DOI: 10.1002/ange.201405637

DNA Structures

Structural Basis for the Identification of an i-Motif Tetraplex Core with a Parallel-Duplex Junction as a Structural Motif in CCG Triplet Repeats

*Yi-Wen Chen, Cyong-Ru Jhan, Stephen Neidle, and Ming-Hon Hou**

Abstract: *CCG triplet repeats can fold into tetraplex structures, which are associated with the expansion of $(CCG)_n$ trinucleotide sequences in certain neurological diseases. These structures are stabilized by intertwining i-motifs. However, the structural basis for tetraplex i-motif formation in CCG triplet repeats remains largely unknown. We report the first crystal structure of a CCG-repeat sequence, which shows that two $dT(CCG)_3A$ strands can associate to form a tetraplex structure with an i-motif core containing four $C{:}C^+$ pairs flanked by two G:G homopurine base pairs as a structural motif. The tetraplex core is attached to a short parallel-stranded duplex. Each hairpin itself contains a central CCG loop in which the nucleotides are flipped out and stabilized by stacking interactions. The helical twists between adjacent cytosine residues of this structure in the i-motif core have an average value of 30°, which is greater than those previously reported for i-motif structures.*

Numerous neurological diseases are associated with the expansion of trinucleotide repeats (TNRs).[1] A notable example is the $(CGG)_n/(CCG)_n$ repeat in the X chromosome; this repeat has been associated with fragile X syndrome.[2] The CGG TNRs in the coding sequence of the *FMR1* (fragile mental retardation 1) gene result in the production of an aberrant protein that plays a critical role in the pathogenesis of the disease.[3] The massive TNR expansion found in neurodegenerative disorders may be caused by the slipped register of the DNA complementary strands in addition to the transient formation of noncanonical DNA structures during DNA replication.[4] The unusual structural characteristics of expandable repeats have been found to determine the threshold length and the large-scale character of the expansions. Single-stranded $(CNG)_n$ repeats are able to form hairpin DNA structures that consist of both Watson–Crick and mismatched base pairs.[5] Besides hairpins, single-stranded $(CGG)_n$ and $(CCG)_n$ repeats can fold into four-stranded helical structures stabilized by intertwining G-quartets and i-motifs.[1] Nucleic-acid sequences capable of folding into an i-motif tetraplex are found in the telomeric DNA sequences of various species.[6] Studies on CCG triplet repeats have led to proposals that these repeats can form four-stranded structures in which two CCG repeats form hairpin-like structures that combine to form an i-motif arrangement.[7] These models were based on an analysis of gel electrophoretic patterns, CD spectra, and UV-light-induced cross-links. However, to date, no single-crystal X-ray analysis of i-motif formation in CCG triplet repeats has been reported.

We report herein the first crystal structure of the oligonucleotide $dT(CCG)_3A$, which is a structure-forming portion of a $(CCG)_n$ triplet-repeat sequence. All atoms of the hairpin molecule exhibit well-defined electron density (see Figure S1 in the Supporting Information) and have been included in the final model. The asymmetric unit contains one DNA strand, which folds back to form a hairpin structure at the central CCG region. The structural unit comprises two $dT(CCG)_3A$ strands related by a crystallographic twofold axis along the pseudohelix axis to form an arrangement resembling two tango dancers (Figure 1 A). The CCG-repeat hairpins in each strand are hydrogen-bonded to one another to form a tetraplex core. The central region of the tetraplex comprises a four-stranded intercalated motif in which the hairpin monomers intersect and are held together by hydrogen bonds from one G:G homopurine base pair and three intercalated $C{:}C^+$ base pairs. Additionally, the three nucleotide residues CGA at the 3′ end of the two oligomers form a short stalk of a right-handed parallel-duplex helix with homopyrimidine and homopurine base pairs, $C{:}C^+$, G:G, and A:A (Figure 1 B,C). Two nucleotides in the central CCG loop of the hairpin moiety are flipped out, as is the 5′ base, T1. Rather than pairing with 3′-terminal A11, 5′-terminal T1 is symmetrically tilted in the wide groove and leaves the stem region to stack with the looped-out C6′ residue from a twofold-related hairpin DNA molecule. These three bases (T1–C6–G7) are perpendicular to the central i-motif core and stack together, with an average interplanar distance of 3.5 Å. The 3′ cytosine residue (C5) of this CCG loop protrudes into the center of the i-motif core. The second cytosine residue (C6) and the 5′ guanosine residue (G7) of the CCG loop repeat are oriented in the opposite direction to C5. The tightly held overall arrangement suggests that this CCG loop in the hairpin structure is inflexible and maintained in the observed conformation: C5 of one strand forms a stacking interaction with G4 of the other strand. This interaction crosses the narrow grooves of the structure, and C6 is stacked on the adjacent G7 base.

[*] Dr. Y. W. Chen, Dr. M. H. Hou
Institute of Genomics and Bioinformatics and Institute of Biochemistry, National Chung Hsing University
No. 250 Kuo-Kuang Road, Taichung (Taiwan)
E-mail: mhho@nchu.edu.tw

C. R. Jhan
Department of Life Sciences, National Chung Hsing University
No. 250 Kuo-Kuang Road, Taichung (Taiwan)

Prof. S. Neidle
The School of Pharmacy, University College London
London WC1N 1AX (UK)

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201405637.

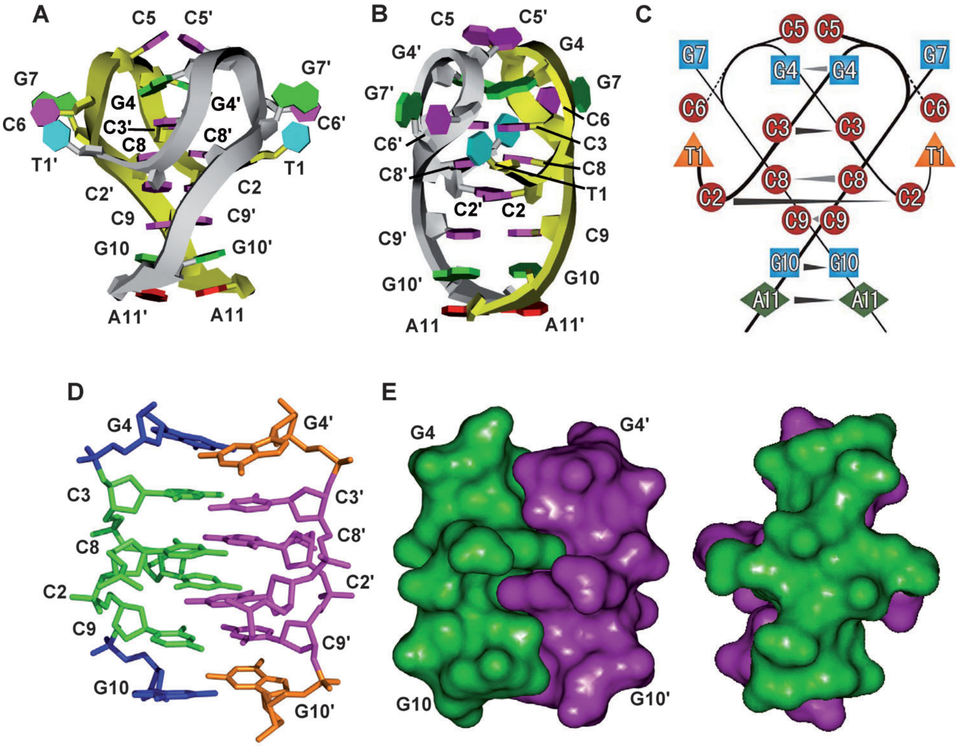

Figure 1. A) Simplified representation of the crystal structure of the dT(CCG)$_3$A DNA strands that fold into an intercalated motif by forming a double hairpin. Guanine bases are green, adenine bases are red, thymine bases are cyan, and cytosine bases are pink. B) View of the structure from the wide-groove direction. C) Topology and base pairs in the structure. Base pairing is indicated by arrows. D) Detailed side view of the i-motif core flanked by G:G homopurine base pairs. E) Solvent-accessible-surface views of the i-motif flanked by the G:G homopurine base pairs from the wide-groove (left) and narrow-groove directions (right).

Figure 1 D presents a detailed view of the two symmetrical hairpins in the i-motif core. The core is formed by two symmetrical dT(CCG)$_3$A hairpins, one from each strand, and comprises four C:C$^+$ pairs flanked by two G:G base pairs. The two trinucleotide halves of the i-motif core are related by a local twofold symmetry. They exhibit global similarities and local differences with a root-mean-square deviation of 0.967 Å between the two halves (see Figure S2). The average stacking interval between the C:C pairs is 3.3 Å, which is larger than the distance observed between the C:C pairs in the d(CCCT) structure.[8] The stacking interval between the C:C pair and the G:G pair is also 3.3 Å (see Table S1). The i-motif core of the structure contains two wide and two narrow grooves (Figure 1 E). From the wide-groove direction, the lathlike central segment is seen to be formed by the cytosine residues from the first and third CCG triplet of dT(CCG)$_3$A in an intercalative manner that is similar to that observed in the crystal structures of d(CCCAAT).[9] The phosphate–phosphate distances across the wide grooves of the i-motif core are similar from top to bottom, and the average phosphate–phosphate distance across the wide grooves of the core is 16.4 Å. The narrow grooves are formed by the proximity of the two individual DNA hairpins. Because of the rotation of the upper phosphate group, the average phosphate–phosphate distance across the upper narrow groove of the i-motif core is 9.1 Å, in contrast to the distance of 7.5 Å across the lower narrow groove. The helical twist between adjacent cytosine residues has an average value of 30°, which is larger than that previously reported for several i-motif structures (see Table S1).[10] Interestingly, the twist angles of the G4–C3 and C9–G10 steps are 56 and 53°, respectively, and generate an overwound DNA conformation. Thus, the four strands in the intercalated stem tightly twist in a clockwise direction and form a right-handed tetrahelix.

Figure 2 shows the observed homobase pairs in the structure, with the hydrogen bonds indicated. The two flanking G:G base pairs form pairs of N2−H⋅⋅⋅N3 hydrogen bonds between the narrow-groove sites with high propeller-twist and buckle angles (Figure 2 A,B; see also Table S2). Hydrogen bonding also occurs between the phosphate groups of G7 and C3/G4 of the opposite strand. This interaction is an important contributor to the stability of the structure and the formation of

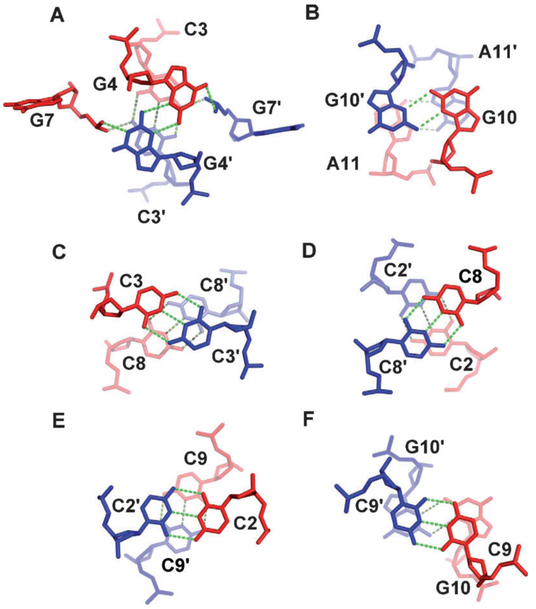

Figure 2. The base pairs in the structure, showing the stacking interactions in the dimeric complex of dT(CCG)$_3$A at A) the C3pG4 step, B) the G10pA11 step, and C–F) the steps in the i-motif core of the refined structure.

highly twisted G4:G4 pairs. The four C:C$^+$ pairs in the i-motif each have three hydrogen bonds between N4 and O2^P (^{P} indicates the parallel strand), between N3 and N3^P, and between O2 and N4^P (Figure 2C–F). In each case, the C:C$^+$ pair must be protonated at N3 of one or the other C residue, with the proton shared between N3 and N3^P. The 5′-terminal A:A pair is formed by two N6–H···N7 hydrogen bonds, between the wide-groove sites (Figure 2B). This pairing is known to occur between both protonated adenine residues and between their neutral forms.[11] The former case was detected in d(TCGA) at pH 4.0 by NMR spectroscopic analysis.[12] Because the present crystal was obtained at pH 6, we assume that the adenine moiety of the A11 residue is not protonated. Additionally, consistent with the results obtained for the G:G homobase pairs, the A:A homobase pairs are nonplanar with high propeller-twist and buckle angles (see Table S2).

The highest peak in the initial (F_o–F_c) difference maps was identified as a Co^{II} ion and was included in the refinement. Two cobalt(II) ions coordinate to the hairpin structure of dT(CCG)$_3$A in the asymmetric unit (Figure 3A,B). One of the Co^{II} ions is clearly coordinated to the N7 nitrogen atom of G4 and is octahedrally coordinated to five water molecules. An additional Co^{II} ion is coordinated to the N7 atom of G7 and the phosphate oxygen atom of C6 with an incomplete hydration shell. Similar coordination to guanine has also been observed in the interaction of Co^{II} in a Z-DNA crystal.[13] Previous studies indicated that the bridging water molecules play an important mediating role in the interactions between the DNA strands, and in turn stabilize the i-motif structures of DNA.[14] Five pairs of symmetry-ordered water molecules mediate the interaction between the DNA hairpins of the independent asymmetric units (see Table S3). Figure 3C shows the association of two hairpin structures of CCG repeats and the associated waters of hydration. At the top of the structure, the flipped-out C5 base is linked by one bridging water molecule (w67) to the C5 phosphate group of the opposite chain. The N4 amino group of C2 and C8 of the i-motif core also hydrogen bonds to two water molecules (w35 and w107), which in turn directly bridge to the oxygen atom of a neighboring phosphate. Besides the cytosine residues, G10 and A11 are linked by one bridging water molecule (w41) to the oxygen atom of the phosphate and ribose group of the opposite chain, respectively. G7 and G4 are linked by one bridging water molecule (w36) to the oxygen atom of the phosphate and the guanine N2 atom of the opposite chain, respectively. Interesting, w36 also plays a key role in stabilizing the hairpin structure of dT(CCG)$_3$A by bridging the G7 phosphate group and G4 (or C3; Figure 3D).

Fragile X syndrome, which is one of the most common forms of inherited mental retardation, originates from the expansion of the trinucleotide motif d(CGG)·d(CCG).[15] Previous studies showed that under physiological conditions, both the G-rich and the C-rich single strands of d(CGG)·d(CCG) repeats are able to form secondary structures, which may cause these expansions.[16] This propensity for secondary-structure formation is more pronounced for d(CCG)$_n$ repeats than for d(CGG)$_n$ strands, and this observation suggests that the d(CCG)$_n$ strand is most likely to form a hairpin or slippage structure and exhibit asymmetric strand expansion during DNA replication. Besides the hairpin structure, the strands of CCG repeats have been reported to adopt a tetraplex structure based on two parallel-oriented hairpins that are held together by hemiprotonated intermolecular C:C$^+$ pairs under physiological conditions.[7a] In this study, we obtained the first crystal structure of a CCG repeat sequence correlated with neurological disease. The crystal structure shows that two dT(CCG)$_3$A strands can associate to form a tetraplex structure with an i-motif core flanked by two G:G homobase pairs and a parallel duplex stem as a structural motif. The structure of a DNA G-quadruplex–duplex junction was reported previously.[17] Furthermore, previous studies have shown that the self-associative properties of cytidine-rich oligonucleotides that favor the formation of symmetrical i-motif tetramers enable these oligonucleotides to form supramolecular structures.[18] The base-pairing interactions in the proton-bound dimer of cytosine is the major force responsible for the stabilization of DNA i-motif conformations.[7b] A model of longer CCG repeats generated from the

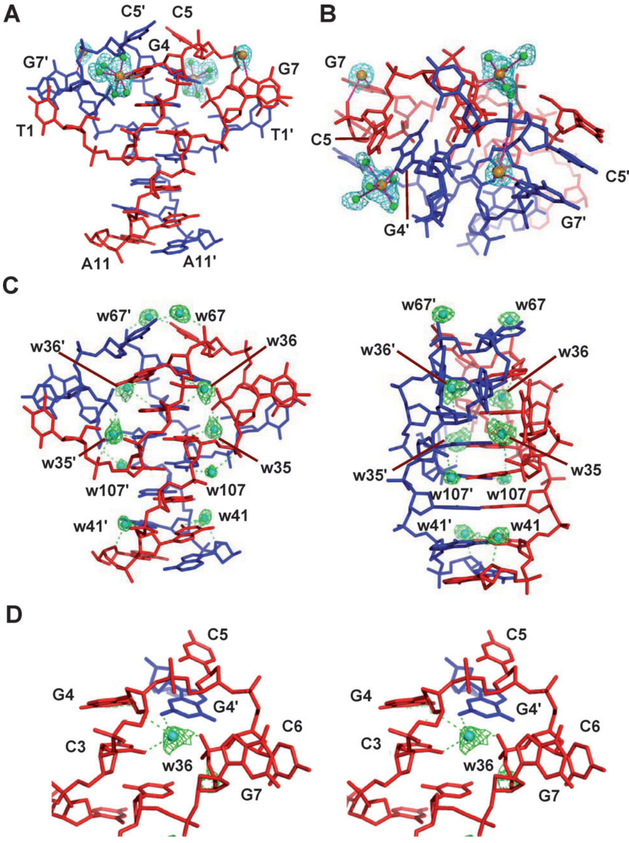

Figure 3. A,B) Coordination of the N7 atoms of guanine residues by two cobalt ions with and without hydration in the dimeric hairpin complex of dT(CCG)$_3$A, as viewed from the narrow-groove direction (A) and the top (B). The cobalt(II) ions are shown as orange spheres, and the coordinated water molecules are colored green. The electron density is contoured at 1.5 σ. C) View of the water cluster formed between the hairpins of dT(CCG)$_3$A from the narrow-groove (left) and wide-groove directions (right). D) Detailed stereoview of the interactions at the w36 binding site. The difference electron density is contoured at 1.0 σ. Dashed lines show direct hydrogen bonds and coordination bonds.

present crystal structure of three CCG repeats is shown in Figure S3. According to the analysis of $(CCG)_9$ by circular dichroism (see Figure S4), the hairpin homodimerization of the CCG repeats is promoted by increasing molecular length. This result supports a model of longer CCG repeats with hairpins that associate and intertwine to form a structure which resembles a supercoil, and which is formed by repeats of four intercalated $C{:}C^+$ pairs flanked by two G:G homobase pairs as a structural motif. This organization is generally consistent with models that have been proposed for the CCG-repeat tetraplex structure.[7a] However, in contrast to the classical G:C pairs of these models, G:G homobase pairs are involved in the present three-repeat crystal structure.

Nucleic acids that contain four stretches of cytidine residues or hairpins with two cytidine stretches can form parallel duplexes stabilized by hemiprotonated $(C{:}C^+)$ base pairs.[6] Several crystal structures of variants of C-rich telomeric DNA sequences, such as d(AACCCC), d(CCCT), and d(TAACCC), have been reported.[8,9,14b,19] These molecules form a more or less symmetrical tetraplex arrangement with cytosine-containing parallel duplexes held together by hemiprotonated $(C{:}C^+)$ base pairs and two duplexes intercalated with each other in the opposite polarity (i-motif). The i-motif cores of these reported structures are similar (see Figure S5) and are in striking contrast to the tetraplex structure of the CCG repeats described herein. It is notable that the twist angle between adjacent $C{:}C^+$ base pairs is approximately 30° in the present structure, and thus greater than the twist angles of about 16–20° observed in previously determined i-motif structures. We also observed a large twist angle between adjacent $C{:}C^+$ and G:G pairs (ca. 55°), as previously found in the parallel-stranded duplex structure of d(GCCAAAGCT).[20] These differences result in the right-handed and overwound conformation of the CCG repeats in the present structure. Additionally, for $dT(CCG)_3A$, d(AACCCC), and d(TAACCC), the 5′- and 3′-end sequences may affect the structures of the boundary regions in the i-motif tetraplex.[14b,19] For example, in the metazoan telomeric sequence d(TAACCC), a stabilized loop is formed by the TAA sequence at the 5′ termini of the molecule, with the i-motif core of the C-rich strand stabilized by the non-Watson–Crick base pairing of A:T, which occurs between the flanking non-cytosine bases of the telomeric DNA sequences.[19] In the *Tetrahymena* telomeric sequence, d(AACCCC), the structure has an adenine cluster at the 5′ terminus of the DNA strands.[14b] In the tetraplex structure of $dT(CCG)_3A$, the CGA sequence near the 3′ terminus has been shown to promote the formation of parallel $d(CGA)_2$ homoduplexes (II-DNA), with C:C, G:G, and A:A pairs stabilizing the i-motif core. The previously described three base-pair formations are identical to those observed in NMR spectroscopic analyses of d(CGA) and d(TCGA).[21]

Herein, we have demonstrated the propensity of CCG repeats to undergo base pairing between the hemiprotonated cytosine residues of one C-rich hairpin duplex and the cytosine residues of a second hairpin duplex to form a stable i-motif tetraplex structure. From previous studies, two possible pathways for the formation of the i-motif tetramers of CCG repeats during DNA expansion may be considered[22] (see Figure S6). In the first model, hairpins and slipped structures consisting of both normal and mismatched base pairs occur at the CCG repeats during DNA replication. Tetraplex i-motif formation subsequently proceeds by the association of two hairpin duplexes. The second model suggests that tetraplex formation proceeds by sequential strand association into duplex and triplex intermediate species. Our observation of this i-motif structure provides a possible molecular-level pathological consequence of CCG-triplet-repeat expansion. The existence of an i-motif tetraplex structure for CCG repeats in vivo is worthy of further study as, for example, a target for therapeutic intervention to inhibit aberrant protein expression.

Received: May 26, 2014
Published online: August 19, 2014

Keywords: DNA structures · i-motifs · neurological disease · tetraplex structures · X-ray diffraction

[1] S. M. Mirkin, *Nature* **2007**, *447*, 932 – 940.

[2] a) H. L. Paulson, K. H. Fischbeck, *Annu. Rev. Neurosci.* **1996**, *19*, 79 – 107; b) S. T. Warren, C. T. Ashley, Jr., *Annu. Rev. Neurosci.* **1995**, *18*, 77 – 99.

[3] Y.-H. Fu, D. P. A. Kuhl, A. Pizzuti, M. Pieretti, J. S. Sutcliffe, S. Richards, A. J. M. H. Verkert, J. J. A. Holden, R. G. Fenwick, Jr., S. T. Warren, B. A. Oostra, D. L. Nelson, C. T. Caskey, *Cell* **1991**, *67*, 1047 – 1058.

[4] a) A. Pluciennik, R. R. Iyer, P. Parniewski, R. D. Wells, *J. Biol. Chem.* **2000**, *275*, 28386 – 28397; b) P. Parniewski, P. Staczek, *Adv. Exp. Med. Biol.* **2002**, *516*, 1 – 25; c) Y. S. Lo, W. H. Tseng, C. Y. Chuang, M. H. Hou, *Nucleic Acids Res.* **2013**, *41*, 4284 – 4294.

[5] a) R. D. Wells, R. Dere, M. L. Hebert, M. Napierala, L. S. Son, *Nucleic Acids Res.* **2005**, *33*, 3785 – 3798; b) A. Kiliszek, R. Kierzek, W. J. Krzyzosiak, W. Rypniewski, *Nucleic Acids Res.* **2012**, *40*, 8155 – 8162; c) Y. W. Chen, M. H. Hou, *J. Inorg. Biochem.* **2013**, *121*, 28 – 36; d) M. H. Hou, H. Robinson, Y. G. Gao, A. H. Wang, *Nucleic Acids Res.* **2002**, *30*, 4910 – 4917.

[6] D. E. Gilbert, J. Feigon, *Curr. Opin. Struct. Biol.* **1999**, *9*, 305 – 314.

[7] a) P. Fojtik, M. Vorlickova, *Nucleic Acids Res.* **2001**, *29*, 4684 – 4690; b) B. Yang, M. T. Rodgers, *J. Am. Chem. Soc.* **2014**, *136*, 282 – 290.

[8] C. H. Kang, I. Berger, C. Lockshin, R. Ratliff, R. Moyzis, A. Rich, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 11636 – 11640.

[9] I. Berger, C. Kang, A. Fredian, R. Ratliff, R. Moyzis, A. Rich, *Nat. Struct. Biol.* **1995**, *2*, 416 – 425.

[10] M. Guéron, J. L. Leroy, *Curr. Opin. Struct. Biol.* **2000**, *10*, 326 – 331.

[11] T. Sunami, T. Kobuna, J. Kondo, I. Hirao, K. Watanabe, K. Miura, A. Takenaka, *Nucleic Acids Res. Suppl.* **2002**, 51 – 52.

[12] D. G. Reid, S. A. Salisbury, T. Brown, D. H. Williams, *Biochemistry* **1985**, *24*, 4325 – 4332.

[13] S. Thiyagarajan, S. S. Rajan, N. Gautham, *Nucleic Acids Res.* **2004**, *32*, 5945 – 5953.

[14] a) J. Weil, T. Min, C. Yang, S. Wang, C. Sutherland, N. Sinha, C. Kang, *Acta Crystallogr. Sect. D* **1999**, *55*, 422 – 429; b) L. Cai, L. Chen, S. Raghavan, R. Ratliff, R. Moyzis, A. Rich, *Nucleic Acids Res.* **1998**, *26*, 4696 – 4705; c) L. Chen, L. Cai, X. Zhang, A. Rich, *Biochemistry* **1994**, *33*, 13540 – 13546.

[15] D. Yudkin, B. E. Hayward, M. I. Aladjem, D. Kumari, K. Usdin, *Hum. Mol. Genet.* **2014**, *23*, 2940.

[16] J. M. Darlow, D. R. Leach, *J. Mol. Biol.* **1998**, *275*, 3 – 16.

[17] K. W. Lim, A. T. Phan, *Angew. Chem.* **2013**, *125*, 8728 – 8731; *Angew. Chem. Int. Ed.* **2013**, *52*, 8566 – 8569.

[18] E. Guittet, D. Renciuk, J. L. Leroy, *Nucleic Acids Res.* **2012**, *40*, 5162 – 5170.

[19] C. Kang, I. Berger, C. Lockshin, R. Ratliff, R. Moyzis, A. Rich, *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 3874 – 3878.

[20] T. Sunami, J. Kondo, T. Kobuna, I. Hirao, K. Watanabe, K. Miura, A. Takenaka, *Nucleic Acids Res.* **2002**, *30*, 5253 – 5260.

[21] H. Robinson, A. H. Wang, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 5224 – 5228.

[22] a) J. L. Leroy, *Nucleic Acids Res.* **2009**, *37*, 4127 – 4134; b) J. Gallego, S. H. Chou, B. R. Reid, *J. Mol. Biol.* **1997**, *273*, 840 – 856.