form in the DNA environment (for the Hoogsteen arrangement studied here). Thus, PB results suggest that the proton that is interchanged with the solvent in the DNA is not the proton at N3, but the amino proton at N4 (this hypothesis agrees with chemical intuition which suggests that N3 cannot interchange a proton while it is bound to N7 of guanine). In other terms, the pK_a measured in water refers to the equilibrium between the oxo and the N3-protonated forms, while the pK_a in the triplex refers to the equilibrium between the imino and the N3-protonated forms. Bearing this in mind, DNA-induced shifts in the pK_a of 4.4 (d(GAGAGAGAGA)) and 1.9 units (d(GAAGGGAAGA)) can be determined from results in Figure 4.

In summary, MD calculations reveal that postulating the combination of protonated and imino forms of cytosine fits with the stability of homoguanine tracts in triplex DNA structures, while restricting consideration to either one of these two forms of cytosine alone cannot. The ratio of protonated/imino cytosines can be different from the ratio of 2:1 considered here, since it should obviously depend on the pH. Thus, a large percentage of cytosines should be protonated at low pH, while the role of imino tautomers is probably more important at higher pH. Note that when the ratio protonated/imino is very small (as should happen at high pH) the triplex is expected to be unstable due to the intrinsic poor binding between guanine and the imino tautomer of cytosine, which explains then the well-known loss of stability of G-rich triplexes at high pH. Note that the model implies that only a portion of the d(G·C·C) trios are protonated (the other cytosines are in the imino form), in accordance with the difference found experimentally between the number of d(G·C·C) trios and the number of "extra" protons in the triplex. The model explains also very clearly the surprising stability of d(G·C·C) containing triplexes at neutral pH.[55]

Most of the structural information of DNA triplexes has been derived from NMR experiments. The NMR data for d(G·C·C)-containing triplexes agree well with protonated cytosine hypothesis, while signals that would correspond to imino forms are not detected. These experimental results have been used to suggest that all the Hoogsteen cytosines are protonated. However, detailed inspection of the literature shows not only that these NMR experiments[13,15,16,20,21] are done at very acidic pH (typically around 4−5), but also that the structures studied are of the type d(GAGA), and not of the type d(GGGG). Note that the majority existence of protonated forms for triplex of the type d(GAGA) is actually predicted from our results, even at pH greater than the pK_a of cytosines. The hypothesis that imino (or other neutral) forms can contribute to the stability of homoguanine containing triplexes should encourage new experimental studies in this field

Classical MD simulations are done in such a way that imino and protonated cytosines are not able to interchange. However, under standard experimental conditions we should expect some interchange between imino and protonated forms. Figure 7 shows that the amino/imino groups of contiguous protonated/imino cytosines are very close (around 3.3 Å on average). Out-of-plane bending of the amino group of cytosine can put one proton of the amino group of the protonated cytosine at a very close distance (less than 2.5 Å) from the nitrogen atom of the imino group (see Figure 7). This close distance should favor a fast intramolecular interchange of the proton between imino and protonated forms of the base, without involving water molecules.

Analysis of the Triplex Structures. MD trajectories of stable triplexes were used to obtain information on the conformational characteristics of d(G·C·C)-containing triplexes. Three trajectories led to stable triplexes: (i) d(GAGAGAGAGA), where all Hoogsteen cytosines were protonated, (ii) d(GAAGGGAAGA), where all Hoogsten cytosines but that bound to guanine 5 were protonated, and (iii) d(AGGGAAGGGA), where

(55) Another nonexplored possibility of neutral d(G·C·C) is the formation of a wobble Hoogsteen-like pairing between G and C. This interaction can be stabilized by a single H-bond, and accordingly it cannot be expected to contribute significantly to the stability of the triplex. Whether or not this type of interaction might exist at neutral pH will be investigated in future works.

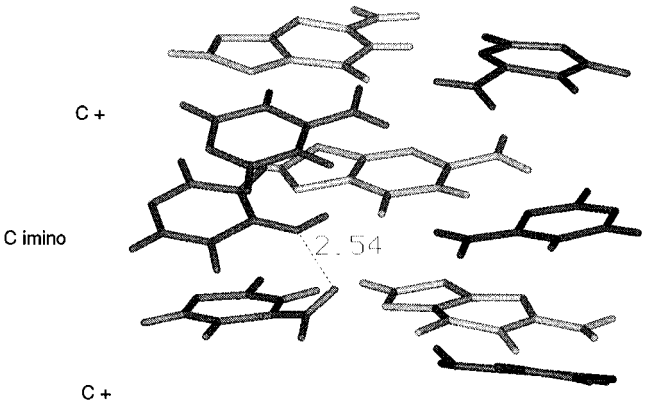

Figure 7. Representation of one arrangement of three cytosines (central one imino, top and bottom in the protonated form) obtained during the trajectory for the d(GAAGGGAAGA) triplex.

all Hoogsteen cytosines but those bound to guanines 3 and 8 were protonated. The most important characteristics of these three structures are displayed in Table 1, where MD averaged values and values for the MD-averaged structure (see Methods) are displayed. It is worth noting that in general MD-averaged values and values for the MD-averaged structure are very similar, which confirms the stability of the trajectories, as well as the goodness of the approach used to obtain the MD-averaged conformation.

The three structures share several general characteristics. Thus, all are right-handed triplexes, and the stacking and H-bond pattern is that expected for a triple helix, with very small values of stretch, shear, and stagger. Slide, shift, roll, and tilt are quite small, in agreement with what is expected from experimental models of the triple helix. As expected, the largest distortions occur at the extremes, where the effect of water is more intense. The three helices show helical twist values around 31−34° (Watson−Crick strands). This range of values agrees well with fiber diffraction data for the d(A·T·T) triplex (average twist of 30° from ref 6). MD results agree also very well with recent NMR data, suggesting twist values in the range 29−31° for d(AGGAAGG), d(GAATAGG), and d(AGAGAGAA) triplexes.[12−14,16] There is also good agreement with previous MD simulations for the d(A·T·T) triplex (twist values around 29° from ref 9).

The three trajectories lead to values of rise between 3.3 and 3.5 Å, which compare well with those obtained by MD for the poly-adenine based triplex (3.5 Å). There is also agreement with fiber diffraction results for the poly-adenine based triplex (3.3 Å from ref 6), as well as with the NMR values for the sequences noted above: 3.3 (13), 3.4 (14), and 3.1 Å.[16] This agreement suggests that the key characteristics of the helices found by MD simulations are correct, since they match the helical parameters (twist and rise), which are accurately determined by experimental techniques. Furthermore, the similarity of the MD estimates of twist and rise obtained for different sequences strongly suggests that the main characteristics of the pyrimidine motif DNA triplex are quite independent of the sequence.

Inspection of other helical properties shows that the Watson−Crick base pairs are in general coplanar, but a maximum of less than 10° loss of coplanarity is allowed by the structure. Buckle is especially important for the d(GAGAGAGAGA) sequence, and propeller twist for d(GAGAGAGAGA) and d(AGGGAAGGGA) sequences. These structural details are difficult to compare against experimental data, since in most cases planarity of the base pairs and ideality of H-bonding scheme are at least partly assumed in the refinement of the models. In all cases negative X-displacements (from −1 to −5 Å) are found. These values are in reasonable agreement with X-displacement values found by fiber diffraction (from −3 to −3.5 Å) and NMR techniques (between −2 and −3 Å from NMR studies) for different sequences. Finally, base pairs are quite perpendicular to the helix axis (see also Figures 3), but nonnegligible positive inclination values are found, specially for the d(GAGAGAGAGA) sequence, which represents a deviation from standard B-type models of the triplex (small inclination), and an approximation to A-type models (inclination around 5°; note the typological error in ref 9).

All the triplexes exhibit deoxyribose puckerings in the *South* or *East* regions, in agreement with the B-type structure. In general the sugar rings are much more rigid in the Watson−Crick strands, where most of the riboses are typically in the *South* region, while the riboses in the Hoogsteen strand are much more flexible. The sugar puckering behavior found in these MD simulations agrees very well with most of the available NMR data, which suggests that most riboses are in the *South* and *East* regions,[10−16] and contrasts with fiber diffraction data,[6] and early NMR studies,[56] which suggested puckerings in the C3′-endo region.

The presence of the third strand divides the major groove of the duplex in two grooves of different size: the minor part of the major groove (mM; M1 groove according to Petitt's nomenclature, Crick−Hoogsteen groove according to Hélène's nomenclature), and the major part of the major groove (MM; M2 according to Petitt's nomenclature, Watson−Hoogsteen groove according to Hélène's nomenclature). The MM groove is very wide (around 15−16 Å for the shortest interstrand P−P distance), and shows a large variation (a standard fluctuation from the average of more than 1 Å). The size of the MM groove is only slightly smaller (around 1−2 Å) than that found previously in MD calculations of the d(A·T·T) triplex, and seems to agree well with NMR structures, even though no quantitative NMR estimates of the size of this groove are available for comparison. In any case, it is the disagreement with models generated from fiber diffraction data.

The width of the MM groove alone is similar to that of the major groove of B-DNA (around 17 Å following the same criteria for measurements). This means that during the binding of the third strand to duplex DNA, the major groove of the duplex is greatly widened (see ref 9). This is of major biological significance, since it suggests that the third strand is not completely blocking the major groove and so precluding any interaction with DNA-binding proteins. On the contrary, our results strongly suggest that the MM groove is wide enough to function as a recognition site for proteins, as recent experimental data suggest.[57]

The minor groove measures about 12 Å in width (P−P distance) for the three structures. This value, which is similar to that found in MD calculations of the d(A·T·T) triplex, agrees well with a rough estimate of 13 Å obtained from NMR data of the d(GAATAGG) triplex, and disagrees with the values obtained from the A-type models of the triplex (around 16 Å). The width of the minor groove of the triplex is almost identical to that found in duplex DNA, although in cross-section the groove shapes are rather different, suggesting that the minor groove of the triplex could be a target for specific minor groove

(56) Umemoto, U.; Sarma, M. H.; Gupta, G.; Luo, J.; Sarma, R. H. *J. Am. Chem. Soc.* **1990**, *112*, 4539.

(57) (a) Guieysse, A. L.; Praseuth, D.; Hélène, C. *J. Mol. Biol.* **1997**, *267*, 289. (b) Jiménez-García, E.; Vaquero, A.; Espinás, M. L.; Soliva, R.; Orozco, M.; Bernués, I.; Azorín, F. *J. Biol. Chem.* In press, 1998.

binders. Finally, the mM groove is narrow (around 9 Å) and quite shallow. Indeed, the width of the mM groove agrees well with NMR estimates by Patel's group (8.9 Å from ref 13), and with fiber difraction data (around 7.5 Å from the structures derived from ref 6).

Molecular Interactions of Triplex Structures. The steric, electrostatic, and hydrogen-bonding characteristics of the grooves in the triplex render them very rich in information and so can be recognition sites for specific drugs or proteins. To analyze this point in more detail, we computed the regions of higher water density (for previous analysis of water density see refs 9, 13, 28, and 29) around the three stable triplexes (data available as Supporting Information), as well as the regions of more negative MIP triplexes (data available as Supporting Information). Both types of calculations provide information on the reconition properties of the triplex grooves.

Water is less structured around d(G·C·C)-based triplexes than around d(A·T·T)-based triplexes. This is detected as lower values of the apparent water density around the d(G·C·C)-based triplexes. Thus, for the d(A·T·T) triplex we found quite large regions in the minor groove where the water density was 8−9 times the density of pure water, while for the d(G·C·C)-containing triplexes the biggest peaks of water density are around 5−6 times the density of pure water. The poorer hydration of d(G·C·C)-based triplexes is probably related to the lower negative charge in these triplexes.

The distribution of water around the d(G·C·C)-based triplexes is very similar to that found in d(A·T·T)-based triplexes. The minor groove is generally the region with the larger apparent water density, it being possible to detect "spines of hydration" for all the structures. The presence of guanines reduces but does not eliminate such regions of highly ordered water in the minor groove. The MM groove is well solvated in general (especially well for the d(GAAGGGAAGA) triplex), but in any case the apparent water density is lower than that found for the minor groove. Finally, the mM groove is generally very well solvated, as found for the d(A·T·T)-based triplex. However, we found that if this goove narrows (as happens in some regions of the d(GAGAGAGAGA) and d(AGGGAAGGGA) triplexes) then hydration is reduced as a consequence.

The general trends in the MIP distribution in d(G·C·C)-containing triplexes are quite similar to those found for pure d(A·T·T)-based triplexes. In all cases the best target for cations (most negative MIP values) is the minor groove, as previously found for the d(A·T·T) triplex. In general the region of negative MIP is wider for d(A·T·T) regions, and narrower for d(G·C·C) trios, especially when they are protonated. However, the differences between d(A·T·T) and d(G·C·C) trios are small, and the region of negative MIP is continuous along the length of the minor groove for all three structures. The mM groove shows a high affinity for cations in all the structures, but the regions of negative MIP are interrrupted in all cases for the steps containing protonated cytosines. Finally, the MM groove represents, in general, a less favored binding site for cations than the other grooves, as previously found for the d(A·T·T)-based triplex.[9]

In summary, structural, hydration, and MIP analyses suggest a recognition pattern for d(G·C·C)-containing triplexes that is not very different from that previously obtained for a pure d(A·T·T)-based triplex. The minor groove of the d(G·C·C)-based triplexes studied here is in general similar to the minor groove of B-form duplex DNA in terms of width, hydration, and MIP, even though the actual shape of the groove is slightly different from that of a B-DNA.[9] This means that the minor groove of d(G·C·C)-containing triplexes might recognize molecules able to interact with the minor groove of d(G·C) sequences of B-DNA, though perhaps with less affinity. The mM groove is narrow, but its size does not preclude completely an interaction with small polar molecules. Finally, the MM groove of the triplex is very wide, in fact similar in dimensions and reactive characteristics to the major groove of a B-DNA duplex. This suggests that it should be able to interact with large polar molecules, such as the DNA-binding motifs of specific proteins.

In summary, these results suggest a general similarity in terms of recognition properties between the d(A·T·T) triplex and the different d(G·C·C)-containing triplexes studied here. This seems surprising, since it might have been expected that the presence of protonated cytosines in the d(G·C·C)-based triplex would result in dramatic differences in triplex recognition possibilities. It seems then that the general susceptibility of triplex DNA to bind other molecules (from small cations to proteins) is very sensitive to its structure, but not too sensitive to chemical composition.

Acknowledgment. We are indebted to Prof. Miquel Pons, Dr. Jiri Sponer, Prof. Ferran Azorín, and Dr. Carlos González for many helpful discussions. This work has been supported by the Centre de Supercomputació de Catalunya (CESCA, Mol. Recog. Project) and the Spanish DGICYT (PB96-1005). R. Soliva is a predoctoral fellow of the Catalonian CIRIT.

Supporting Information Available: MIP distribution plots and water density plots for the stable triplex structures (2 pages, print/PDG). See any current masthead page for ordering informaiton and Web access instructions.

JA981121Q