Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 3874–3878, April 1995
Biochemistry

Stable loop in the crystal structure of the intercalated four-stranded cytosine-rich metazoan telomere

(C·C$^+$ base pairs/Hoogsteen base pairing/parallel stranded duplex/trinucleotide loop/cruciform extrusion)

CHULHEE KANG†, IMRE BERGER†‡, CURTIS LOCKSHIN†, ROBERT RATLIFF§, ROBERT MOYZIS§, AND ALEXANDER RICH†

†Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; ‡Department of Biophysical Chemistry, Hannover Medical School, 30623 Hannover, Germany; and §Center for Human Genome Studies, Los Alamos National Laboratories, Los Alamos, NM 87545

Contributed by Alexander Rich, December 29, 1994

ABSTRACT **In most metazoans, the telomeric cytosine-rich strand repeating sequence is d(TAACCC). The crystal structure of this sequence was solved to 1.9-Å resolution. Four strands associate via the cytosine-containing parts to form a four-stranded intercalated structure held together by C·C$^+$ hydrogen bonds. The base-paired strands are parallel to each other, and the two duplexes are intercalated into each other in opposite orientations. One TAA end forms a highly stabilized loop with the 5′ thymine Hoogsteen-base-paired to the third adenine. The 5′ end of this loop is in close proximity to the 3′ end of one of the other intercalated cytosine strands. Instead of being entirely in a DNA duplex, this structure suggests the possibility of an alternative conformation for the cytosine-rich telomere strands.**

Telomere DNA at chromosome ends has a large number of repeating sequences in which one DNA strand typically has guanine clusters and the other has cytosine clusters (1). These clusters are interspersed with short sequences containing adenine and thymine residues. The telomere G-rich strand is capable of forming a four-stranded structure with four guanines hydrogen-bonded in one plane (2, 3). It is likely that this structure may be found *in vivo*, as several proteins have been cloned that interact with it in a specific manner (4–7). In multicellular animals, the C-rich strand repeating sequence is generally d(TAACCC). NMR studies of C-rich telomere sequences show that they can also form a four-stranded structure with cytosine-containing parallel duplexes held together by hemiprotonated (C·C$^+$) base pairs and two duplexes intercalated with each other in opposite polarity (8–10). Recent x-ray diffraction studies of C-rich deoxynucleotides have revealed many features of this intercalative system (11, 12). If the cytosines in d(TAACCC) are involved in forming a four-stranded structure, what is the structural significance of the TAA component? Here we report the crystal structure of d(TAACCC).¶ This molecule forms a four-stranded cytosine-intercalated structure. Moreover, it utilizes the sequence TAA at the ends to form a novel loop in which the 5′ end of one molecule is in very close proximity to the 3′ end of another molecule. This highly stabilized loop facilitates a potential manner in which C-strand telomere segments could associate with each other.

The sequence d(TAACCC) and its complement were initially isolated as the human telomere sequence (13), and its evolutionary conservation among vertebrate species was later demonstrated (14, 15). The functional interchangeability of repeats of this sequence with the more variable yeast telomere suggests an ancient origin for this conserved DNA sequence (16). Recent experiments on a large number of invertebrate species show that this telomere repeat originated prior to the evolution of multicellular animals and, hence, has been conserved for >1 billion years (E. C. McCanlies, J. Meyne, and R.M., unpublished data).

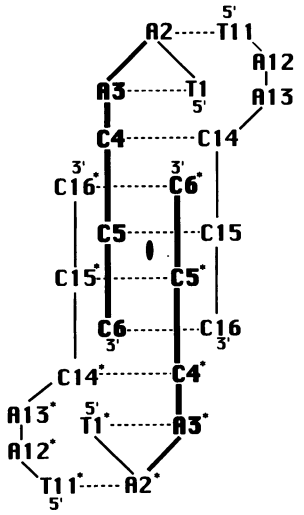

FIG. 1. Schematic diagram illustrating the structure of d(TAACCC) from the metazoan telomere. The asymmetric unit has two chains, numbered 1–6, 11–16. The numbers without asterisks represent one asymmetric unit, while those with asterisks are another asymmetric unit. They are related by a twofold axis between the base pairs C15·C5 and C5*·C15* indicated by the solid oval. Dashed lines represent the hydrogen-bonding interactions; solid lines represent the covalent links.

MATERIALS AND METHODS

Crystals of d(TAACCC) were grown from a solution containing 5% (vol/vol) 2-methyl-2,4-pentanediol (MPD), 20 mM $MgCl_2$, 0.10 mM spermine, 80 mM KCl, 40 mM potassium cacodylate buffer (pH 6.0), and 2 mM DNA (single-strand concentration) equilibrated against a reservoir of 35% MPD. Within 24 hr, crystals appeared as trapezoidal plates ≈0.4 × 0.7 × 0.3 mm in dimension. Data were collected at 4°C with an *R*-AXIS IIc imaging plate (Rigaku). The crystals did not show any detectable decay over a data collection period of 48 hr. We obtained crystals of similar size of d(TAA^{Br}CCC) and d(TAAC^{Br}CC) from the same condition as the unbrominated DNA sequence. Crystals of the unbrominated sequence belong to the space group *F*222 with cell parameters a = 59.94, b = 81.33, and c = 26.86 Å. Unfortunately, d(TAA^{Br}CCC) crystallized in a hexagonal space group (a = 35.27 and c = 51.12 Å). The sequence d(TAAC^{Br}CC), even though crystallizing in

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

¶The atomic coordinates and structure factors have been deposited in the Protein Data Bank, Chemistry Department, Brookhaven National Laboratory, Upton, NY 11973 (reference 200D).

*F*222, yielded cell dimensions of $a = 59.05$, $b = 81.21$, and $c = 26.75$ Å, thus showing a deviation of up to 1.5% in cell axis length. Consequently, the Harker sections of the difference Patterson map of d(TAAC^{Br}CC) were noisy and difficult to interpret because of the lack of isomorphicity. Recently, we solved the crystal structure of d(C_3T) at 1.4-Å resolution (11). In an attempt to solve the structure of d(TAACCC) by molecular replacement, we used the four-stranded C_3 part of d(C_3T) as a starting model and added bromine to the central cytidine base at position 5. Rotation and translation search with this model at various resolution ranges of the d(TAAC^{Br}CC) diffraction data always led to the same orientation of the molecule in the lattice, clearly indicating that the asymmetric unit contained two strands of d(C_3) arranged as a parallel duplex involving C·C$^+$ base pairs stacked at ≈6.4-Å intervals. The position of the molecule showed an orientation of the helical axis parallel to the diagonal of the crystallographic *a*/*c*-plane in accordance with the predominant Patterson pattern. After several cycles of rigid-body refinement of two strands of d(C^{Br}CC) using 12- to 2.7-Å data, the difference map allowed us to identify most of the missing parts of the molecule. After we fit in the remaining residues, we carried out simulated annealing refinement, leading to an *R* factor of 24% with data between 12 and 1.85 Å above the 1σ level (based on F_o). At this stage, we switched to the native data and performed rigid-body refinement followed by positional and individual temperature-factor refinement. During the entire process of refinement, no constraints were applied. The *R* factor of the current model is 19% for 2122 reflections above the 1σ level (based on F_o) between 12- and 1.85-Å resolution. The free *R* factor (17) value based on a random 10% subset of the reflections is 24%. The rms deviation is 0.03 Å for bonds and 4.0° for angles based on the nucleic acid dictionary of XPLOR (18). Thirty-six water molecules were included in the structure.

RESULTS

The oligonucleotide d(TAACCC) crystallizes in the orthorhombic space group *F*222, and the crystals diffract to 1.9-Å resolution. There are two strands in the asymmetric unit, and in Fig. 1 symmetry-related strands are numbered identically except for the asterisks. The center of the molecule (Fig. 2 *Upper*) shows that the three cytosines in each of the four chains are organized into an intercalation motif (8). At the top and bottom, the 5′ TAA sequences have two different conforma-

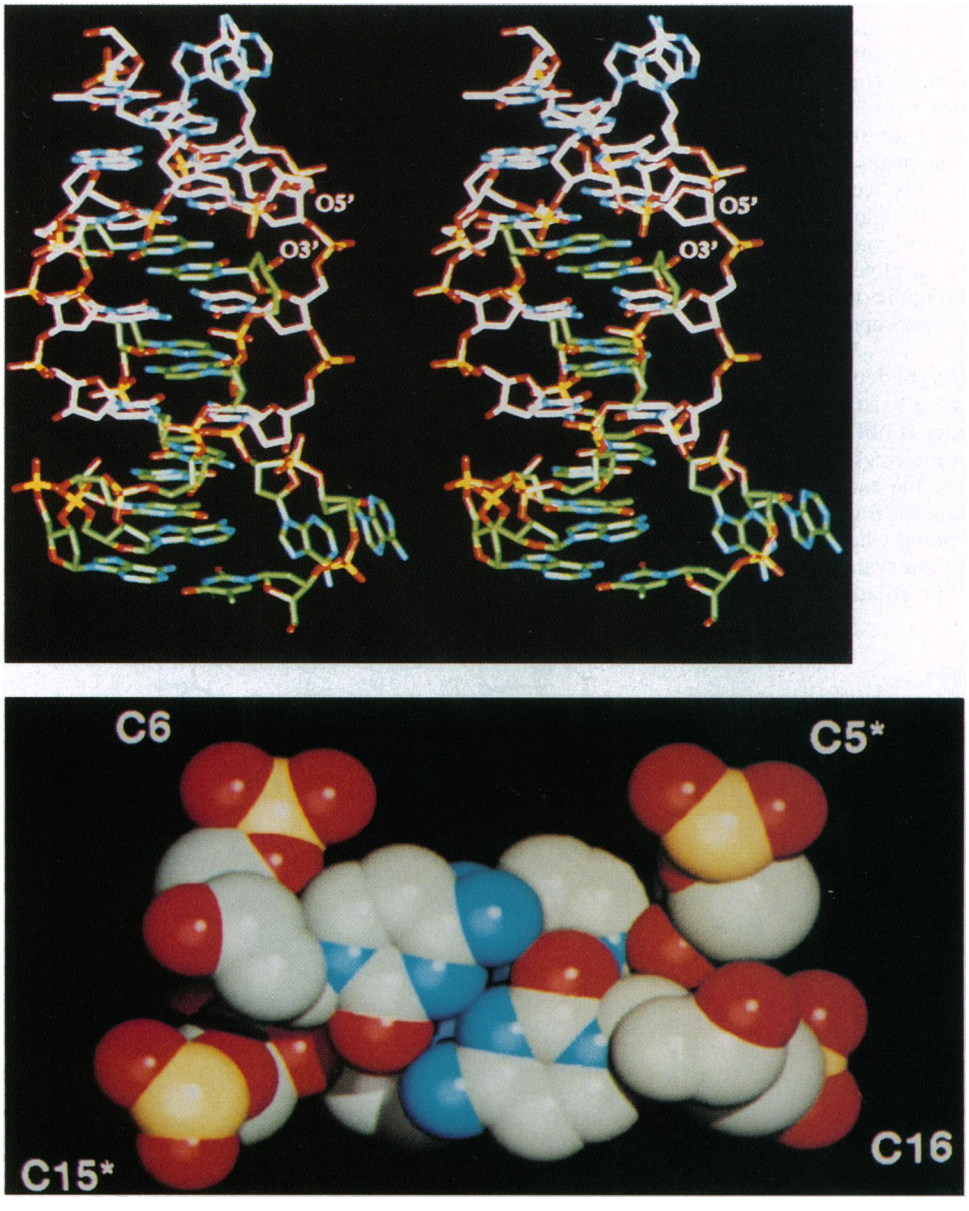

FIG. 2. Views of d(TAACCC). (*Upper*) A stereo diagram showing the entire molecule. The two pairs of strands in the asymmetric unit have different color. The upper pair has a white backbone, and the lower one has a green backbone. The center of the molecule is composed of intercalating cytosine residues held together by hemiprotonated C·C$^+$ base pairs. Near the top the 5′ thymine (labeled O5′) can be seen to form a Hoogsteen base pair with adenine A3. Adenine A2 is stacked above it. The other 5′ end (upper right) has the two adenine residues A12 and A13 projecting away from the molecule in a stacked conformation. Those are used in building the lattice. The terminal 5′ thymine T11 of that strand loops back to form a reverse Watson–Crick base pair with A2. It should be noted that the O5′ of thymine T1 is close to the O3′ that is labeled for cytosine C6*. Oxygen is red, nitrogen blue, and phosphorous yellow. (*Lower*) A van der Waals model showing two base pairs viewed perpendicular by the helix axis. Cytosine C6* is hydrogen-bonded to cytosine C16*. The base pair C5*·C15* is behind it. The phosphate groups of C5* and C6 on the top are rotated so that they are pointing almost vertically, while the phosphate groups from C15* and C16 at the bottom are pointing away from the helix axis. A bridging water molecule (not shown) is found between the amino group of C6 on the left to the phosphate group of C5* at the upper right.

tions. The strand at the upper left curls about and ends with its 5′ end very close to the 3′ end of another molecule. Adenine residue A3 and thymidine residue T1 form a Hoogsteen base pair, which is stacked upon the cytosine C·C$^+$ base pair C4·C14. Adenine A2 caps the loop and stacks upon the Hoogsteen base pair. In the other strand, A13 and A12 project away from the molecule where they stack on each other. These two bases provide the interactions used for building up the lattice as they are stacked on molecules above and laterally to build a three-dimensional system. Thymine T11 curves back to the molecule where it forms a reverse Watson–Crick base pair with adenine A2.

Careful inspection of the center of the molecule with the intercalated cytosine residues reveals that there are two broad grooves and two narrow grooves in the molecule. Unlike the NMR structure (8–10), the broad grooves differ from each other in a significant way. The groove at the right rear (Fig. 2 *Upper*) has the phosphate groups extending away from the center of the molecule. The broad groove at the left front has the phosphate groups bent over toward each other. This is shown more clearly in Fig. 2 *Lower*, which is a van der Waals view down the center of the molecule in which cytosine C16 on the lower right is paired to cytosine C16 at the upper left. Immediately behind it, cytosine C5* is paired to cytosine C15*. When looking down the axis of the molecule, it is clear that the phosphate groups at the top are rotated upwards toward the broad groove, while on the bottom they are oriented away from the center. The upper groove of the molecule in Fig. 2 *Lower* is heavily hydrated. The N4 amino group on cytosine C6 is hydrogen bonded to a water molecule (not shown), which bridges over and hydrogen-bonds to the rotated phosphate oxygen on C5 from an opposite strand. This occurs systematically in the upper groove in Fig. 2 *Lower*. Bridging water molecules of this type are not found in the bottom broad groove (Fig. 2 *Lower*). A similar system of bridging water molecules is found in the crystal structure of d(C_3T) (11), but not in d(C_4) (12). The C_3 segment of d(TAACCC) is virtually superimposable on that segment of d(C_3T).

As seen in Fig. 2 *Upper*, the four-stranded intercalated cytosine segment has a small right-handed twist. The angular twist between covalently linked cytosines is not uniform, but the average rotation is 19.7°. The cytosine rings do not stack upon each other; stacking is confined to the exocyclic amino and carbonyl groups. The stacking distance is found to be 3.13 Å, somewhat reduced from the 3.4 Å found when rings stack over each other. However, the stacking interval of the T1·A3 base pair on the paired cytosine rings immediately below it is 3.4 Å.

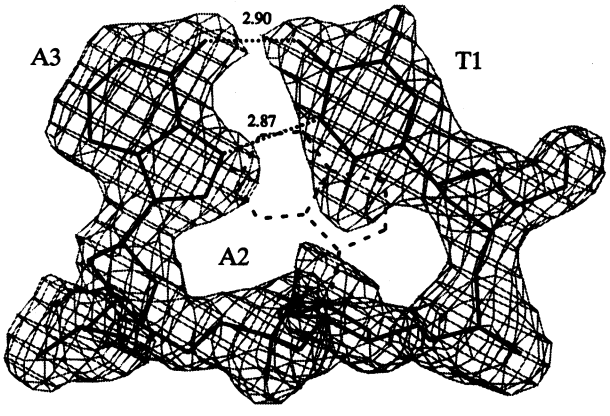

FIG. 3. $2F_o$–F_c electron density map shows the position of the Hoogsteen base pair between T1 and A3. The electron density map is contoured at 1σ. The position of the capping A2 behind the Hoogsteen base pair is shown in dashed lines. Hydrogen bond lengths are shown in Å.

The effect of rotating the phosphate groups in the upper broad groove in Fig. 2 *Lower* leads to a phosphate–phosphate distance across the upper groove of 12.6 Å, compared with this distance across the lower groove of 16.9 Å. The phosphate rotation stabilized by bridging water molecules has the effect of separating the phosphate groups in the two antiparallel chains that are in van der Waals contact in the narrow grooves. The average phosphate–phosphate separation across the narrow grooves (on the sides of Fig. 2 *Lower*) is 6.7 Å. All of the glycosyl bonds are in the *anti* conformation in this structure. The most common pucker found in the sugar rings is a C3′ *endo* pucker, although some residues also have C2′ *endo* configuration. Full details of the conformation will be described elsewhere.

DISCUSSION

It is interesting to note that isomorphous crystals of d(TAACCC) were grown over a wide range of pH, ranging from pH 5.5 to pH 7.5. The C·C$^+$ base pairs depend upon hemiprotonation of the cytosine. This hemiprotonation was first observed in small-molecule crystals (19) but later was established in polymers of both ribo- and deoxycytidylic acid (20–22). For deoxycytidylic acid polymers, the protonated cytosine structure was stable up to pH 7. The stability of the crystal lattice may have raised the pK for hemiprotonation even higher than pH 7.

The loop at the end of the C quartet is stabilized by three interactions. (*i*) T1 is hydrogen-bonded to A3 (Fig. 3). (*ii*) The loop is stabilized by stacking interactions. Fig. 3 shows in dotted outline the position of A2, which is stacked upon the T1·A3 Hoogsteen base pair. (*iii*) In addition, the phosphate group found between A2 and A3 is stabilized by a hydrogen-bonded interaction with the N4 amino group of cytosine C4 (Fig. 4); likewise, the phosphate group between A12 and A13 is hydrogen-bonded to the N4 amino group of C14. All of these interactions combine to form a very tight loop involving the TAA sequence at the 5′ end of the molecule. Fig. 4 also has a jagged arrow showing that the 5′-OH of T1 is very close to the 3′-OH of C6* (also see Fig. 2 *Upper*). These two residues are so close to each other that a small rotation of the dihedral angle around the C4′–C5′

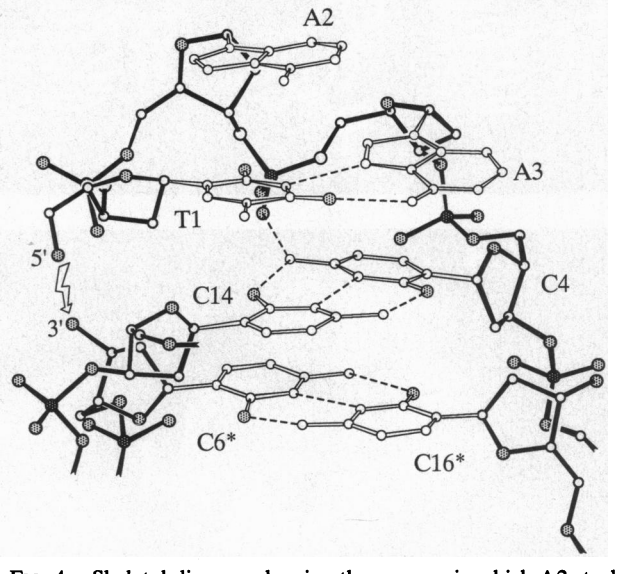

FIG. 4. Skeletal diagram showing the manner in which A2 stacks upon the Hoogsteen base pair involving T1 and A3. These, in turn, are stacked upon the C14·C4 hemiprotonated base pair. It can be seen that the amino group of C4 is hydrogen-bonded to the phosphate group between A2 and A3 in a manner that stabilizes the conformation. The jagged arrow shows the close proximity of the 5′-OH of T1 and the 3′-OH of C6*.

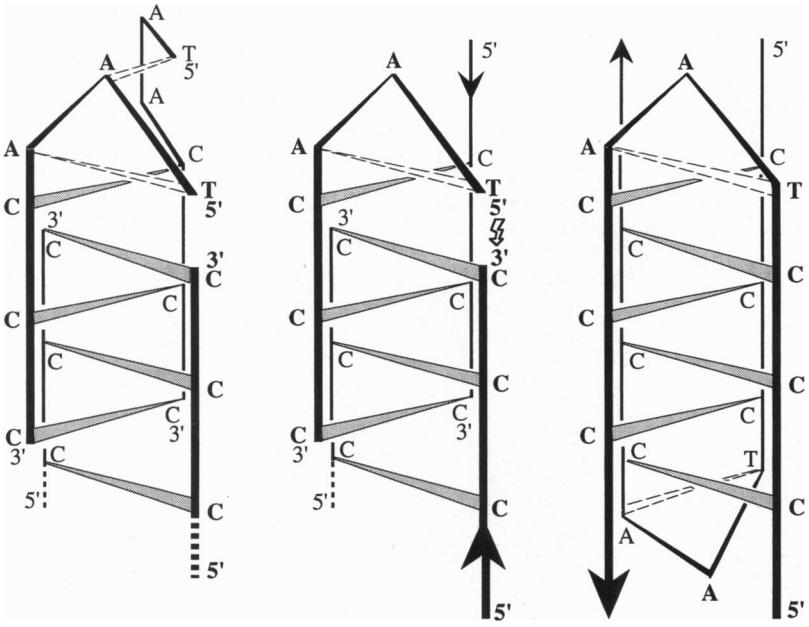

FIG. 5. Diagram of telomere C-strand interactions. (*Left*) Schematic diagram showing the conformation illustrated in Fig. 4. Tapering bonds that are stippled and bounded by solid lines represent the C·C$^+$ hemiprotonated base pairs; those that are open and bounded by dashed lines represent hydrogen bonding between adenine and thymine residues. (*Center*) The 5′ TAA that is not involved in the loop interaction is removed, and a jagged arrow indicates that the 5′-OH of T and the 3′-OH of C are close enough to form a phosphodiester bond. This bond is formed in Fig. 5 *Right*. (*Right*) Diagram illustrating the manner in which two loops of the metazoan telomere can interact to form a stable intercalated four-stranded motif capped by the TAA loops at either end.

bond of T1 puts its O5′ close enough to the C6* O3′ so that a phosphate group could be found in that position. The total movement of O5′ is less than 2 Å. This structure with a close proximity between 5′-OH and 3′-OH of two different strands is similar in some ways to the structure of a DNA duplex built out of a dodecamer paired to two different DNA hexamers (23). In that case, the 5′-OH and the 3′-OH were close enough so that a continuous chain could be formed if a phosphate group were present to fill the gap.

The C-strand telomere structure strongly suggests that the loop may provide a mechanism for stabilizing the interaction of two pairs of C-strand repeats, as shown schematically in Fig. 5. Fig. 5 *Left* illustrates the close proximity of the 5′-OH of T1 and the 3′-OH of C6*. In Fig. 5 *Center* the TAA segment of the molecule in the rear is removed, and the jagged arrow shows that a phosphate group could be introduced there, making a loop with A at the apex on top of the Hoogsteen A·T base pair. If a similar loop were formed at the bottom (Fig. 5 *Right*), this would suggest that the telomeres have the potential to stably interact and thus to recognize other C-strand telomere sequences.

The crystal structures of the *Oxytricha* G-rich telomere (2) show that two looping strands can interact to make a stable four-stranded structure with four guanines hydrogen-bonded in a plane. This is topologically similar to the four-stranded intercalated structure shown in Fig. 5 *Right*, yet there is an important distinction. Where it has been studied, the telomere end has a single-strand extension of two repeats of the G-rich segment (1). Two of these ends could pair together to make a four-stranded structure (2). However, the C-rich strands are all paired in duplexes with G-rich strands. For the C-rich strands to loop out, negative supercoiling would be required to extrude two cruciform loops, one containing cytosines and the other guanines (12). If these cruciforms each involved two repeats, then they would form four-stranded structures by interacting with similar cruciforms on other molecules. Alternatively, if they involved four repeats, each could form its own four-stranded structure.

At present there is no conclusive evidence demonstrating that either four-stranded guanine or four-stranded cytosine structures are found in biological systems. However, it is likely that the DNA duplex in the telomere may differ from standard B-DNA. In yeast, the transcriptional regulator repressor/activator protein (RAP1) is found to bind to the telomeric repeats, where it induces a distortion of the double-helical structure (24). Further, intrachromosomal telomeric repeats in the Chinese hamster ovary cell line are found to have increased sensitivity to irradiation compared with the rest of the genome (25). These findings suggest that the structure may be altered. Finally, since several proteins have been identified that interact specifically with G-quartet structures (4–7), it may be reasonable to anticipate that other proteins binding specifically to four-stranded cytosine structures will also be discovered.

The presence of a stable loop associated with the four-stranded cytosine structure reinforces the possibility that conformations of this type may be found in the telomere. We also wonder if the stability of the loop is why the TAA sequence found in the multicellular animal C-rich-strand telomere repeats has been preserved for such a long evolutionary period.

This research was supported by grants at the Massachusetts Institute of Technology from the National Institutes of Health, the National Science Foundation, the Office of Naval Research, the American Cancer Society, the National Aeronautics and Space Administration, and a U.S. Department of Energy grant to R.M.

1. Blackburn, E. H. (1991) *Nature (London)* **350,** 569–573.
2. Kang, C. H., Zhang, X., Ratliff, R., Moyzis, R. & Rich, A. (1992) *Nature (London)* **356,** 126–131.
3. Smith, F. W. & Feigon, J. (1992) *Nature (London)* **356,** 164–168.
4. Weisman-Shomer, P. & Fry, M. (1993) *J. Biol. Chem.* **268,** 3306–3312.
5. Fang, G. & Cech, T. R. (1993) *Cell* **74,** 875–885.
6. Liu, Z., Frantz, J. D., Gilbert, W. & Tye, B.-K. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 3157–3161.
7. Schierer, T. & Henderson, E. (1994) *Biochemistry* **33,** 2240–2246.
8. Gehring, K., Leroy, J.-L. & Guéron, M. (1993) *Nature (London)* **363,** 561–565.
9. Ahmed, S., Kintanar, A. & Henderson, E. (1994) *Nature Struct. Biol.* **1,** 83–88.
10. Leroy, J.-L., Guéron, M., Mergny, J.-L. & Hélène, C. (1994) *Nucleic Acids Res.* **22,** 1600–1606.
11. Kang, C. H., Berger, I., Lockshin, C., Ratliff, R., Moyzis, R. & Rich, A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 11636–11640.
12. Chen, L., Cai, L., Zhang, X. & Rich, A. (1994) *Biochemistry* **33,** 13540–13546.
13. Moyzis, R. K., Buckingham, J. M., Cram, L. S., Dani, M., Deaven, L. L., Jones, M. D., Meyne, J., Ratliff, R. L. & Wu, J.-R. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 6622–6626.
14. Meyne, J., Ratliff, R. L. & Moyzis, R. K. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 7049–7053.

15. Meyne, J., Baker, R. J., Hobart, H. H., Hsu, T. C., Ryder, O. A., Ward, O. G., Wiley, J. E., Wurster-Hill, D. H., Yates, T. L. & Moyzis, R. K. (1990) *Chromosoma* **99,** 3–10.
16. Riethman, H. C., Moyzis, R. K., Meyne, J., Burke, D. T. & Olson, M. V. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 6240–6244.
17. Brünger, A. T. (1992) *Nature (London)* **355,** 472–474.
18. Brünger, A. T., Karplus, M. & Petsko, G. A. (1989) *Acta Crystallogr. A: Found. Crystallogr.* **45,** 50–61.
19. Marsh, R. E., Bierstedt, R. & Eichhorn, E. L. (1960) *Acta Crystallogr.* **15,** 310–316.
20. Langridge, R. & Rich, A. (1963) *Nature (London)* **198,** 725–728.
21. Inman, R. B. (1964) *J. Mol. Biol.* **9,** 624–637.
22. Hartman, K. A., Jr., & Rich, A. (1965) *J. Am. Chem. Soc.* **87,** 2033–2039.
23. Aynami, J., Coll, M., van der Marel, G. A., van Boom, J. H., Wang, A. H.-J. & Rich, A. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2526–2530.
24. Gilson, E., Roberge, M., Giraldo, R., Rhodes, D. & Gasser, S. M. (1993) *J. Mol. Biol.* **231,** 293–310.
25. Alvarez, L., Evans, J. W., Wilks, R., Lucas, J. N., Brown, J. M. & Giacca, A. J. (1993) *Genes Chromosomes Cancer* **8,** 8–14.