Several neuromuscular and neurodegenerative diseases are caused by genetically unstable triplet repeat sequences (CTG-CAG, CGG-CCG, or AAG-CTT) in or near the responsible genes. We implemented novel cloning strategies with chemically synthesized oligonucleotides to clone seven of the triplet repeat sequences (GTA-TAC, GAT-ATC, GTT-AAC, CAC-GTG, AGG-CCT, TCG-CGA, and AAG-CTT), and the adjoining paper (Ohshima, K., Kang, S., Larson, J. E., and Wells, R. D. (1996) J. Biol. Chem. 271, 16784–16791) describes studies on TTA-TAA. This approach in conjunction with in vivo expansion studies in Escherichia coli enabled the preparation of at least 81 plasmids containing the repeat sequences with lengths of 16 up to 158 triplets in both orientations with varying extents of polymorphisms. The inserts were characterized by DNA sequencing as well as DNA polymerase I sequencing, two-dimensional agarose gel electrophoresis, and technical probe analyses to evaluate the capacity to adopt negative supercoiled non-B DNA conformations. AAG-CTT and AGG-CTT form intramolecular triplexes, and the other five repeat sequences do not form any previously characterized non-B structures. However, long tracts of TCG-CGA showed strong inhibition of DNA synthesis at specific loci in the repeats as seen in the cases of CTG-CAG and CGG-CCG (Kang, S., Ohshima, K., Shimizu, M., Amirhaeri, S., and Wells, R. D. (1995) J. Biol. Chem. 270, 27014–27021). This work along with other studies (Wells, R. D. (1996) J. Biol. Chem. 271, 2875–2878) on CTG-CAG, CGG-CCG, and TTA-TAA makes available long inserts of all 10 triplet repeat sequences for a variety of physical, molecular biological, genetic, and medical investigations. A model to explain the reduction in mRNA abundance in Friedreich's ataxia based on intermolecular triplex formation is proposed.

The molecular basis of nine genetic diseases (including fragile X syndrome (FRAXA and FRAXE), myotonic dystrophy (DM), Kennedy's disease, Huntington's disease (HD), spinal muscular atrophy (SMA), Machado-Joseph disease, dentatorubral-pallidoluysian atrophy (DRPLA), and Friedreich's ataxia type 1 (SCA1), and some cases for simplicity, the duplex insert is referred to by the first three letters (i.e. GTA for GTA-TAC). The abbreviations used are: TRS, triplet repeat sequences; pur sequences form triplexes (H-DNA) and (in certain cases) nodular DNA, alternating pur-pyr sequences adopt left-handed Z-DNA, inverted repeats form cruciforms, and repeating A tracts exist in bent (curved) conformations. Some unusual structures were proven to exist in vivo in plasmids (5, 21, 22) and in chromosomes (23). However, no DNA structural studies have been reported on long tracts of TRS in plasmids except for nucleosome positioning by EM (24, 25). A number of biophysical studies have appeared recently (26–32) on short (generally <50 bp) synthetic oligonucleotides with CTG, CAG, CGG, or CGG sequences which are the basis for the concepts of quasi-stable hairpin loops as well as tetraplexes and other ordered conformations. A wide range of investigations on DNA polymers with TRS (reviewed in Ref. 33) that first revealed the influence of DNA sequences on of repeating GTA, which may also be written AGT or TAG; TAC, the complementary strand, may also be written as CTA or ACT. The orientation is 5' to 3' for both designations of the antiparallel strands. In some cases for simplicity, the duplex insert is referred to by the first three letters (i.e. GTA for GTA-TAC).
Seven Triplet Repeat DNA Sequences

properties and structures was described—25 years ago.

Herein, we report novel cloning strategies for long tracts of TRS and the properties of seven cloned TRS. Similar data were described for CTG-CAG\(^3\) (5, 8, 9, 24, 34, 35) and for CGG-CCG\(^3\) (5, 34, 36), and the adjoining paper (37) reports parallel work on TTA-TAA. These investigations were undertaken to understand the properties and conformations of each of the 10 TRS and to serve as a basis for the interpretation of results with the disease sequences, CTG-CAG, CGG-CCG, and AAG-CTT (reviewed in Ref. 5). Although a number of diseases have been characterized recently as TRS syndromes (1–7), AAG-CTT is the only new TRS to be implicated with a disease gene. However, other diseases also show anticipation (reviewed in Refs. 1–6); if a correlation exists between anticipation and triplet repeats, many more diseases may be identified that show this behavior since at least 40 genes containing TRS have been found (reviewed in Ref. 5). Perhaps these disease genes may contain some of the other seven sequences.

MATERIALS AND METHODS

Synthetic Oligonucleotides and Plasmid Construction—Oligonucleotides containing a triplet repeat sequence and a restriction site that was absent in the natural triplet repeat sequence (Table I) were synthesized by the Gene Technology Laboratory, Institute of Developmental and Molecular Biology, Texas A&M University, using an Applied Biosystems 394 synthesizer. 5 μg of oligonucleotide were phosphorylated in 15 μl of a solution containing 50 mM Tris-Cl (pH 7.6), [γ-\(32\)P]ATP (150 μCi, 25 pmol) (DuPont NEN), T4 polynucleotide kinase (30 units) (U. S. Biochemical Corp.), 10 mM MgCl\(_2\), and 10 mM 2-mercaptoethanol at 37°C for 15 min followed by addition of 1 mM ATP and T4 polynucleotide kinase (30 units) and incubated at 37°C for 45 min. The complementary oligonucleotides were mixed, heated at 90°C for 15 min, and slowly cooled to 25°C for 12 h. The annealed oligonucleotides shown in Fig. 1 were purified from a 20% polyacrylamide gel. All inserts were cloned into pUC19 unless indicated otherwise. pUC19 was digested with the restriction enzyme, and calf intestine alkaline phosphatase as described above. The linearized plasmids were digested with the restriction enzyme, the alkaline lysis method (39), and the above experiments repeated. After the ligation reaction by the addition of 95% formamide and 20 mM EDTA, the DNA was fractionated on a 12% denaturing polyacrylamide gel, and the bands were visualized by autoradiography.

Two-dimensional Agarose Gel Electrophoresis—To generate topoisomers of plasmids, 6 μg of DNA was incubated in 100 μl of a solution containing Tris-Cl (pH 7.6), 50 mM KCl, 10 mM 2-mercaptoethanol, 1 mM EDTA, 0–4 μM ethidium bromide, and chicken erythrocyte topoisomerase (40) for 60 min at 37°C. The ethidium bromide and topoisomerase were removed with phenol extraction twice and ether once, and the DNAs were ethanol-purified. Mixtures of topoisomerase populations were subjected to first-dimension gel electrophoresis in 1.25% agarose at 3.3 V/cm at 25°C in 45 mM Tris borate, 1 mM EDTA, the bands were visualized by autoradiography. The second dimension was carried out at a 90° angle to the first dimension at 3.3 V/cm at 25°C in 0.5 × TBE buffer (pH 8.3) containing 20 μM chloroquine diphosphate (Sigma) for 3 h. Electrophoresis in the second dimension was carried out at a 90° angle to the first dimension at 3.3 V/cm at 25°C in 0.5 × TBE buffer (pH 8.3) containing 20 μM chloroquine diphosphate.

Chemical Modifications—The modifications of the plasmids by OsO\(_4\) (Aldrich), diethyl pyrocarbonate (DEPC) (Aldrich), and chloroacetalddehyde (CAA) (Fluka) were performed as described previously (39, 41). CAA was used after distillation. For the OsO\(_4\) modifications, 3 μg of DNA in 100 μl of 0.5 × TBE buffer (pH 8.3) or TAE buffer (pH 4.5) was incubated for 30 min at 25°C with 2 μg OsO\(_4\). The DEPC and CAA reactions were performed for 30 min at 25°C with 10% DEPC and for 60 min at 37°C with 2% CAA, respectively. The DNA samples were terminated by chilling on ice and washed twice with cold ether.

After recovery by ethanol precipitation, the DNAs were divided into two samples and digested with either Sph and HindII or EcoRI and SphI. The overhangs were labeled with [γ-\(32\)P]dATP (10 μCi, 1.7 pmol) (Amersham Corp.) and the Klenow fragment of E. coli DNA polymerase I (10 units) (U. S. Biochemical Corp.). The DNA fragments were isolated by polyacrylamide gel electrophoresis. The purified fragments were dissolved in 100 μl of 10% piperidine (Aldrich), heated to 90°C for 30 min, and then the piperidine was removed by lyophilization. The DNA samples were then fractionated on a 12% denaturing polyacrylamide gel, and the bands were visualized by autoradiography.

RESULTS

Cloning of Triplet Repeat Sequences—We have previously cloned three TRS from natural sources, CTG-CAG\(^1\) (9, 24), CGG-CCG\(^3\) (34, 36, 42), and TTA-TAA (37), with various lengths up to 250 repeats. Long tracts of other triplet repeats have been identified in the human genome (11, 12), but no cloning has been reported yet. For studies on the other seven triplet repeats, it is desirable to do long tracts (\(\approx 200\) bp). Synthetic oligonucleotides have been useful for cloning the desired sequences but only up to a length of \(\approx 100\) bp. Thus, in this study, new cloning strategies with synthetic oligonucleotides were implemented to do much longer tracts of TRS.

The goal of these strategies is to utilize chemically synthe-

R. Gellibahian, M. Shimizu, S. Amirhaeri, S. Kang, K. Oshima, J. E. Larson, Y.-H. Fu, C. T. Caskey, B. A. Oostra, and R. D. Wells, manuscript in preparation.
Seven Triplet Repeat DNA Sequences

| Triplet Repeat | Unique Cleavage Enzyme | Duplex Sequence with Recognition Site |
|---------------|------------------------|-------------------------------------|
| GTA•TAC       | BsaAl                  | (TAC) 15 ATG TAGTACCC ATGCATGCC      |
| GAT•ATC       | EcoRV                  | GGATATCC ATGGTAGT ATG 15            |
| GTT•AAC       | Hpal                   | GGCTATAC ATGATTG AAAGTAGG ACGTGAAC  |
| CAC•GTG       | BsaAl                  | (CAC) 15 CACGTAC ATGCATGCC          |
| AGG•CCT       | Stul                   | AGGCCTGG AGGGCTGG TGCGGACC          |
| TCG•CGA       | NruI                   | GCGCGCGA CGAGCGCT CGT 15            |
| AAG•CTT       | Stul                   | (AAG) 15 AGGCGTG AGGCCTGG GCCTGGGAC |

![Fig. 1. Unique cleavage enzyme sites and oligoduplexes used for cloning of triplet repeat sequences.](Image)

The bars on the top and bottom of the duplex sequences indicate the recognition sites of the unique restriction enzymes. The arrows indicate the cleavage sites.

-sized deoxyribo-oligonucleotides of lengths that can be prepared conveniently containing a unique restriction enzyme site at one end that does not exist in the pUC19 vector (Fig. 1). The duplex oligonucleotide was cloned in both orientations into the HindIII site of pUC19 via blunt-ended ligation. Digestion of the ensuing plasmid with the appropriate restriction enzyme (see Fig. 1, BsaAl in the case of GTA•TAC) provided a linear “vector” for the subsequent cloning of a second unit of the synthetic duplex (Fig. 2A). In general, this process was repeated four times to give inserts of ~60 triplet repeats. In principle, the inserts could be cloned in either orientation. However, we predominantly found direct repeat inserts, not inserts with inverted repeats that might have been deleted due to cruciform formation in vivo (22). This strategy (Figs. 1 and 2A) should give inserts with pure TRS, due to the judicious choices of the unique cleavage enzymes, with no interruptions (also called polymorphisms or mutations). However, after DNA sequencing of some of the plasmids (Table I), interruptions were found that obviously were introduced during their replication. In several cases, longer inserts (Table II) were generated in plasmids derived from pUC18 by utilizing the expansion procedures in E. coli as reported previously (8, 9).

For the cloning of GTA•TAC repeats (Fig. 2A), the complementary synthetic oligonucleotides were annealed to produce a duplex containing a BsaAl site along with TAC repeat units. The duplex was cloned into the HindIII site of pUC19 to produce pRW3151 which has a G to A mutation or the oppositely oriented pRW3152 (Table I). Both plasmids were digested with BsaAl, and the same duplex oligomer used for the first cloning was cloned into the BsaAl sites to produce pRW3153 or pRW3154 (Table I), respectively. For pRW3153, small deletions of the repeats occurred during the cloning. In general, deletions occurred in multiples of 3 bp. The above procedure based on digestion with BsaAl and insertion of a duplex oligomer was repeated to produce long tracts of GTA•TAC repeats (Table I). pRW3155 and pRW3156 were obtained from pRW3153 and pRW3154, respectively. pRW3157 and pRW3159 were produced from pRW3155, pRW3451 containing 77 repeats of GTA•TAC was produced from pRW3157 by the insertion of an additional 16 repeats. pRW3851 was obtained by recloning of pRW3451 into pUC18NptI. For pRW3159 and pRW3156, deletions of 1 and 7 repeats, respectively, occurred from the desired products.

The same procedure was used for the cloning of GAT•ATC, GTT•AAC, CAC•GTG, AGG•CCT, and TCG•CGA triplet repeats, using the appropriate restriction enzymes and synthetic oligonucleotides (Fig. 1). According to the cloning strategy, 16 pure repeats should be introduced into the plasmid in the first cloning cycle, and the repetitive cloning results in the introduction of an additional 16 repeats for each subsequent cloning cycle. However, for the plasmids shown in Table I, polymorphisms including base mutations, insertions, and deletions were observed as seen in long tracts of TRS in the human genome (43–50). For example, pRW3161 has one A deletion in the case of GTA•A (22). This strategy (Figs. 1 and 2A) should give inserts with pure TRS, due to the judicious choices of the unique restriction enzyme site along with TAC repeat units. pRW3162 was produced from pRW3161 with the insertion of (GAT)16, pRW3163 has ((GAT)10AT(GAT)3) added to (GAT)16 from pRW3161, pRW3164, pRW3165, pRW3166, and pRW3169 are expanded products with small expansions in the longest GAT tracts. pRW3441 was formed by elongation of pRW3169 with the (GAT)16 oligomer, but the ensuing plasmid has one T to A mutation, and pRW3442 has an additional (GAT)3 insertion in addition to (GAT)16.

pRW3171 had an insertion of extra repeats; 26 triplet repeats containing one G to A mutation were found instead of the 16 repeats expected in the first cloning. pRW3172 and pRW3173 were predicted products, but one T to G and one G to A mutation occurred in the cloning of pRW3174. pRW3871 was derived from pRW3173 by recloning into pUC18NptI. For both GAT•ATC and GTT•AAC, only inserts in one of the two possible orientations were found (Table I); the reason for this behavior is uncertain. Furthermore, repeated attempts to clone SacI-HindIII fragments containing the GAT-ATC inserts into pUC18NptI were unsuccessful. Interestingly, this sequence was proposed (51) to have an unusual structure since it does not bind actinomycin D.

For CAC•GTG, the first cloning cycle produced pRW3181 and pRW3182. pRW3182 was the starting material for pRW3183 and pRW3184 with 16 and 14 repeats of GTG inserted, respectively. For pRW3184, a new flanking unit (GTACG) was observed, but this was not seen for pRW3183. However, this phenomenon is not the same as that seen for pRW3163 described above. For pRW3185 and pRW3186, derived from pRW3183, (GT)2 and GTG, respectively, were deleted from the predicted repeats. pRW3186 was the starting material for pRW3187, pRW3188, pRW3189, and pRW3421. pRW3424 was obtained from pRW3421 with the insertion of (GTG)14. pRW3425 is an expanded product of pRW3421 with 12 more repeats. pRW3427, pRW3428, and pRW3429 were elongation products from pRW3425. The plasmids derived from the cycle following the production of pRW3185 and pRW3186 from pRW3183 do not have additional GTACG flanking units, indicating that in vivo expansion of the triplet along with mutations of a few base pairs has occurred as seen in the case of pRW3183.

For AGG•CCT repeats, the first cloning cycle produced
pRW3191 and pRW3192, pRW3193, pRW3194, and pRW3195 were elongated products from each cloning step. These new inserts contained an additional 14, 12, and 9 repeats, respectively, indicating that the perfect AGGCCCT repeats were likely to be unstable to give deletions from the predicted repeats. For pRW3196, two duplex units were inserted to give CCAGG polymorphisms in the CCT repeats. pRW3197, pRW3198, and pRW3891 were produced by recloning of pRW3194, pRW3195, and pRW3196, respectively, into pUC18NotI.

For TCGCA repeats, pRW3411 and pRW3412 were obtained by the first cloning cycle. pRW3413 had one C and a TCGG triplet deletion in the insert. pRW3414 and pRW3415 were elongation products from each cloning step.

In general, as the length of triplet repeats increases, polymorphisms are more likely to occur, probably due to instability (5, 9). Therefore, polymorphisms seem to stabilize the triplet repeats, as seen in the case of human disease genes (43–45, 49, 50).

Cloning of AAGCTT Triplets—Since there are no appropriate restriction enzymes for cloning the AAGCTT repeat sequence using the method described above, we modified the strategy used for the other six TRS (Fig. 2B). In this case, two duplexes were formed by annealing two pairs of complementary oligonucleotides. Each pair contained an overhanging end complementary to that of the other duplex, were mixed together in the cloning procedure to give a longer duplex containing a StuI site (pRW3436 and pRW3437). For both plasmids, deletions of Gs were found to give polymorphisms. In
addition, in some cases only one of the component duplexes was cloned (pRW3433 and pRW3434). Either pRW3436 or pRW3437 was digested with Stul, and the duplex for AAG repeats (Fig. 1, bottom line) was cloned into the Stul site. pRW3438 was obtained from pRW3436 with a 4-repeat deletion from the expected number, and pRW3439 and pRW3440 were

| Triplet repeat | Top strand | Name    | n   | Sequence                  |
|----------------|------------|---------|-----|----------------------------|
| GTA-TAC        | (GTA)$_n$  | pRW3151 | 16  | (GTA)$_{16}$ATA(GTA)$_1$   |
|                |            | pRW3153 | 29  | (GTA)$_{29}$ATA(GTA)$_7$   |
|                |            | pRW3155 | 45  | (GTA)$_{45}$ATA(GTA)$_7$   |
|                |            | pRW3159 | 60  | (GTA)$_{60}$ATA(GTA)$_7$   |
|                |            | pRW3157 | 61  | (GTA)$_{61}$ATA(GTA)$_7$   |
|                |            | pRW3451 | 77  | (GTA)$_{77}$ATA(GTA)$_7$   |
|                | (TAC)$_n$  | pRW3152 | 16  | (TAC)$_{16}$               |
|                |            | pRW3154 | 32  | (TAC)$_{32}$               |
|                |            | pRW3156 | 41  | (TAC)$_{41}$               |
|                |            | pRW3851 | 77  | (TAC)$_{77}$ATA(TAC)$_{69}$|
| GAT-ATC        | (GAT)$_n$  | pRW3161 | 15  | (GAT)$_{15}$GT(GAT)$_9$    |
|                |            | pRW3162 | 31  | (GAT)$_{31}$GT(GAT)$_2$    |
|                |            | pRW3163 | 44  | (GAT)$_{44}$GT(GAT)$_{10}$AT(GAT)$_3$ |
|                |            | pRW3164 | 49  | (GAT)$_{49}$GT(GAT)$_{10}$AT(GAT)$_3$ |
|                |            | pRW3165 | 55  | (GAT)$_{55}$GT(GAT)$_{10}$AT(GAT)$_3$ |
|                |            | pRW3166 | 59  | (GAT)$_{59}$GT(GAT)$_{10}$AT(GAT)$_3$ |
|                |            | pRW3167 | 62  | (GAT)$_{62}$GT(GAT)$_{10}$AT(GAT)$_3$ |
|                |            | pRW3159 | 66  | (GAT)$_{66}$GT(GAT)$_{10}$AT(GAT)$_3$ |
|                |            | pRW3441 | 82  | (GAT)$_{82}$GT(GAT)$_{10}$AT(GAT)$_{14}$GAA(GAT)$_{14}$ |
|                |            | pRW3442 | 91  | (GAT)$_{91}$GT(GAT)$_{10}$AT(GAT)$_{28}$ |
| GTT-AAC        | (GTT)$_n$  | pRW3171 | 26  | (GTT)$_{26}$ATT(GTT)$_{13}$ |
|                |            | pRW3172 | 42  | (GTT)$_{42}$ATT(GTT)$_{21}$ |
|                |            | pRW3173 | 58  | (GTT)$_{58}$ATT(GTT)$_{25}$ |
|                |            | pRW3174 | 74  | (GTT)$_{74}$ATT(GTT)$_{13}$GT(GTT)$_{12}$ATT(GTT)$_3$ |
|                | (AAC)$_n$  | pRW3871 | 58  | (AAC)$_{58}$ATA(AAC)$_{16}$ |
| CAC-GTG        | (CAC)$_n$  | pRW3181 | 16  | (CAC)$_{16}$               |
|                | (GTG)$_n$  | pRW3182 | 16  | (GTG)$_{16}$               |
|                |            | pRW3184 | 30  | (GTG)$_{30}$               |
|                |            | pRW3183 | 32  | (GTG)$_{32}$               |
|                |            | pRW3185 | 46  | (GTG)$_{46}$GT(GTG)$_{16}$  |
|                |            | pRW3186 | 46  | (GTG)$_{46}$               |
|                |            | pRW3187 | 56  | TG(GTG)GT(GTG)$_{10}$AT(GTG)$_{2}$ |
|                |            | pRW3189 | 58  | (GTG)TG(GTG)$_{10}$(GTG)$_{14}$ |
|                |            | pRW3188 | 62  | (GTG)GG(GTG)$_{10}$TG(GTG)$_{52}$ |
|                |            | pRW3421 | 62  | (GTG)$_{62}$AT(GTG)$_{48}$  |
|                |            | pRW3425 | 74  | (GTG)$_{74}$AT(GTG)$_{48}$  |
|                |            | pRW3424 | 76  | (GTG)$_{76}$AT(GTG)$_{48}$  |
|                |            | pRW3428 | 79  | (GTG)$_{79}$AT(GTG)$_{48}$  |
|                |            | pRW3427 | 81  | (GTG)GG(GTG)$_{10}$AT(GTG)$_{52}$ |
|                |            | pRW3429 | 81  | (GTG)GG(TGT)$_{10}$AT(GTG)$_{52}$ |
| AGG-CCT        | (AGG)$_n$  | pRW3192 | 13  | (AGG)$_{13}$AG(AGG)G(AGG)  |
|                |            | pRW3197 | 42  | (AGG)$_{42}$               |
|                |            | pRW3198 | 53  | (AGG)$_{53}$               |
|                |            | pRW3891 | 74  | (AGG)$_{74}$CCAGG(AGG)$_{16}$ |
|                | (CCT)$_n$  | pRW3191 | 16  | (CCT)$_{16}$               |
|                |            | pRW3193 | 30  | (CCT)$_{30}$               |
|                |            | pRW3194 | 42  | (CCT)$_{42}$               |
|                |            | pRW3195 | 53  | (CCT)$_{53}$               |
|                |            | pRW3196 | 74  | (CCT)$_{74}$CCAGG(CCT)$_{58}$ |
| TCG-CGA        | (TCG)$_n$  | pRW3411 | 16  | (TCG)$_{16}$               |
|                |            | pRW3413 | 30  | (TCG)$_{30}$TCG(TCG)$_{2}$ |
|                |            | pRW3414 | 46  | (TCG)$_{46}$TCG(TCG)$_{14}$ |
|                |            | pRW3415 | 62  | (TCG)$_{62}$TCG(TCG)$_{34}$ |
|                | (CGA)$_n$  | pRW3412 | 16  | (CGA)$_{16}$               |
| AAG-CTT        | (AAG)$_n$  | pRW3433 | 20  | (AAG)$_{20}$               |
|                |            | pRW3436 | 37  | (AAG)$_{37}$AA(AAG)AA(AAG)$_3$ |
|                |            | pRW3437 | 38  | (AAG)$_{38}$AA(AAG)$_3$    |
|                |            | pRW3440 | 48  | (AAG)$_{48}$AA(AAG)$_3$AGG(AAG)$_3$ |
|                |            | pRW3438 | 52  | (AAG)$_{52}$AA(AAG)AA(AAG)$_3$AGG(AAG)$_{15}$ |
|                |            | pRW3439 | 58  | (AAG)$_{58}$AA(AAG)$_2$AGG(AAG)$_{19}$ |
|                | (CTT)$_n$  | pRW3434 | 20  | (CTT)$_{20}$               |

* Cloned into pUC18NotI.
Seven Triplet Repeat DNA Sequences

Table II

| Triplet repeat | Name     | n  | Origin     |
|----------------|----------|----|------------|
| (GTA)$_n$      | pRW3453* | 110| pRW3451    |
| (GTA)$_n$      | pRW3482* | 53 | pRW3186    |
| (GTC)$_n$      | pRW3483* | 80 | pRW3482    |
| (GTC)$_n$      | pRW3480 | 115| pRW3428    |
| (TCG)$_n$      | pRW3481 | 140| pRW3428    |
| (AAG)$_n$      | pRW3471  | 103| pRW3439    |
| (AAG)$_n$      | pRW3801* | 103| pRW3439    |

*Cloned into pUC19Ncl1.

**Cloned into pUC18Ncl1.

The lengths observed exceed the number of triplet repeats known to cause neuromuscular diseases (1–7).

Pausing of DNA Polymerase in Triplet Repeat Sequences—

Table III

Summary of pausing of DNA polymerase for the seven triplet repeats

| Triplet repeat | n  | Plasmid | Template |
|----------------|----|---------|----------|
| GA-TAC         | 61 | pRW3157 |          |
| GT-AAC         | 42 | pRW3172 |          |
| GT-ATC         | 62 | pRW3167 |          |
| GT-ATC         | 74 | pRW3174 |          |
| AGG-CCT        | 16 | pRW3191 | +c       |
| TCG-CGA        | 46 | pRW3414 |          |
| AAG-CCT        | 16 | pRW3431 | +c       |

*+a, pausing; —, no pausing.

**The strong pausings were observed at 29–33 triplets from the beginning of the repeats.

***The pausings were observed in the 3'-half of the synthesized strand.

****The pausings were observed throughout the repeats.

The lengths observed exceed the number of triplet repeats known to cause neuromuscular diseases (1–7).

Pausing of DNA Polymerase in Triplet Repeat Sequences—

Primer extension analyses were performed as described under "Materials and Methods." n indicates the total number of triplets. pRW3431, which contains (AAG)$_{16}$, were obtained by cloning synthetic dinucleotides, GCTCT(AAG)$_{16}$, into the HindI site of pUC19.

The lengths observed exceed the number of triplet repeats known to cause neuromuscular diseases (1–7).

Pausing of DNA Polymerase in Triplet Repeat Sequences—

Primer extension analyses were performed as described under "Materials and Methods." n indicates the total number of triplets. pRW3431, which contains (AAG)$_{16}$, were obtained by cloning synthetic dinucleotides, GCTCT(AAG)$_{16}$, into the HindI site of pUC19.
In our previous study (34), we found that the pausing phenomenon seen for the CTG-CAG and CGG-CCG repeats was influenced by temperature; preincubation at 70 °C for 10 min abolished the pausing for CTG-CAG, but not for CGG-CCG, indicating that the heat treatment destroyed a structure (probably H-bonded) that blocked polymerase movement. The structure in the CGG-CCG repeats must be more thermally stable than that in the CTG-CAG repeats. In the case of (TCG-CGA)_{98}, the pausing was abolished by the preincubation at 70 °C for 10 min as found for CTG-CAG (data not shown). Likewise, treatment of (TCG-CGA)_{98} at 60 or 50 °C also abolished the pausing sites. These results indicate that similar thermal structural stabilities of the DNA structures exist for CTG-CAG and TCG-CCG repeats, as expected.

On the other hand, no pausing of this type were seen for the other six triplet repeats (Table III). GTA-TAC, GAT-ATC, GTT-AAC, and CAC-GTG had no pausings at all, whereas AGG-CCT and AAG-CTT showed different types of pausings (data not shown). Extensions of purine strands (AGG and AAG) as the templates were terminated throughout the repeats. In contrast, extensions of pyrimidine tracts (CTT and CTT) showed pausings preferentially in the 3'-half of the extended strands. These terminations of DNA polymerization are likely due to inhibition of DNA synthesis by intramolecular triplex formation with the G-G-C and A-A-T base triads (54, 58-60), since the primer extension analyses were performed at pH 7.5 in the presence of Mg^{2+} (60), and these results are diagnostic for triplexes. Hence, these data indicate that TCG-CGA, the sequence isomer of CTG-CAG, has the same property for the pausing of DNA polymerase as CTG-CAG and CGG-CCG. Thus, out of all 10 triplet repeats, these three repeats may form an unusual DNA structure(s), in agreement with other results, that inhibits DNA polymerization. However, these structure(s) differ from the triplexes formed by AGG-CCT and AAG-CTT repeats.

**DNA Structural Analyses of Triplet Repeat Sequences**—The supercoil stress-induced transitions for non-B DNA structures (such as left-handed Z-DNA, triplexes, cruciforms, and AT-rich bp unpairing) can be monitored by the use of two-dimensional agarose gel electrophoresis (22, 61-68). In addition, chemical probing can identify B-Z junctions, single-stranded, or perturbed regions in other unusual DNA structures (22, 41, 61-67, 69, 70). The combination of the two analyses has provided powerful diagnostic tools for analyzing DNA structures at the bp level. We performed these analyses previously on CTG-CAG (up to 130 repeats) and CGG-CCG (up to 240 repeats) triplet repeats observing neither agarose gel transitions nor chemical modifications (36).

On the other hand, TTA-TAA triplet repeat sequences formed unpaired regions throughout the repeats (37) as detected by these methods. Herein, similar studies are described to extend these DNA structural analyses to the other seven TRS.

Topoisomer populations of plasmids were subjected to two-dimensional gel electrophoresis (Table IV). For GTA-TAC (pRW3155, the total number of triplets (n) is 45), GAT-ATC (pRW3163, n = 44), GTT-AAC (pRW3172, n = 42), CAC-GTG (pRW3185, n = 46), and TCG-CGA (pRW3414, n = 46), no relaxations were observed at pH 8.3 and/or 4.5 (data not shown). The chemical probe analyses for these repeats, GTA-TAC (pRW3155, n = 45), GAT-ATC (pRW3163, n = 44), GTT-AAC (pRW3171, n = 26 and pRW3172, n = 42), CAC-GTG (pRW3181, n = 16 and pRW3185, n = 46), and TCG-CGA (pRW3411, n = 16), showed no reactivities with OsO4 or DEPC at pH 4.5 and/or 8.3 at negative supercoil densities (~n) of ~0.060 as isolated from E. coli (data not shown). These results

**Fig. 3. DNA polymerase pausing sites in TCG-CGA and CTG-CAG repeat sequences in plasmids.** Primer extensions on the bottom strands of pRW3416 containing (TCG-CGA)_{98} and pRW1981 (9) containing (CTG-CAG)_{30}, were performed with primer 1201 as described under "Materials and Methods." The arrows represent the beginning sites of the triplet repeat inserts. G, A, T, and C on the top of the gel designates primer extension analyses with the Klenow fragment of E. coli DNA polymerase I.

previously observed for CGG-CCG (34).

The pausing phenomenon was length-dependent since (TCG-CGA)_{46} did not show any pausings (Table III). This agrees with previous data for CTG-CAG and CGG-CCG in which the repeats required a continuous length of >60 for the pausing to occur (34). In the case of (CTG-CAG)_{30}, pausings were seen at 35-39 triplets, whereas in the case of (TCG-CGA)_{98} pause sites were at 29-33 (Fig. 3). This difference is due to the primer locations. The location of the primer binding sites was found to influence the pausing sites for CTG-CAG repeats (34); as the distance between the initiation site of the CTG repeat and the 5'-end of the primer increased, the pausing site was further from the initiation site. The distance between the pausing site and the first CTG is about 20 bp longer than the distance between the first CTG and the 5'-end of the primer. In this study, the 5'-end of the primer is located 64 bp from the first TCG triplet for (TCG-CGA)_{98} and 89 bp for (CTG-CAG)_{130}.
indicate that the above five triplet repeat sequences do not form supercoil-dependent non-B DNA conformations.

On the other hand, AGG<z>CCT and AAG<z>CTT repeats showed relaxations by two-dimensional gel electrophoresis and modifications by chemical probings under certain conditions. For AGG<z>CCT repeats, pRW3191 (n = 16), at pH 4.5, exhibited a transition even at low supercoil density (−s = 0.010) with a relaxation of about 2.5 supercoil turns, and CAA reactivities in the 5'-half of the (AGG)16 strand at −s = −0.060, whereas at pH 8.3, neither relaxations for pRW3191, pRW3193 (n = 30), and pRW3198 (n = 53) nor chemical modifications for pRW3191 were observed (data not shown).

For AAG<z>CTT repeats, the 5'-half of the (AAG)16 strand of negatively supercoiled pRW3431 (n = 16) (−s = −0.060) was modified by CAA at pH 4.5, but not at pH 8.3, and a supercoil-induced transition was observed only at pH 4.5 at −s = 0.027 with a relaxation of 3.5 supercoil turns (data not shown).

Thus, in both cases of the AGG<z>CCT and AAG<z>CTT repeats, the modifications observed at pH 4.5 show that the pur-pyr tracts formed intramolecular triplexes which consisted of C<z>G<z>C1 and T<z>A<z>T base triads. A number of studies by these methods (58, 59, 61, 62, 69, 70) with other intramolecular triplexes validate this conclusion. It is likely that these triplexes differ from those observed in the primer extension study at pH 7.5 described above.

Longer AAG<z>CTT repeats showed relaxations even at pH 8.3 (Table IV). For pRW3437 (n = 38), the transition occurred at topoisomer −16 (−s = 0.060) with 8.5 supercoil turns relaxed, corresponding to an unpairing or re-pairing into a triplex of 89 bp. At topoisomer −18 (−s = 0.067), an additional two supercoils were relaxed, corresponding to 111 bp, or involvement of the entire repeat region (data not shown). For pRW3439 (n = 58), the transition was initiated at topoisomer −14 (−s = 0.051) with six supercoils relaxed (corresponding to 63 bp), and with the addition of more supercoil density, the relaxation continued until topoisomer −21 (−s = 0.077), a total of 13 supercoil turns corresponding to 137 bp (Fig. 4). For pRW3193, pRW3437, and pRW3439, multiple spots at each linking number were observed when samples were run at pH 4.5. As shown in Fig. 4 for pRW3439, topoisomers between −4 (−s = 0.015) (the beginning of the transition) and −18 (−s = 0.066) (the end of the transition) showed the presence of more than one spot for DNA molecules having the same linking number, indicating that different conformers of the triplexes are forming due to different nucleation sites as seen previously for the (GA-TC)37
tract (62) and for (GAA)$_3$(TTC)(GAA)$_8$ and (GGA)$_n$(TCC)(GGA)$_n$ (61).

Previous studies on shorter AGG-CCT and AAG-CTT repeat sequences (8 repeats or 17 repeats interrupted by 3 bp) revealed that no modifications were observed in negatively supercoiled plasmids at pH > 7.6 (41, 61), whereas, in this study, longer AAG-CTT triplet repeats (38 and 58 repeats) showed relaxations to form triplexes even at pH 8.3. This is due to reducing the dependence on low pH by increasing the length of the pur-pyr tract, as found previously (62). Relaxation was not found for longer AGG-CCT repeats (30 and 53 repeats) at pH 8.3. These differences between the AGG-CCT and AAG-CTT repeats may be due to the different GC content, since the base composition of pur-pyr sequences affects the thermostability and the amount of supercoiling needed for intramolecular triplex formation (70).

In summary, considering the seven TRS studied herein, AGG-CCT and AAG-CTT form intramolecular triplexes, and the other five TRS do not form any previously characterized non-B DNA structures.

**DISCUSSION**

The role of DNA structure and properties in gene expression has been the principal emphasis of this lab (5, 33, 69). On the basis of in vitro and in vivo investigations with simple repeating nucleotide sequences (mono- through tetra- and dodeca-) in DNA oligonucleotides and polymers, restriction fragments, plasmids, and chromosomes, we anticipated that the long tracts of certain TRS (CTG-CAG, CGG-CCG, and AAG-CTT) that elicit human hereditary neuromuscular and neurodegenerative diseases might adopt non-B conformations. A series of physical, biochemical, and genetic investigations (5) reveal several significant factors that provide insights into our understanding of genetic instabilities which elicit the clinical characteristics of anticipation (1-4, 6). We feel that studies on all 10 TRS are important since, first, it is possible that one or more of the other seven TRS (GTA-TAC, GAT-ATC, GTC-AAC, CAC-CTG, AGG-CCT, TCG-CGA, and TTA-TAA) may be identified in the future with a genetic disease gene (e.g. AAG-CTT was very recently described (7) as the cause of Friedreich's ataxia) and, second, they serve as controls for the interpretation of data with the other TRS. This contribution and the accompanying paper on TTA-TAA (37) along with prior work on CTG-CAG (9, 24) and CGG-CCG (34, 36, 42) describe the doning and partial characterization of all 10 TRS.

Our doning strategy was designed to employ synthetic oligonucleotides that would not slip (due to the presence of the restriction site (Fig. 1)) which would enable the cloning of long inserts of pure TRS (no polymorphisms) by repetitive steps (Fig. 2). Since long tracts (hundreds of repeats) have only been found to be associated with the three TRS implicated in the hereditary diseases (Introduction), it was necessary to employ synthetic oligomers for the other TRS. Our doning strategies (Fig. 2) successfully provided all seven TRS with 60–90 repeats. However, some of the plasmids contained polymorphisms, especially for the longer TRS. This behavior in E. coli is similar to the observations in eucaryotes including humans (44, 45) that the presence of a few non-perfect repeats is apparently necessary for genetic stability. Also, in the case of the fragile X and SCA1 genes, it is thought that polymorphisms might cause the stabilization of TRS and that the loss of polymorphisms might cause instability (expansions) (36, 43–45). Interestingly, in the Friedreich's ataxia case (AAG-CTT), the polymorphisms always maintain the homopurines on one strand with the pyrimidines located exclusively on the complementary strand (7); we observed the same behavior in E. coli (Table I).

Another doning strategy with synthetic oligonucleotides containing BbsI and Bsal sites along with TRS was reported (71) for the preparation of tracts of CTG-CAG. The method involves the multimerization by ligation of fragments with TRS to give longer tracts by repetitive cloning steps and was proposed to be useful for the doning of any repeat sequence. We attempted a similar strategy with BbsI as well as SapI, instead of Bsal, to clone long tracts of CGG-CCG and were not successful. However, other methodologies worked well (36).

For AAG-CTT repeats, the strategy was modified from that used for the doning of the other six TRS since no appropriate enzyme site exists for AAG-CTT (Fig. 2B). As a result, AGG-CCT interruptions were necessarily present in the AAG-CTT tracts (Table II). Since AAG-CTT repeats showed a propensity, albeit small, to be expanded in E. coli (8), (AAG-CTT)$_{103}$ (pRW3471) was obtained from (AAG-CTT)$_{58}$ (pRW3439) by in vivo expansion. Investigations on the mechanism(s) of genetic expansion (8, 9, 72) of this sequence are likely to be interesting since (a) it forms a triplex (61, 70) and (b) Friedreich's ataxia (FRDA) patients have 200–900 repeats in the first intron of the frataxin (210-amino acid) gene, whereas normal individuals have 7–20 repeats (7).

We propose that the mechanism of reduction of abundance of mature X25 mRNA in individuals with FRDA (7) is the formation of an intermolecular triplex between the AAG-CTT in the first X25 DNA intron and the RNA segment with the GAA tract (Fig. 5) removed by splicing. Prior work (73, 74) showed that the presence of a triplex inhibits transcription. In the case of long r(AAG) tracts (600–800 repeats) from FRDA cases, the triplex may be sufficiently stable thermodynamically to cause the reduction in abundance of the FRDA mature mRNA, whereas for shorter r(AAG) stretches from normal individuals (6–20 repeats), the triplex may be unstable and will not cause an inhibition. Possible pairing schemes (reverse Hoogsteen) are shown for the T-A-A and C-G-G triads.

Our E. coli in vivo expansion system (8, 9, 72) has been valuable for the preparation of long TRS. Prior work showed that CTG-CAG repeats could be expanded about nine times more frequently than any of the other nine TRS; TCG-CGA, CGG-CCG, and CAC-CTG were also prone to be expanded, but less frequently; of the other TRS, only AAG-CTT was found, but at a very low level (8). In the present study, besides those TRS, GTA-TAC expansion was observed in an individual experiment. For GAT-ATC, small expansions (~six repeats) occurred. Since it is known that any TRS might be sufficiently unstable to give slippage(s) depending on the sequence (75), we propose that expansions occurred during DNA replication. In the present and other studies (8, 72) on in vivo expansion, we focused on expanded products larger than 20 repeats rather than on small expansions as seen for GAT-ATC in this study. Since the expansion in E. coli of CTG-CAG repeats is distal to the replication origin as a single large event (72), and since the expansion occurs more frequently with increasing repeat numbers (9), this strategy is highly desirable for the doning of longer expansion products. Once the appropriate repeat lengths are cloned from synthetic oligomers, in vivo expansion could be applied. An advantage of the in vivo expansion approach is that fewer polymorphisms are observed when than synthetic oligonucleotides are used and also that the expansion products will be triplexes except for the cases of mono- or dinucleotide interruptions in the TRS. Thus, the combination of the doning strategies with synthetic oligonucleotides and in vivo expansion was useful for the doning of long tracts of TRS.

CTG-CAG and CGG-CCG repeats showed length-dependent

---

5 R. Gellibolian and R. D. Wells, unpublished data.
strong pausings by DNA polymerases within the repeats (34). Herein, we showed that TCG\_CGA, the sequence isomer of CTG\_CAG, also showed similar pausing (Fig. 3). The pausing was destroyed by heat treatment (70 °C). AAG\_CTT and AGG\_CCT had different types of pausings from the cases seen for the previous three TRS, indicating that the pausings were derived from triplex structures. No pausings were observed for the other TRS. We previously proposed that the pausings were due to the existence of a non-B DNA structure(s) that blocks DNA polymerase progression; the resultant idling polymerase may catalyze slippages to give expanded sequences. Although CAC\_GTG, AAG\_CTT, and GTA\_TAC repeats were expanded in E. coli, pausings, as seen for CTG\_CAG, CGG\_CCG, and TCG\_CGA, were not found for these inserts except for AAG\_CTT. The locations of the pause sites were related to the location of the primer (34). Recent studies\(^6\) on the DNA sequences of the paused newly synthesized DNA products revealed that template switching had occurred. We envision that the snap-back structure was caused by the polymerase encountering a non-B DNA structure in the TRS.

Simple repeating DNA sequences adopt non-B DNA conformations (reviewed in Refs. 5, 21, 22, 33). For TRS, long tracts of CTG\_CAG, CGG\_CCG, and TTA\_TAA have been shown to form non-B DNA structures. Our present studies with chemical probes and two-dimensional gel electrophoretic analyses showed fundamental DNA structural features in vitro for TRS. Neither specific reactivities nor relaxations were observed except for AGG\_CTT and AAG\_CTT, which formed triplex structures. At acidic pH, it has been shown that pur\_pyr tracts, including short tracts of AAG\_CTT and AGG\_CCT (41, 61, 70), adopt triplexes. In this study, the longer AAG\_CTT tracts (n = 38 and 58) showed relaxations even at pH 8.3, presumably due to the formation of triplexes. Prior studies (62) showed that low pH is not required for intramolecular triplex formation for certain sequences of the appropriate length. Thus, long AAG\_CTT repeats (~800 repeats) in the Friedreich's ataxia patients (7) might be expected to form a non-B DNA conformation(s) in vivo.

Acknowledgments—We thank Drs. J. Klysik and A. Jaworski for helpful discussions and M. Pandolfo for a preprint of his paper prior to publication.

REFERENCES

1. Davies, K. E., and Warren, S. T. (eds) (1993) Genome Analysis, Vol. 7, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

\(^6\) K. Ohshima and R. D. Wells, manuscript in preparation.
