J|A|C|S COMMUNICATIONS

Published on Web 04/08/2009

A Direct and Nondestructive Approach To Determine the Folding Structure of the I-Motif DNA Secondary Structure by NMR

Jixun Dai,[†] Attila Ambrus,[†] Laurence H. Hurley,[†,‡,§,||] and Danzhou Yang*[,†,‡,§,||]

College of Pharmacy, The University of Arizona, Tucson, Arizona 85721, BIO5 Institute, The University of Arizona, Tucson, Arizona 85721, Arizona Cancer Center, Tucson, Arizona 85724, and Department of Chemistry, The University of Arizona, Tucson, Arizona 85721

Received February 12, 2009; E-mail: yang@pharmacy.arizona.edu

We report here a direct and nondestructive method that can be utilized to unambiguously determine the folding structure of an I-motif DNA secondary structure formed from a native, nonmutated DNA sequence. The I-motif DNA secondary structure is a four-stranded structure consisting of parallel-stranded DNA duplexes zipped together in an antiparallel orientation by intercalated, hemiprotonated cytosine$^+$−cytosine base pairs[1] (Figure 1A). Since its first report in 1993, the biological role of I-motif structures as well as their potential in nanotechnological applications have been extensively explored.[2,3] For its nanotechnological applications, research revealed that non-DNA based C-rich sequences can form I-motifs as well. Different modifications or inclusions of RNA residues, locked nucleic acids, 2′-fluoro substitutions, and peptide nucleic acids can be incorporated into the I-motif structures.[4−7] The biological significance of I-motif structures is still being heavily investigated. Oligonucleotide fragments using the sequences from naturally occurring C-rich strands in the genome of human and other species have been shown to form intramolecular and intermolecular I-motifs *in vitro*, such as the *Tetrahymena thermophila* and human telomeric repeats, centromeric sequences, the human insulin minisatellite, the fragile X repeat, and oncogene promoters.[8−12] In particular, polyguanine/polycytosine (polyG/polyC) tracts have recently been demonstrated to be highly prevalent in human proximal promoter regions, especially those of oncogenes that are related with growth and proliferation.[13−17] These G/C-rich promoters are highly dynamic in their structures and are found to associate with nuclease hypersensitive elements (NHEs). Under superhelicity conditions, these G/C-rich regions can form alternative conformations different from the typical B-DNA structure.[18] While the G-rich strands can form DNA G-quadruplex structures, which have been demonstrated in a number of oncogene promoters and are suggested to function as regulatory elements in gene transcription, the complementary C-rich strands have the potential to form I-motif structures which may also be associated with transcriptional regulation.[13,14,19]

Nuclear magnetic resonance (NMR) is a major tool for structural studies of I-motifs.[1,10−12,20] In contrast to the number of G-quadruplex structures in the public domain, the molecular structures of I-motifs are more limited. Similarly to DNA G-quadruplexes, I-motif structures can be formed by one, two, or four strand(s). A tetrameric I-motif formed by short oligonucleotide strands appears to be more straightforward for NMR structural characterization, because the NMR spectra are simpler due to the symmetry of the system and thus the smaller number of resonances observed in the equivalence strands. For the more biologically relevant unimolecular

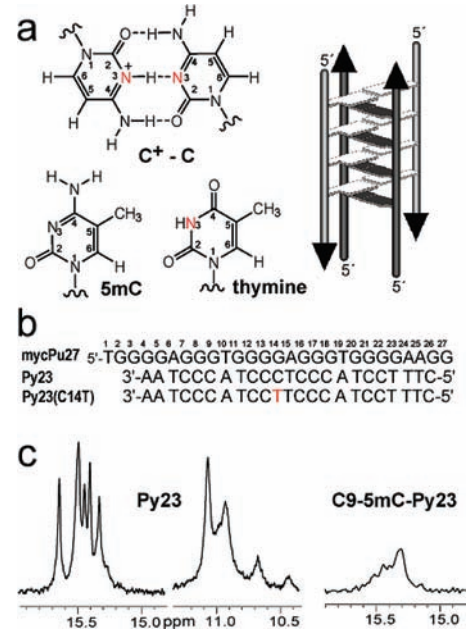

Figure 1. (a) (Left upper) A hemiprotonated C$^+$−C base pair; (left lower) 5-Methylcytosine base used in the chemical substitution for assignment and Thymine base; (right) A schematic drawing of a four-stranded I-motif structure. (b) The wild-type 27-mer G-rich sequence (mycPu27) of the NHE III$_1$ element of the c-Myc gene promoter and the DNA sequences used in this study (Py23 and Py23(C14T). (c) 1D ^{1}H NMR of cytosine- and thymine-imino regions of Py23 (left) and cytosine-imino region of C9-5mC-Py23, Py23 with a 5mC substitution at the C9 position (right). It is clear that the 5mC at C9 destabilizes the I-motif structure.

(intramolecular) I-motif structure formed by single-stranded DNA, the lack of symmetry of the structure, the resonance overlapping, and possible conformational exchange make the NMR spectral assignment much more challenging.[1,10−12,20] This is similar to G-quadruplexes; however, the NMR spectral analyses of the unimolecular I-motifs could be more perplexing. While the DNA G-quadruplex is a four-stranded structure of stacked G-tetrads, each G-tetrad is formed by four cyclically H-bonded guanine residues in the same plane and therefore the four guanines involved in one tetrad are connected and exhibit specific NOE connections with each other. In contrast, an I-motif structure consists of the intercalated hemiprotonated cytosine$^+$−cytosine base pairs (Figure 1A). The intercalated cytosine$^+$−cytosine base pairs are from two parallel duplexes and are not connected with each other and, thus, are more difficult to be spectrally assigned by NMR. In addition, not all cytosines of a C-run can be used simultaneously in C$^+$−C hemiprotonated base-pair formation, and the base-pairing partnership between cytosines of the two parallel strands can be more

[†] College of Pharmacy, The University of Arizona.
[‡] BIO5 Institute, The University of Arizona.
[§] Arizona Cancer Center.
[||] Department of Chemistry, The University of Arizona.

10.1021/ja900967r CCC: $40.75

variable. Thus the degree of sequence redundancy is higher in the C-rich sequence for the formation of an I-motif than in the G-rich sequence for the formation of G-quadruplex, which makes it more difficult to determine even the folding structure of an I-motif without full NMR spectral assignment and NOE connections.

To date, all structural studies of I-motifs follow the full proton NMR assignment process for nucleic acids,[1,10−12,20] which requires a complete set of COSY, TOCSY, and NOESY spectra, and sometimes further help from $^{1}H-^{31}P$ correlation spectroscopy and HMBC at natural abundance.[21] To determine the folding topology, the characteristic proton connectivities for I-motif structures observed in NOESY spectra are used. The possible $C^{+}-C$ pair combination can be deduced based on the specific short inter-residue $^{1}H-^{1}H$ distances corresponding to the characteristic intercalation topology, such as those for imino−imino, H1′−H1′, H1′−H2′/H2″, amino-H2′/H2″, H1′−H4′, and H4′−H4′. Due to a more severe NMR spectral overlapping and a larger number of $C^{+}-C$ base pairs in I-motif structures, chemical substitutions are always needed to deconvolute the NMR spectral assignment, assuming the substitutions do not cause structural alteration. These chemical substitutions are used to break the quasi-symmetry and provide a marker readily identifiable from other overlapping peaks. 5-Methyl-C (5mC) (Figure 1A) is the most commonly used for cytosines in the cytosine^{+}−cytosine base pairs. However, the incorporation of 5mC for cytosine can potentially affect the I-motif structures (see below). Here we report a direct method to unambiguously determine the folding structure of an intramolecular I-motif. This affordable method makes use of the site-specific low enrichment (6%) of uniform ^{15}N-labeled nucleotides, which is nondestructive and can be used in a native, nonmutated DNA sequence that forms the I-motif structure. Furthermore, this method can also be applied to unambiguously determine multiple equilibrating I-motif structures coexisting in a sequence.

We use the promoter sequence of the *c-myc* oncogene for the reported method. This sequence is used because DNA secondary structures formed in the c-myc NHE III_1 have been extensively studied.[13,15,22] *c-myc* is the most commonly overexpressed gene in human cancers whose proximal promoter region contains an NHE III_1 element, which can form DNA secondary structures that regulate 75−85% of the total transcription activity. The NHE III_1 element comprises five consecutive runs of guanines on one strand and cytosines on the other strand. The major G-quadruplex is formed within the 3′- four G-runs whose molecular structure has been determined by NMR.[22] The C-rich strand in the c-MYC NHE III_1 element has been shown to form I-motif structures at near neutral pH.[13] We have identified a sequence of the C-rich strand containing the modified 5′- four C-runs, complementary to the G-rich strand that forms the major c-myc G-quadruplex (Py23, Figure 1B). The one-dimensional ^{1}H NMR spectrum of the Py23 is shown in Figure 1C left. The well-resolved imino proton resonances located at 15−16 ppm indicate the formation of stable I-motif structure(s).[1,20] The sharp NMR spectral line widths indicate the I-motif structure(s) is of intramolecular monomeric nature. The monomeric nature of the I-motif structure(s) was confirmed by variable temperature studies by NMR and CD, in which Py23 shows a concentration-independent melting temperature (data not shown).

An essential feature for the I-motif structure is that one cytosine of every $C^{+}-C$ base pair is protonated at N3, and this proton is shared by the two base-paired cytosines. The imino proton resonances of hemiprotonated $C^{+}-C$ base pairs located at 15−16 ppm are characteristic of I-motif structures. The amino protons of the base-paired cytosines are found at approximately 9.5 and 8.4 ppm. The distinct chemical shifts at 15−16 ppm for the I-motif

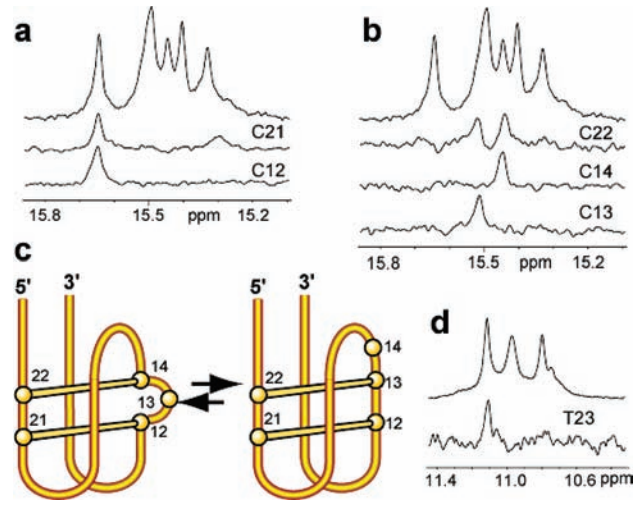

Figure 2. Assignments of imino protons based on site-specific low-enrichment (6%) ^{15}N-labeling. In each panel, the 1D ^{1}H spectrum (upper) and 1D ^{15}N-filtered HMQC spectra (lower) of different site-specific labeled samples are aligned as to the chemical shift. Each site-specifically labeled residue is shown above its corresponding spectrum. (a) Imino proton assignments of C21 and C12 of Py23 using 1D ^{15}N-filtered experiments on site-specific C21- and C12-labeled oligonucleotides. The hemiprotonated $C^{+}-C$ base pair C21−C12 in the DNA sequence Py23 was determined based on the same imino proton shared by C21 and C12. (b) Imino proton assignments of C22, C13, and C14 of Py23 using site-specific C-labeled oligonucleotides. The two hemiprotonated $C^{+}-C$ base pairs in equilibrium, C22−C14 and C22−C13, were determined based on the shared imino protons. (c) Schematic drawing of the folding structures of the two equilibrating I-motifs formed in the Py23 sequence. (d) The assignment of imino proton of T23 based on T23-^{15}N-labeled DNA oligonucleotide Py23(C14T). The T23 residue is thus indicated to be involved in the H-bonded capping structure. Experimental conditions: 1 mM DNA oligomer, 7 °C, pH 5.5.

cytosine imino protons result from the downfield shifting by a combination of hydrogen bonding and a positively charged protonation site. Using site-specific low-concentration (6%) incorporation of a ^{15}N-labeled cytosine nucleoside, the imino protons of cytosine residues can be unambiguously assigned. This is similar to the incorporation of ^{15}N-labeled guanines used for G-quadruplex structures;[23−26] however, for the I-motif structure, the hemiprotonated $C^{+}-C$ base pairs can be directly determined by this method. For example, each cytosine of the sequence 5′-CTTTCCTAC-CCTCCCTACCCTAA (Py23) (Figure 1B) is 6% labeled by ^{15}N-cytosine one at a time. As the two cytosines forming a hemiprotonated $C^{+}-C$ base pair share one imino H3 hydrogen, the imino proton involved in the hydrogen bonding distributes equally between the two base-paired cytosines (Figure 1A left). The imino proton in a $C^{+}-C$ base pair has one-bond coupling to the N3 atoms of both cytosines and can be readily detected by 1D ^{15}N-filtered HMQC experiments.[27] Through the $^{1}H-^{15}N$ one bond coupling, this proton will be detected as the same ^{1}H resonance in the 1D HMQC experiments for the two DNA samples site specifically ^{15}N labeled at each base-paired cytosine. The assignment of a $C^{+}-C$ base pair in Py23 is shown as an example in Figure 2A. The site-specific substitution of ^{15}N-labeled cytosine at the C21 and C12 positions, respectively, gives rise to a peak with the same cytosine imino proton resonance at 15.65 ppm, indicating C21 and C12 form a hemiprotonated $C^{+}-C$ base pair and shared the same imino proton. This leads to the direct identification of the $C^{+}-C$ base pair between C21 and C12 (Figure 2C). In comparison, the employment of chemical substitution with 5mC destabilizes the I-motif structure (Figure 1C right and Figure S1), while the method

of low-enriched site-specific labeling is direct, nondestructive, and affordable. The assignment of each cytosine imino proton involved in the hemiprotonated C^+-C base pairs enables the identification of each partner C^+-C base pair involved in an I-motif structure and thus the determination of the folding topology. For an I-motif structure, once the folding topology is determined, the molecular structure of the I-motif core can be reasonably calculated by computer modeling because of the nature of such a structure. For example, the neighboring strands are always antiparallel and connected by lateral loops, while the cytosines of the #n C-run are always base pairing with the cytosines of the #($n \pm 2$) C-run, and the widths of the adjacent grooves are always alternating between wide and narrow.

More significantly, this method can directly detect the equilibrating multiconformations of I-motif structures coexisting in a DNA sequence (Figure 2B). The coexisting multiple conformations are very difficult to directly determine by conventional assignment strategies using homonuclear 2D spectra only. Furthermore, the chemical substitution with 5mC often associated with conventional assignment strategies may destabilize a specific I-motif structure or shift the equilibrium between multiple conformations. As shown in Figure 2B, the C22-labeled DNA sequence Py23 (Figure 1B) shows two imino peaks at 15.5 ppm and 15.45 ppm in the ^{15}N-filtered experiment, indicating that C22 is involved in two different conformations. Each of the C22 imino peaks corresponds to the single imino peak arising from the C13- (15.5 ppm) and C14- (15.45 ppm) labeled Py23, indicating that C22 is base pairing with C13 in one conformation and with C14 in another, respectively. This unexpected result directly indicates that the Py23 sequence forms two stable I-motif conformations in slow equilibrium on the NMR time scale (ms), as they have sharp and well-resolved NMR peaks (Figure 2B).

This method can also be applied to the direct assignment of imino protons of thymine residues that are involved in H-bonding interactions in an I-motif structure. The conformation of thymines in the loop regions and the flanking regions are essential to the structure and stability of an I-motif structure; in particular, for a specific I-motif structure, specific thymines are found to be involved in different capping structures when they are hydrogen bonded with other residues. Only the imino protons from the hydrogen-bonded thymines are clearly observable in NMR at temperatures above 0 °C. In a unimolecular I-motif structure, it is very challenging to assign the multiple thymine residues using the conventional assignment strategy. The low-abundance (6%) site-specific labeling of thymines can be used to solve this problem and determine the thymine residues involved in the H-bonded capping structures. As shown in Figure 2D, using the DNA sequence 5′-CTTTCCTAC-CCTCCCTACCCTAA-3′, Py23(C14T), the resonance of imino protons of T23 was unambiguously assigned based on the one-bond $^{15}N-^{1}H$ coupling in the ^{15}N-filtered experiments.

In summary, the reported approach using site-specific low-enrichment ^{15}N-labeled cytosine provides a direct and unambiguous determination of the hemiprotonated C^+-C base pairs in an I-motif structure by NMR with affordability. This direct detection of the C^+-C base pairs can unambiguously determine the folding topology of a unimolecular I-motif structure. Because the C-rich strand possesses an inherent sequence redundancy in the formation of unimolecular I-motif structures, the unambiguous determination of the folding topology of I-motif has been a challenging and arduous task using conventional NMR spectral assignment strategies. More significantly, the reported method can directly and unambiguously determine the equilibrating multiple I-motif conformations coexisting in a single DNA sequence, which would be a very difficult task using the conventional assignment strategy. This method can also be applied to the direct detection of the H-bonded thymines that are involved in the capping structures. The reported method is direct and easy to use and can provide direct folding topology and specific capping structure information. In addition, this method can aid the full spectral assignment for the complete NMR structure determination; e.g., the assignment of the base H5 and H6 protons can be obtained by long-range connections with the imino protons. The direct assignment of the cytosine and thymine imino protons can also provide important internucleotide NOEs of the stacking cytosines and capping thymines for NMR structure determination. Furthermore, the approach can be applicable to I-motif structures involving non-DNA residues and provides a direct and affordable method to tackle related structure problems.

Acknowledgment. This research was supported by the National Institutes of Health (1S10 RR16659, CA122952, and CA94166) and the Arizona Biomedical Research Commission (0014). We acknowledge Tiffanie Bialis, who assisted with the NMR experiments. We thank Megan Carver for proofreading the paper.

Supporting Information Available: Experimental Methods. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Gehring, K.; Leroy, J. L.; Gueron, M. *Nature* **1993**, *363* (6429), 561–5.
(2) Chakraborty, S.; Modi, S.; Krishnan, Y. *Chem. Commun. (Camb.)* **2008**, (1), 70–2.
(3) Gilbert, D. E.; Feigon, J. *Curr. Opin. Struct. Biol.* **1999**, *9* (3), 305–14.
(4) Snoussi, K.; Nonin-Lecomte, S.; Leroy, J. L. *J. Mol. Biol.* **2001**, *309* (1), 139–53.
(5) Fenna, C. P.; Wilkinson, V. J.; Arnold, J. R.; Cosstick, R.; Fisher, J. *Chem. Commun. (Camb).* **2008**, *30*, 3567–9 (Epub 2008 Jun 5).
(6) Modi, S.; Wani, A. H.; Krishnan, Y. *Nucleic Acids Res.* **2006**, *34* (16), 4354–63 (Epub 2006 Aug 26).
(7) Brazier, J. A.; Fisher, J.; Cosstick, R. *Angew. Chem., Int. Ed.* **2006**, *45*, 114–117.
(8) Gueron, M.; Leroy, J. L. *Curr. Opin. Struct. Biol.* **2000**, *10* (3), 326–31.
(9) Mills, M.; Lacroix, L.; Arimondo, P. B.; Leroy, J. L.; Francois, J. C.; Klump, H.; Mergny, J. L. *Curr. Med. Chem. Anticancer Agents* **2002**, *2* (5), 627–44.
(10) Esmaili, N.; Leroy, J. L. *Nucleic Acids Res.* **2005**, *33* (1), 213–24.
(11) Nonin-Lecomte, S.; Leroy, J. L. *J. Mol. Biol.* **2001**, *309* (2), 491–506.
(12) Phan, A. T.; Gueron, M.; Leroy, J. L. *J. Mol. Biol.* **2000**, *299* (1), 123–44.
(13) Simonsson, T.; Pribylova, M.; Vorlickova, M. *Biochem. Biophys. Res. Commun.* **2000**, *278* (1), 158–66.
(14) Rustighi, A.; Tessari, M. A.; Vascotto, F.; Sgarra, R.; Giancotti, V.; Manfioletti, G. *Biochemistry* **2002**, *41* (4), 1229–1240.
(15) Siddiqui-Jain, A.; Grand, C. L.; Bearss, D. J.; Hurley, L. H. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (18), 11593–11598.
(16) Huppert, J. L.; Balasubramanian, S. *Nucleic Acids Res.* **2007**, *35* (2), 406–413.
(17) Maizels, N. *Nat. Struct. Mol. Biol.* **2006**, *13* (12), 1055–1059.
(18) Kouzine, F.; Sanford, S.; Elisha-Feil, Z.; Levens, D. *Nat. Struct. Mol. Biol.* **2008**, *15* (2), 146–154.
(19) Guo, K.; Pourpak, A.; Beetz-Rogers, K.; Gokhale, V.; Sun, D.; Hurley, L. H. *J. Am. Chem. Soc.* **2007**, *129* (33), 10220–10228.
(20) Han, X.; Leroy, J. L.; Gueron, M. *J. Mol. Biol.* **1998**, *278* (5), 949–65.
(21) Cromsigt, J.; van Buuren, B.; Schleucher, J.; Wijmenga, S. *Methods Enzymol.* **2001**, *338*, 371–99.
(22) Ambrus, A.; Chen, D.; Dai, J.; Jones, R. A.; Yang, D. Z. *Biochemistry* **2005**, *44* (6), 2048–58.
(23) Dai, J. X.; Carver, M.; Punchihewa, C.; Jones, R. A.; Yang, D. Z. *Nucleic Acids Res.* **2007**, *35* (15), 4927–4940.
(24) Ambrus, A.; Chen, D.; Dai, J.; Bialis, T.; Jones, R. A.; Yang, D. Z. *Nucleic Acids Res.* **2006**, *34* (9), 2723–2735.
(25) Dai, J.; Dexheimer, T. S.; Chen, D.; Carver, M.; Ambrus, A.; Jones, R. A.; Yang, D. Z. *J. Am. Chem. Soc.* **2006**, *128* (4), 1096–1098.
(26) Phan, A. T.; Patel, D. J. *J. Am. Chem. Soc.* **2002**, *124* (7), 1160–1161.
(27) (a) Sklenar, V.; Bax, A. *J. Magn. Reson.* **1987**, *74* (3), 469–479. (b) Szewczak, A. A.; Kellogg, G. W.; Moore, P. B. *FEBS Lett.* **1993**, *327* (3), 261–264.

JA900967R