Analysis of DNA Replication Intermediates Suggests Mechanisms of Repeat Sequence Expansion*

Janaki Veeraraghavan, Marie L. Rossi, and Robert A. Bambara†
From the Department of Biochemistry and Biophysics and the Cancer Center, University of Rochester School of Medicine and Dentistry, Rochester, New York 14624

We previously developed a system to investigate the mechanism of repeat sequence expansion during eukaryotic Okazaki fragment processing. Upstream and downstream primers were annealed to a complementary template to overlap across a CAG repeat region. Annealing by the competing primers lead to structural intermediates that ligated to expand the repeat segment. When an equal number of repeats overlapped on the upstream and downstream primers, a 2-fold expansion was expected, but no expansion occurred. We show here that such substrates do not expand irrespective of their repeat length. To reveal mechanism, we tested different hairpin loop intermediates expected to form and facilitate ligation. Substrates configured to form large loops in either the upstream or downstream primer allow expansion. Large or small fixed position single loops allowed expansion when located at least six nucleotides up- or downstream of the nick. Fixed loops in both primers, simulating a double loop intermediate, allowed expansion as long as each loop was nine nucleotides from the nick. Thus, neither the double loop configuration required to form with equal length overlaps nor the large single loop configuration are fundamental structural impediments to expansion. We propose a model for the expansion mechanism based on the relative stabilities of single loop, double loop, hairpin, and flap intermediates that is consistent with the observed expansion efficiency of equal and unequal overlap substrates. The model suggests that the equilibrium concentration of double loop intermediates is so vanishingly small that they are not likely contributors to sequence expansion.

Repeat sequences are distributed widely in all organisms, forming the micro- and mini-satellite regions of their chromosomal DNAs (1). Triplet repeat sequences have attracted particular attention, because they are involved in pathogenesis of at least 14 neurological disorders (2). Of interest are the CAG, CGG, and GAA repeat tracts that are present in the normal population in lengths of 10–25 triplets and show significant length polymorphisms. In a subset of the population repeat lengths reach the relatively stable pre-mutational length of 30–50 repeats, which then undergo large intergenerational expansion by mechanisms that are poorly understood. Repeat sequences present in coding regions expand to a smaller extent than repeats that are located in 3′-untranslated region or regulatory regions that show expansion into thousands of repeats (2).

Locus-specific expansion of CAG/CTG and CGG/CCG sequences suggests that the instability is an inherent property of repeat DNA (3). One characteristic of such DNA is the ability of repeat sequences to slip and mispair because of partial self-complementarity in the region. This gives rise to secondary structures in DNA (4, 5). Although slip mispairing can occur at all repeat sequences, only GAA, CGG, and CTG repeats exhibit expansion in vivo (6). These sequences have the additional quality that they can form secondary structures with high stability. Single strands of CTG and CGG repeats have higher melting temperatures than repeats of the other triplets (7–9). Furthermore, the single strands of these sequences form more stable hydrogen-bonded fold-back structures than those of the other triplets, as determined by NMR (7, 10). In *Escherichia coli* and yeast, CTG and CGG repeats show significantly higher propensity for expansions and deletions than repeat tracts of non-structure-forming sequences, indicating that DNA structure plays a role in instability in vivo (11, 12).

Slip mispairing can occur when DNA is in a single stranded form during replication, repair, and recombination (13, 14). Examination of imperfect repeat sequences in patients showed that repeats were most frequently added on the 3′ end of the segment (15). This polarity in expansion suggested that the mechanism is associated with synthesis of DNA. Moreover, synthesis by DNA polymerase *in vitro* was found to pause frequently across repeat sequences indicating the presence of secondary structures (13, 16). This decreased rate of synthesis was expected to further facilitate structure formation.

Additional orientation dependence was observed for repeat instability first in yeast and *E. coli* and more recently in mammalian cell extracts and cell lines (17–19). Expansion and deletion of CTG repeats occurred mostly on the lagging strand, implicating lagging strand synthesis as a source of instability (12, 18, 20, 21). CTG sequences tended to expand when present on the Okazaki fragment strand, whereas they underwent deletions when present on the lagging strand template (12, 18, 22). This correlates with the particular stability of secondary structure in CTG hairpins, which were found to be more stable than CAG hairpins.

Based on the ability of repeat sequences to form stable secondary structure and their propensity to expand when present on the lagging strand, Gordenin et al. (23) proposed a model for repeat sequence expansion. They hypothesized that during lagging strand synthesis, strand displacement within the repeat sequence will result in generation of a flap that could form stable secondary structure. Such secondary structures, they proposed, would be resistant to processing by the structure-
specific 5’ flap endonuclease (FEN1) that is involved in resolution of Okazaki fragments (23, 24). Re-annealing of the flap without the removal of repeats within the flap sequence would then result in generation of sequence expansion in the daughter strand. Indeed yeast FEN1 null mutants show an increased propensity to expand repeat sequences (20, 25). Further, CTG and CCG containing flaps were found to be resistant to FEN1 cleavage in vitro (26).

We have previously described a model system designed to recapitulate strand displacement within a CAG/CTG repeat region in vitro, to study the cis and trans factors that influence sequence expansion (27). The oligonucleotide model consists of a template strand containing 10 CAG repeats that are flanked by random sequences. To its 3’ end we annealed a complementary downstream primer with CTG repeats at its 5’ end. An upstream primer with a varying number of CTG repeats at its 3’ end was also annealed, such that it strand-displaced varying lengths of downstream primer into a 5’ flap or caused slip mispairing that could result in formation of expansion intermediates (27).

We found that the repeat sequence substrate and DNA ligase I alone were sufficient to allow expansion by ligation of the overlapping strands (27). Addition of FEN1 to this reaction resulted in a decrease in expanded product and increase in correctly sized DNA segment. However, as the number of repeats on the upstream primer was increased such that more repeats were displaced into a 5’ flap on the downstream primer, FEN1 cleavage of these flaps and the subsequent formation of correct sized product were greatly decreased (27). These results were consistent with the proposed model for sequence expansion resulting from decreased processing of repeat containing flaps. Surprisingly, whereas flap cleavage decreased with increase in primer overlap, ligation of the overlapped segments did not committantly increase. On the contrary, as the repeat number on the upstream primer approached that on the downstream primer, ligation efficiency dropped precipitously. This suggested that repeat sequence expansion to lengths of 2-fold or greater may not occur by simple joining of unprocessed flaps that re-annal to ligatable intermediates.

In the current study, we examine why ligation efficiency of triplet repeat sequences decreases with increase in the length of overlap. To establish mechanism, we determined the properties of slipped DNA intermediates that can mediate ligation-based expansion during lagging strand synthesis and processing. We find that very few slipped DNA intermediates are refractory to ligation-based expansion. The factor limiting expansion of certain length overlaps during Okazaki fragment processing is a feature of the DNA itself and not of the transacting factors. We propose a model that is based on the propensity of repeat DNA to form certain types of intermediates and the stability of these intermediates as the basis for the sequence expansion.

**EXPERIMENTAL PROCEDURES**

**Materials—Oligonucleotides** were synthesized either by Integrated DNA Technologies (Corvalis, IA) or by Genosys Biotechnologies (Woodlands, TX). Radionucleotides [α-32P]dUTP and [γ-32P]ATP (3000–6000 Ci/mmol) were obtained from PerkinElmer Life Sciences. T4 polynucleotide kinase and the Klone fragment of E. coli DNA polymerase I (labeling grade) were from Roche Diagnostics. Recombinant N-terminal His-tagged human DNA ligase I was cloned into the pHis expression vector and was expressed and purified from E. coli as reported previously (27). Yeast FEN1 expression and purification was performed as described previously (28). All other reagents were the best available commercial grade.

**Oligonucleotide Substrates—**Repeat sequence model substrates were generated as described previously (27). Briefly, oligonucleotides were designed to mimic the last steps of Okazaki fragment processing. Each substrate consisted of a template strand annealed with a downstream primer at its 5’ end and an upstream primer at its 3’ end. The sequence for oligonucleotides used are listed in Table I. The template and downstream primers contained CAG/CTG repeats, having complete complementarity. Upstream primers containing varying number of repeats, which overlap with the CTG repeats of the downstream primer, are added to complete substrate formation. The (CTG)n substrates each contained n number of CTG repeats at the 3’ ends of the upstream primer. In the final substrate, the upstream and downstream primers contain sequences complimentary to the template and overlap across the repeat region. These overlapping regions should result in a dynamic equilibrium between the primers. This would produce intermediates involving strand displacement of one primer to form flap and loop structures involving one or both primers. A nick substrate that lacked overlapping CTG repeats was used as a control for ligation efficiency.

Fixed CTG bubbles on downstream or upstream primer were configured by annealing a primer with six internal CTG repeats, flanked by non-repeat sequences, to a template complementary to the non-repeat sequences. Such primers were aligned to create a nick with an adjacent primer. This annealing configuration would result in the formation of a six-CTG repeat hairpin at varying distances from the nick.

Prior to annealing the downstream primer was radiolabeled at its 3’ end. The downstream primer (10 pmol) was annealed to labeling template T1 (25 pmol) generating a recessed 3’ end and was then extended with [α-32P]dUTP using Klenow polymerase at 37° C for 3 h. Unincorporated radionucleotides were removed using Micro Bio-spin 30 chromatography columns (Bio-Rad). Upstream primers were labeled at the 5’ end with [γ-32P]ATP in an end labeling reaction with T4 polynucleotide kinase. The labeled primers were isolated and purified on a 10% 7 M urea denaturing polyacrylamide gel.

Substrates were generated by annealing the downstream primer, template, and upstream primer at a molar ratio of 1:2:4, respectively. DNA ligase, template, and upstream primers were diluted into 50 μl of TE (10 mM Tris-Cl, pH 8, and 1 mM EDTA) and heated to 100 °C for 5 min. The reaction was placed at 70° C and allowed to slowly cool to room temperature.

**Enzyme Assays—**Assays contained indicated amounts of substrate, DNA ligase I, or yeast FEN1 in reaction buffer in a final volume of 20 μl. Reaction buffer contained 30 mM Hepes, pH 8.0 (diluted from a 1 M stock), 40 mM KCl, 4 mM MgCl2, 0.5 mM ATP, 0.01% Nonidet P-40, 0.5% inositol, 0.1 mg/ml bovine serum albumin, and 1 mM dithiothreitol. Assays were assembled on ice and then incubated at 37°C for 15 min. The reactions were stopped by the addition of 10 μl of termination dye (95% formamide (v/v), 1 mM EDTA with 0.5% bromphenol blue and 1 mM EDTA) and heated to 100 °C for 5 min. The reaction was placed at 70°C and allowed to slowly cool to room temperature.

**Analysis for Stability of Structural Intermediates—**Free energies for various structural intermediates formed by overlapping CTG repeats was calculated using Mfold, a software used to determine structures formed by single stranded DNA. The software calculates the free energy of formation for various structures formed by user specified sequences because of intra-strand base pairing (29). The free energy estimates predict the stability of various intermediates and the abundance in solution of these structures (30). The algorithm allows users to force the formation of specific base pairs and therefore various structures formed by a given sequence. This makes possible the estimation of free energy of various intermediates expected to form by overlapping repeats.

The three primers used to form the overlapping repeat substrate were entered into M-fold as one single stranded DNA molecule. To one single stranded DNA molecule, the 5’ end of the downstream primer was used as the 5’ end of the single stranded DNA. To its 3’ end, the 5’ end of the template followed by the 5’ end of the upstream primer was linked using a six-nucleotide random sequence designated the N string. The N string allows for the formation of hairpin loops at the junction of upstream and downstream primer to generate a unique structure. The structures formed by 10 CTG repeats on the downstream and upstream primers competing to base pair to the complimentary CAG repeat on the template was obtained at 150 mM NaCl and 10 mM MgCl2. Double loop and flap-loop intermediates were formed by forcing specific bases in the sequence to base pair as permitted by the software. The free energy of flap formation was obtained by subtracting the free energy of structure formed by a string of 10 CTG repeats from the free energy of the fold-back flap structure.

1 J. SantaLucia, Jr., M. Zuker, A. Bommarito, and R. J. Irani, unpublished results.
RESULTS

We set out to test the hypothesis that equal length overlapping repeat segments lead to the formation of intermediates unfavorable for ligation. The intention was to use the results of that analysis to discern the intermediates and mechanisms that favor expansions.

To assess expansion efficiency, complementary oligonucleotides were annealed to simulate the template, downstream, and upstream primers. The results of this analysis allowed us to discern the intermediates and mechanisms that favor expansions.

### Table I

| Oligonucleotide sequence                  |
|------------------------------------------|
| Upstream and downstream primers are aligned in 5'-3' to template primers shown in the 3'-5'. Repeat and mismatch regions are indicated in bold. n, number of repeats; MM, mismatch; R, repeats. |

**Overlapping repeat substrates**

**Upstream primer (U).** n = 1–20.

- (CTG)ₙ
  - TCCCAAGTCACAGCC(CTG)ₖ₃’
  - TCCCAAGTCACAGCC(CTA)ₖ₃’
- Nick
  - TCCCAAGTCACAGCC

**Downstream primer (D).** n = 5 for template primer (T) and 5, 6, 8, and 10 for downstream primer (D).

- (CTG)ₙ
  - TCCCAAGTCACAGCC(CTG)ₖ₃’
  - TCCCAAGTCACAGCC(CTA)ₖ₃’
  -尼克
  - TCCCAAGTCACAGCC
  - TCCCAAGTCACAGCC

**5 Repeat primers.**

- n = 1–20.
  - (CTG)ₙ
  - TCCCAAGTCACAGCC(CTG)ₖ₃’
  - TCCCAAGTCACAGCC(CTA)ₖ₃’
- Nick
  - TCCCAAGTCACAGCC

**10 or 20 repeat primers.**

- n = 10 or 20 for template and 10, 15, 18, and 20 for downstream primer.
  - (CTG)ₙ
  - TCCCAAGTCACAGCC(CTG)ₖ₃’
  - TCCCAAGTCACAGCC(CTA)ₖ₃’
- Nick
  - TCCCAAGTCACAGCC

**Fixed loop and mismatch substrate**

- (6R - 18In)CGACCGTGCCAGCCTAAACCACTG(CTG)₆TGCTACACTTGCCCGTGC.₃’
- (6R - 9In)CGACCGTGCCAGCCTAAACCACTGTGCTACACT(CTG)₆TGCCCGTGC.₃’
- (6R - 6In)CGACCGTGCCAGCCTAAACCACTGTGCTACACTTGC(CTG)₆CCGTGC.₃’
- (6R - 3In)CGACCGTGCCAGCCTAAACCACTGTGCTACACTTGCCCG(CTG)₆TGC.₃’
- (6R)CGACCGTGCCAGCCTAAACCACTGTGCTACACTTGCCCGTGC.₃’

**Fig. 1.** Triplet repeat expansion decreases as overlap length increases to match the template repeat length. a, a template primer (T) containing 10 CAG repeats was annealed to a downstream primer (D) with 10 CTG repeats and upstream primers (U) of various repeat lengths to simulate strand displacement within the repeat segment. Overlap length was increased on the upstream primer from 0–10 repeats ((CTG)ₙ), b, 5 fmol of substrate with increasing repeat overlap was incubated at 37 °C for 10 min in the presence of increasing amounts of human DNA ligase I (LIG I) (0.5, 1, and 2 fmol). The number of repeats on the upstream primer is indicated above the figure. The nick substrate has no repeats on the upstream primer. The numbers on the right indicate the base pair size of substrate and products.

**RESULTS**

We set out to test the hypothesis that equal length overlapping repeat segments lead to the formation of intermediates unfavorable for ligation. The intention was to use the results of that analysis to discern the intermediates and mechanisms that favor expansions.
and upstream fragments in the configuration formed during Okazaki fragment synthesis and processing (27). The template oligonucleotide included an internal repeat segment of 10 CAG repeats, to which a complementary 3′/H11032 downstream primer containing a 10-CTG repeat and 5′/H11032 upstream primers with varying numbers of repeats were annealed such that they overlapped within the repeat segment (Fig. 1a). The overlapping configuration of upstream and downstream primer results in transient strand displacement of the downstream primer into a 5′ flap. However, as the substrates contain sequence repeats, they can slip and mispair to generate a nick that can be joined by DNA ligase to expand the daughter strand. Our experimental design does not address the potential effects of template folding. Such folding might be expected to promote repeat sequence contraction.

**Decreased Expansion Efficiency with Increase in Overlap Length**—In the presence of increasing amounts of human DNA ligase I, overlapping repeat segments on the upstream and downstream primers were joined, resulting in expansion of the daughter strand (Fig. 1b). The ligation efficiency increased as
the repeat number on the upstream primer was increased to six CTGs (Fig. 1b, lanes 1–24). However, as the repeat length on the upstream primer was increased to seven, there was a sharp decline in the formation of expanded product. Expansion by ligation was completely suppressed as the overlap length was increased from 8 to 10 repeats even at large enzyme excess (Fig. 1b, lanes 29–36). The same decrease in ligation efficiency with length of overlap also occurred when the number of repeats was fixed for the upstream primer and varied on the downstream primer (data not shown). This indicates some form of symmetry in the phenomenon. Both results would appear to be inconsistent with the acceleration of repeat expansion with increasing repeat length observed in vivo (2). Further, this result suggests a size limitation on expansions based on direct ligation of DNA replication intermediates.

Overlapping repeat segments of equal length must form double loop intermediates to be joined by DNA ligase I for expansion. For example, in a 10-10 repeat overlap, a 9-1 intermediate will carry a 9-repeat loop on one primer and a 1-repeat loop on the other. Theoretically, many such loop intermediates may be formed. The lack of expansion products with 8-10 or 10-10 overlap (Fig. 1b, lanes 29–36) indicates that either the formation of a suitable double loop intermediate or the position of these loops might interfere with ligation and limit expansion. This issue lead us to examine the substrate structures that influence ligation.

Suppressed Ligation Is Independent of Repeat Length—We first determined whether the decrease in ligation efficiency correlated with size of the repeat segment on the template. Templates were designed to contain either 5 or 20 CAG repeats, and each was annealed to a downstream primer containing 5 or 20 CTG repeats, respectively. These substrates were then annealed to upstream primers with 3' repeats of varying lengths and tested for expansion by DNA ligase (Fig. 2). Surprisingly, the substrates with both 5 and 20 repeat-containing templates showed results similar to those with the 10-CAG substrate. Upstream primers producing small overlaps allowed formation of expanded products on both the 5- and 20-repeat substrate (Fig. 2, a and b, lanes 1–15). However, increasing the overlap length to 5 repeats on the (CAG)<sub>5</sub> substrate resulted in complete prevention of expansion, as did increasing the repeat length to 18 in the case of (CAG)<sub>20</sub> substrate (Fig. 2, a, lanes 16–20, and b, lanes 21–25). Again this phenomenon was independent of the enzyme concentration. These results verified that expansion does not occur when overlap regions contain about equal numbers of repeats, irrespective of the size of the repeat segment. This suggests that an equal length overlap across a repeat region, when produced during lagging strand DNA replication, would also inhibit ligation.

A Mechanism for 2n Expansion—Lack of 2n expansion when repeat segments overlap suggested that the position or size of a loop on either upstream or downstream primer is inhibitory to the activity of DNA Ligase I. To manipulate the size and position of the loop in a systematic manner, we employed primers such that all of the additional repeats for expansion (up to 2n) were contributed by only one primer. In different experiments the extra repeats would be only on either the upstream or the downstream primer. For example, in the substrate used in Fig. 3a, we annealed downstream primers with 10, 15, 18, and 20 CTG repeats to a template with 10 CAG repeats and an upstream primer that has no repeats. The upstream primer is complementary to the random sequence on the 3' end of the template. We tested repeat loops of various sizes on both up- and downstream primers that were annealed to (CAG)<sub>5</sub> and (CAG)<sub>10</sub> templates.

We observed that on these substrates DNA ligase could join the two primers irrespective of the size and site of the loop relative to the nick (Fig. 3). Interestingly, a 2n expanded product was readily formed, resulting from joining of primers when (CTG)<sub>20</sub> and (CTG)<sub>20</sub> primers were annealed to a (CAG)<sub>5</sub> and (CAG)<sub>10</sub> template, respectively (Fig. 3, a, lanes 16–20, and b and c). Apparently, template size-independent 2n expansion can occur in our system when there is slip mispairing on only one of the fragments and may occur by a similar mechanism in vivo. A possible reason for the observed lack of 2-fold expansion in the previous experiments is the overlapping configuration in which the substrates were annealed across the repeat region. The overlapping segments may interact with the template and each other in a way that prevents the formation of stable double loop intermediates that could be substrates for expansion.

Comparing the efficiency of ligation of the loops formed on upstream and downstream primers on (CAG)<sub>5</sub> and (CAG)<sub>10</sub> substrates, there were some subtle but noteworthy differences. We observed lower levels of ligation on the (CAG)<sub>5</sub> template (Fig. 3b). Downstream and upstream (CTG)<sub>10</sub> repeat-containing primers ligated, but at less than 30% efficiency, whereas the nick substrates were joined at 80% efficiency. Ligation efficiency of upstream and downstream (CTG)<sub>20</sub> primers annealed to the (CAG)<sub>10</sub> template was about 50%.

Slip mispairing can result in formation of either a single large loop comprising the entire excess repeat, anywhere relative to the nick, or multiple smaller loops distributed across the repeat region. The fact that (CTG)<sub>20</sub> primers ligated with comparatively high efficiency (about 50%) suggested that a single large loop is formed, as smaller loops would be equally as inhibitory as larger ones, in both smaller and larger repeat segments, if located at the same distance from the adjacent DNA fragment. Further, the lower ligation efficiency on the smaller repeat segment suggests that a relatively large loop, formed closer to the nick on smaller repeat segments, interferes with the ligase or is unstable so as not to form ligatable intermediates.

Loop Position and Not Size Modulates DNA Ligase Activity—The substrates used thus far in the study are dynamic in that they allow the formation of multiple loop intermediates. The effect of any one intermediate is indistinguishable from that of the others. To study the effect of a specific intermediate on the activity of DNA ligase, we designed substrates that contain repeat loops of specific size at precise distances from the nick. This was achieved by making primers that contained a central segment of two or six CTG repeats flanked on both sides by non-repeated sequence. Each primer was then annealed to a template complementary only to its non-repeat sequences. This fixed the annealing position of both sides of the primer and produced a repeat-containing loop in the middle. Another completely complementary primer was then annealed in an adjacent position on the template to form a nick. With this approach the loop could be positioned at various distances to only one side of the nick. The substrate could also be designed so that both fragments contained non-complementary repeats, giving rise to a double loop, one in each primer. A repeat loop could then be placed at a fixed distance from the nick on downstream and/or upstream primers (Fig. 4). DNA ligase activity was tested on all these substrate configurations.

A six-CTG repeat-fixed loop on the downstream primer formed expansion products in the presence of DNA ligase with increasing efficiency as the loop distance was increased from six to 18 nucleotides relative to the nick site (Fig. 4a). Moreover, with increasing loop distance, the expansion products formed with comparable efficiency to the ligation of a nick site (Fig. 4a). Although a loop located six nucleotides away showed...
Fig. 3. 2n expansion occurs in a non-overlapping configuration. Repeat length was increased either on the upstream or the downstream primer alone such that slip mispairing occurred on only one segment. a, expansion was assayed as described above when repeat length was increased on the downstream primer alone D(CTG)\textsubscript{10+n}. b and c, quantitation of expansion products for (CAG)\textsubscript{5} and (CAG)\textsubscript{10} repeat substrates as repeat length (nr) was increased on the upstream (U) or downstream (D) primer alone. Expansion products were quantitated and plotted as percent ligation at increasing concentrations of DNA ligase I for various substrates.
Fig. 4. **Distance of fixed repeat loops from the nick impact ligation.** The effect of structures formed by repeat sequences as a result of slip mispairing was tested by configuring substrates to contain six non-complimentary CTG repeats at various distances (n) on either downstream (a) or both sides of a nicked DNA (b). Substrates were incubated with increasing amounts of DNA ligase I (1, 5, 10, and 20 fmol) for 10 min at 37 °C. A schematic of the substrates employed is indicated above the figure.
some expansion product, when the loop was located three nucleotides either 5' or 3' to the nick, no DNA ligase activity was observed. Similar results were obtained when (CTG)$_n$ loops were located on the upstream primer, as well as with two CTG repeat-fixed loop substrates (data not shown). This suggested that when the loop is located three to six nucleotides from the nick it interferes with the ability of DNA ligase to carry out its reaction. One possible reason is that annealing of only three bases is rather unstable and that this intermediate might spend most of its time as a flap structure. Although six-base pair annealing may not be significantly more stable, detection of some ligation suggests that it does form the loop intermediate with sufficient stability for the ligatable intermediate to be captured by the enzyme.

Various fixed (CTG)$_n$ double loop intermediates were assayed for generation of expansion products in the presence of DNA ligase I (Fig. 4b). As with the single loops located at a distance of three nucleotides from the nick, double loops at this distance also inhibited ligation (Fig. 4b, lanes 1–5). Consistently, we also observed that loops located at a distance of nine or more nucleotides relative to the nick site did not interfere with the activity of DNA ligase (Fig. 4b, lanes 11–20). However, unlike with single loops of (CTG)$_n$ with two (CTG)$_n$ loops present six nucleotides from the nick, very little ligation was observed (Fig. 4b, lanes 6–10). Comparable results were obtained when double (CTG)$_n$ loops were tested in this assay, ruling out a role for exact characteristics of the (CTG)$_n$ loop as a factor affecting ligation (data not shown). The ability of DNA ligase to seal nicks with single but not double loops at a distance of six nucleotides implies that the double loop configuration can interfere with DNA ligase binding, if the loops are sufficiently close to the nick site. This result is in agreement with the published footprint for Chlorella virus DNA ligase, which spans 19–21 nucleotides on either side of the nick (31).

Our data, in fact, indicate that DNA ligase binding is flexible and can accommodate single loops located at the far reaches of its binding region. Significantly, this result also shows that ligation of a double loop intermediate is not limiting for the formation of 2n expansion when repeat segments overlap, provided they are located beyond the footprint for DNA ligase binding.

In addition to steric interference to DNA ligase binding, the lack of ligation with single/double (CTG)$_n$ loops at three-six nucleotides from the nick might be attributable to a decrease in annealing efficiency or stability for these substrates. However, it would be experimentally difficult to determine the exact contribution of these factors.

**Location and Size of the Loop on the Upstream Primer Influences DNA Ligase Activity**—To measure whether the size of the loop or its location alone influence ligation we designed substrates to contain mixed sized and positioned double loops. Substrates contained (CTG)$_2$ and (CTG)$_6$ loops at a distance of six nucleotides on either side of the nick (Fig. 5). (CTG)$_2$ loops at a distance of six nucleotides on the upstream primer formed more expansion product than when (CTG)$_6$ loops were located on the upstream primer (Fig. 5, lanes 1–10). The effect of relative distance of the loop from the nick was measured using substrates with (CTG)$_6$ loops located at six and nine nucleotides from the nick. (CTG)$_6$ loops at a distance of six nucleotides on the upstream primer and nine nucleotides on the down-
stream primer showed lower product formation than the mirrored substrate (Fig. 5, lanes 16–25). Apparently, larger loops on the upstream primer six nucleotides from the nick are inhibitory to ligation. This suggests that the flexibility in DNA ligase binding is greater on the downstream primer than on the upstream primer. The enzyme must tolerate abnormal structures on the downstream side of the binding site more readily than on the upstream side. This analysis gives an apparent footprint of six nucleotides on the downstream and nine or more nucleotides on the upstream primer for the human DNA ligase I. This is the first estimation of the human DNA ligase footprint to our knowledge. Further, the binding of the ligase to nicked DNA seems to be oriented differently than suggested by the footprint published for the \textit{Chlorella} virus enzyme (31). The \textit{Chlorella} virus ligase had a larger footprint on the downstream primer than the upstream primer.

**Instability of Three-base Pair Annealing**—We proposed above that absence of DNA ligase activity on loops located three nucleotides from the nick resulted from inherent instability in annealing only three annealed nucleotides, further exacerbated by the presence of a mismatched loop. To confirm this we designed substrates to contain either six mismatches (Table I, primers D(MM) annealed to T2 and Un) or a six-nucleotide T loop beginning three nucleotides from a nick on the downstream primer (Fig. 6). The mismatch substrate would determine the ability of only three annealed nucleotides to ligate, whereas the T loop substrate would measure the effect of a loop on annealing of three base pairs. As predicted the mismatch substrate generated very little ligation product, and ligation was undetectable with the T-loop substrate (Fig. 6). This suggests that any structure or mismatch located three nucleotides from the nick could destabilize base pairing at the nick at 37°C to an extent that is very inhibitory to ligation.

Only a subset of potential intermediates are formed when CTG repeats overlap—The data presented so far suggest that some of the double loop intermediates that can be formed by overlapping repeats should in theory be joined by DNA ligase. For example, on a (CTG)\textsubscript{10} overlap, all intermediates that have two (CTG)\textsubscript{5} loops on the upstream and downstream primers at a distance of greater than six nucleotides from the nick should be ligated. The number of such loop intermediates in reality should be finite and detectable by ligation. We detect expansion products with (CTG)\textsubscript{7} overlaps but not on (CTG)\textsubscript{8} or (CTG)\textsubscript{10}. However, the number of possible intermediates that fulfill the apparent criteria for ligation that can be formed by each of these substrates is unlikely to be greatly different. The inability to detect expansion with overlaps greater than (CTG)\textsubscript{8} suggests that with increasing overlaps the fraction of all structural forms of the substrate that are ligatable is reduced. This would occur if the ligatable intermediates were particularly unstable compared with non-ligatable intermediates.

One of the possible intermediates is the loop-flap intermediate in which the downstream or upstream primer carries the excess repeats in un-annealed flaps in addition to interior loops. The presence and persistence of flap intermediates can be measured by the activity of the 5′-flap structure-specific endonuclease FEN1 (24). Theoretically, if all possible loop-flap intermediates are present in equilibrium, in the presence of FEN1, cleavage products of these intermediates should be detectable in the form of a three-base pair ladder. Our previously published results showed that with the (CTG)\textsubscript{10} overlaps, increasing amounts of FEN1 produced only one major cleavage product that corresponds to the formation of a 5′-flap containing all the excess repeats. Only a very minor amount of intermediate cleavage was detected (27). This suggests that not all loop-flap intermediates were formed with equal probability. To verify that this is a feature of structure-forming repeats, we compared cleavage patterns of (CTG)\textsubscript{10} overlaps with non-structure-forming (CAA)\textsubscript{10} overlaps. CAA repeats, although capable of slip mispairing, do not form stable hairpin loops with

---

**Fig. 6.** Three-base pair annealing is unstable. A substrate with a six-nucleotide mismatched segment (dotted line) or a T loop (dotted loop) located at a distance of three nucleotides downstream from the nick were configured by annealing either the D(MM) or the D\textsubscript{r} primers to T\textsubscript{2} and U\textsubscript{n} primers, to test for the stability of annealing three-base pair sequence. Substrates were incubated with DNA ligase as described under “Experimental Procedures.”
internal base pairing like CTG repeats. We propose that CAA overlaps would contain all possible intermediates with equal probability as they are energy neutral for formation. However, with the CTG repeats, larger, self-annealed loops with many hydrogen bonds are likely to be more stable than flap or loop-flap intermediates, which are likely to have fewer hydrogen bonds to stabilize them. This would make folded-back flaps much more probable than other intermediates. We expect that the cleavage pattern for CAA repeats would reflect its ability to form intermediates whereas that of CTG would not.

An assay of cleavage products for (CAA)10 and (CTG)10 substrates with FEN-1 yielded a single product corresponding to cleavage at the base of a single large flap of about 30 nucleotides for the CTG repeats at all enzyme concentrations used (Fig. 7). At lower enzyme concentrations CAA repeats also showed cleavage at a similar position (Fig. 7, lanes 2 and 9). However, at higher enzyme concentration products that correspond to cleavage of smaller flaps was observed with the CAA repeat but not the CTG repeat substrates (Fig. 7, lanes 5–7 versus 12–14). Further, the smaller CAA cleavage products appeared to be a three-base pair ladder, as there were 10 major bands corresponding to all the possible 5' flap intermediates that can be formed on the downstream primer by slip mispairing. This cleavage pattern indicates that only one 5' flap intermediate is formed by the (CTG)10 repeats whereas the (CAA)10 repeats form all possible flap-loop intermediates. Closer examination of FEN1 activity on these substrates revealed more efficient cleavage of the (CTG)10 substrate in comparison with the (CAA)10 substrate, especially at lower FEN1 concentrations. This was contrary to our expectation that non-structure-forming repeats would be better FEN1 substrates. However, it can be explained by the ability of (CAA)10 repeats to form all possible flap intermediates, many of which are unlikely to be ideal substrates for FEN1. We also observe that cleavage activity of FEN1 is significantly lower on the repeat sequences compared with an ideal double flap substrate used as control for this experiment (Fig. 7, lanes 15 and 16). At the lowest concentration (0.5 fmol) of FEN1 (Fig. 7, lane 16), about 90% of the double flap substrate is cleaved whereas substantially lower amounts of the repeat substrates are cleaved. This is consistent with our observation that FEN1 activity is inhibited in flaps containing repeat sequences.

These results confirm our hypothesis that overlapping CTG repeat sequences do not form all possible intermediate structures. The result suggests that whereas non-structure-forming repeats generate all the possible intermediates in the overlapping configuration, CTG repeats form a limited set of structures that are inaccessible to DNA ligase or FEN1 activity.

Structural Intermediates and Their Free Energy for Formation—To determine the relative stability and free energies of various structural intermediates, which can theoretically form by overlapping repeats, we applied the program DNAfold (29). This program was used to analyze structures formed by single stranded repeat containing DNA as described under “Experimental Procedures.” Calculation of free energies showed significant differences among the various intermediates. The most stable intermediates were the 5' and 3' flaps that had formed hairpin intermediates (Fig. 8) with free energy of –91.2 kcal/mol. The most unstable intermediate was the double loop intermediate with free energy of –70 kcal/mol. 5' and 3' flaps and flap-loop
intermediates, which have smaller flaps and an internal loop, were of intermediate stability at -83 and -75 kcal/mol, respectively (Fig. 8).

With our expansion substrate, the algorithm generated hairpin flap structures for the repeat region that were located either on the 3′ or the 5′ repeat. As we had suspected, this analysis indicated that the (CTG)_{10} hairpin flap is the most stable and energetically most favorable intermediate. This structure, if formed at the 3′ end of the upstream primer of the actual substrate, would be inaccessible to FEN1. If present on the 5′ end of the downstream primer, it would interfere with FEN1 activity as shown previously (27, 32). This is in agreement with our previous experiment, which indicated that the (CTG)_{10} substrate only formed a 30-nucleotide flap that was refractory to FEN1 cleavage consistent with the formation of structure. At the same time these hairpin flap intermediates are not ligatable and therefore unable to produce a 2n expansion. Additionally, the double loop intermediate has a significantly higher free energy of formation (70.2 versus 91.2 kcal/mol) indicating its instability.

The difference of 21 kcal/mol in the formation of two intermediates translates into a 10^{14} lower abundance in solution. Moreover, the free energy difference between an unstructured repeat flap and the hairpin flap is also high (8 kcal/mol) so that the hairpin flap is likely to be far more abundant than the flap itself. This explains both our failure to obtain 2n expansion in experiments with DNA ligase and the inhibition of FEN1 cleavage. Furthermore, there is a significant difference in the free energies among the various flap and flap-loop intermediates. This suggests that the substrate shows a preference for formation of the fold-back hairpin structure that exists in a free energy trough, rather than being in a dynamic equilibrium among the various intermediates (Fig. 8).

**DISCUSSION**

We have investigated the mechanism of sequence expansion during Okazaki fragment processing within a repeat region. Analyzing the expansion properties of repeats in vitro we had observed a curious property of overlapping repeats (27). They do not expand when the number of repeat units in upstream and downstream primers are matched to that in the template. We present data here showing that many types of loops formed by slip mispairing in repeat regions can produce nicks that are susceptible to the activity of human DNA ligase. Contrary to our expectation, even double loop intermediates did not interfere with the activity of DNA ligase I when located outside of its footprint across the nick. However, expansion of the repeats to 2n length was observed only when all of the excess repeats were contained on either upstream or downstream primer. Additionally, we determined that CTG repeats fail to form flap-loop intermediates that are formed by similarly configured CAA repeats. Taken together, these results suggest that the inability to expand to 2n length by direct ligation is a function of the DNA structures assumed by repeats in overlapping configurations.

The results presented here lead us to propose a model for DNA replication-associated sequence expansion. It is based on the stability and abundance of the structural intermediates that would be formed by overlapping repeats on upstream and downstream primers during lagging strand synthesis. We propose that unequal overlaps form internal hairpins that can be readily ligated and cause expansion of the daughter strand, whereas the larger equal length overlaps form hairpin flaps located on one or the other primer (Fig. 8). Implicit in this proposal is that a double loop intermediate required to generate 2n expansion is not formed by equal, or nearly equal, length overlapping primers. This is evident in higher AG values for the double loop intermediates, showing higher energy requirement for formation. The higher free energies are possibly because of the loss of freedom of movement required to form multi-branch loops and hairpin loops (29). This is consistent with the observation that a single large loop, with the excess repeats located on either side of the nick, is favored for ligation over the distribution of excess repeats into numerous smaller loops when the repeats overlap. Indeed, excess CTG repeats in a heteroduplex DNA form a single large hairpin at the 3′ end of a repeat region, as determined by mung bean nuclease digestion and electron microscopy (33). Formation of this structure may result from stabilization because of maxi-
minal base pairing and base stacking interactions that directly contribute to the lower free energy and therefore stability of the DNA.

The high stability of the single hairpin structure suggests that during strand displacement, excess repeats are extruded as a flap that folds over to base pair via internal complementarity into a hairpin. Breslauer and co-workers (34) report that, once formed, a sufficiently long single stranded hairpin state is stable under physiological conditions and requires thermal disruption for it to revert to duplex DNA, despite the higher stability of the latter state. Thus, once formed, the hairpin flap intermediate may not equilibrate readily to other intermediates. The large difference in free energies between flaps and flap-loop structures (17.1 kcal/mol) suggests that flap-loop intermediates are rarely, if ever, formed. These latter structures also possess multi-branch loops and hairpin loops that destabilize the structure.

Previously, structurally intermediates formed by repeat sequences were examined using single stranded repeat DNA or heteroduplexed DNA with excess CAG or CTG repeats using nuclease mapping or NMR techniques (5, 7, 33). However, none of these analyses determined dynamics of overlapping repeats and the structures resulting from competitive annealing to the template. Our ability to assess DNA dynamics makes this study especially relevant to the mechanism of expansion resulting from lagging strand synthesis.

Polarity of expansions in vivo and an abundance of evidence from yeast genetics points to a role of lagging strand synthesis in destabilizing repeat sequences (20, 25, 36). We show here that if strand displacement synthesis were to occur within a repeat sequence, slip mispairing followed by ligation can result in moderate length expansion. However, expansion to 2-fold or greater lengths may not occur by direct joining when the repeats are distributed in both primers equally. The observation that in yeast most expansion in triplet repeats during lagging strand synthesis frequently resulted in <2n, and rarely in >2n lengths, supports this view (36, 37).

According to our model, strand displacement of a large number of repeats will result in long-lived hairpin flaps that are refractory to FEN1 and DNA ligase. These nick-flap sites may be prone to double strand breakage. This idea is in agreement with the increase in double strand breakage and recombination in yeast FEN1 null mutants (38, 39). Repair of the long-lived nick-flap by repair or recombination pathways may result in large expansions in vivo. This idea is supported by the fact that CTG/CAG repeat sequences induce recombination and gene conversion in yeast and E. coli (40–43). Alternatively, our ability to visualize 2n expansion when loops are located to any one side of the nick (Fig. 3) suggests that large scale expansion can occur when slip mispairing results during synthesis of the repeat-containing sequence. This is in agreement with previous results from biochemical analysis using self-priming repeat sequences, which shows that CTG repeats slip mispair efficiently into even numbered repeat loops causing expansion during synthesis (14). Moreover, in yeast, frequently expansion was polar and occurred directionally when the repeat segment was interrupted by non-repeat sequence (44). This suggests that slip mispairing during replication was the source of instability.

Slip mispairing during replication alone does not address the mechanism of repeat expansion satisfactorily. Although we present information on mechanisms for expansion of short stretches of triplet repeats, it is well documented that very long segments (>30) are particularly unstable in vivo. Data suggest that apart from lagging strand synthesis, repair and recombination are also a source of repeat instability (40, 42, 45) that might be more effective at promoting expansion of longer sequences. Although a role for structures formed by repeat DNA in producing instability is clear, the exact role that they play in expansion remains unknown. Do these structures signal/recruit trans factors that in turn destabilize the repeat? Evidence for participation of trans factors in repeat instability has been hard to come by. Among cis factors the location of the repeat region relative to origin of replication and methylation status of DNA have been determined to play a role in maintaining stability (17, 19, 35). How these processes may affect expansion is intriguing. Further research into the interaction of repeat DNA with trans acting factors will be important to untangle the influence of the various DNA metabolic pathways on repeat instability.

Acknowledgments—We thank Dr. Leigh A. Henricksen and also the members of the Bambara laboratory for useful discussions. We thank Dr. Jeffrey Hayes and Dr. Sayura Aoyagi for kindly sharing purified human DNA ligase I. We are grateful to Dr. Douglas Turner for invaluable guidance regarding the use of the DNAfold algorithm.

REFERENCES

1. Bennett, P. (2000) Mol. Pathol. 53, 173–183
2. Usdin, K., and Grabcyk, E. (2000) Cell. Mol. Life Sci. 57, 914–931
3. Gehrke, G. M., Tester, D., Thibodeau, S., Almovist, E., Goldberg, Y. P., Hayden, M. R., and McMurray, C. T. (1997) Am. J. Hum. Genet. 60, 879–890
4. Chablain, P. D., and Sinden, R. R. (1998) J. Mol. Biol. 275, 405–411
5. Pearson, C. E., Wang, Y. H., Griffith, J. D., and Sinden, R. R. (1998) Nucleic Acids Res. 26, 816–823
6. Sutherland, G. R., and Richards, R. I. (1995) Proc. Natl. Acad. Sci. U. S. A. 92, 3636–3641
7. Gacy, A. M., Goellner, G., Juranic, N., Macura, S., and McMurray C. T. (1995) Cell 81, 533–540
8. Mitas, M. Y., Aill, D., Kamp, T. J., Chambers, E. J., and Haworth, I. S. (1995) Nucleic Acids Res. 23, 1050–1059
9. Mitas, M. Y., Aill, D., and Haworth, I. S. (1995) Biochemistry 34, 12893–12897
10. Gacy, A. M., and McMurray, C. T. (1998) Biochemistry 37, 9426–9434
11. Ohshima, K., Kang, S., and Wells, R. D. (1996) J. Biol. Chem. 271, 1853–1856
12. Miret, J. J., Pessoa-Blandino, L., and Lahue, R. S. (1998) Proc. Natl. Acad. Sci. U. S. A. 95, 12438–12443
13. Ohshima, K., and Wells, R. D. (1997) J. Biol. Chem. 272, 16798–16806
14. Hartenstein, M. J., Goodman, M. F., and Petruska, J. (2000) J. Biol. Chem. 275, 18162–18180
15. Eichler, E. E., Holden, J. J., Popovich, B. W., Reiss, A. L., Snow, K., Thibodeau, S. N., Richards, C. S., Ward, P. A., and Nelson, D. L. (1994) Nat. Genet. 8, 85–94
16. Kang, S., Ohahima, K., Shimizu, M., Arimbaer, S., and Wells, R. D. (1995) J. Biol. Chem. 270, 27014–27021
17. Cleary, J. D., Nichol, K., Wang, Y. H., and Pearson, C. E. (2002) Nat. Genet. 31, 17–26
18. Freudneunreich, C. H., Stavenhagen, J. B., and Zakian, V. A. (1997) Mol. Cell. Biol. 17, 2090–2098
19. Panagiri, H., and Cleary, J. D., and Pearson, C. E. (2002) J. Biol. Chem. 277, 13926–13934
20. Freudneunreich, C. H., Kantrow, S. M., and Zakian, V. A. (1998) Science 279, 853–856
21. Kang, S., Jaworski, A., Ohahima, K., and Wells, R. D. (1995) Nat. Genet. 10, 213–218
22. Maurer, D. J., O’Callaghan, B. L., and Livingston, D. M. (1996) Mol. Cell. Biol. 16, 6617–6622
23. Gordenin, D. A., Kunkel, T. A., and Resnick, M. A. (1997) Nat. Genet. 16, 116–118
24. Bambara, R. A., Murante, R. S., and Henricksen, L. A. (1997) J. Biol. Chem. 272, 4447–4450
25. Ireland, M. J., Reinke, S. S., and Livingston, D. M. (2000) Genetics 155, 1657–1665
26. Spira, C., Pelletier, R., Rolfsmeyer, M. L., Dixon, M. J., Lahue, R. S., Gupta, G., Park, M. S., Chen, X., Mariappan, S. V., and McMurray, C. T. (1999) Mol. Cell. 4, 1079–1085
27. Henricksen, L., Veeraragavan, J., Chaﬁn, D. R., and Bambara, R. A. (2002) J. Biol. Chem. 277, 22361–22369
28. Xie, Y., Liu, Y., Argue, J., Henricksen, L. A., Kao, H. I., Bambara, R. A., and Alani, E. (2001) Mol. Cell. Biol. 21, 4889–4899
29. Zuker, M. (2000)Curr. Opin. Struct. Biol. 10, 303–310
30. SantaLucia, J., Jr. (1998) Proc. Natl. Acad. Sci. U. S. A. 95, 1460–1465
31. Odell, M. M., and Shuman, S. (1999) J. Biol. Chem. 274, 14033–14039
32. Henricksen, L. A., Tom, S., Liu, Y., and Bambara, R. A. (2000) J. Biol. Chem. 275, 16420–16427
33. Pearson, C. E., Tan, M., Wang, Y. H., Montgomery, S. E., Dar, A. C., Cleary, J. D., and Nichol, H. (2002) Nucleic Acids Res. 30, 4534–4547
34. Volker, J., Makute, N., Plum, G. E., Klump, H. B., and Breslauer, K. J. (2002) Proc. Natl. Acad. Sci. U. S. A. 99, 14700–14705
Replication Intermediates in Repeat Expansion

35. Nichol, K., and Pearson, C. E. (2002) *Genome Res.* **12**, 1246–1256
36. Rolfsmeier, M. L., Dixon, M. J., Pessoa-Brandao, L., Pelletier, R., Miret, J. J., and Lahue, R. S. (2001) *Genetics* **157**, 1569–1579
37. Schweitzer, J. K., and Livingston, D. M. (1998) *Hum. Mol. Genet.* **7**, 69–74
38. Jin, Y. H., Obert, R., Burgers, P. M., Kunkel, T. A., Resnick, M. A., and Gordenin, D. A. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5122–5127
39. Debrauwere, H., Loeillet, S., Lin, W., Lopes, J., and Nicolas, A. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98**, 8263–8269
40. Jankowski, C., Nasar, F., and Nag, D. K. (2000) *Proc. Natl. Acad. Sci. U. S. A.* **97**, 2134–2139
41. Jakupciak, J. P., and Wells, R. D. (2000) *IUBMB Life* **50**, 355–359
42. Jakupciak, J., and Wells, RD. (2000) *J. Biol. Chem.* **275**, 49003–49013
43. Jakupciak, J. P., and Wells, R. D. (1999) *J. Biol. Chem.* **274**, 23469–23479
44. Rolfsmeier, M. L., and Lahue, R. S. (2000) *Mol. Cell. Biol.* **20**, 173–180
45. Richard, G. F., Goellner, G. M., McMurray, C. T., and Haber, J. E. (2000) *EMBO J.* **19**, 2381–2390