Cloning, Characterization and Properties of Plasmids Containing CGG Triplet Repeats from the FMR-1 Gene

Miho Shimizu¹, Robert Gellibolian¹, Ben A. Oostra² and Robert D. Wells¹*

¹Institute of Biosciences and Technology, Texas A & M University, Department of Biochemistry and Biophysics Texas Medical Center, 2121 W. Holcombe Houston TX 77030-3303 USA
²Department of Clinical Genetics, Erasmus University Rotterdam, P.O. Box 1738 3000 DR Rotterdam The Netherlands

The FMR-1 gene for the human fragile-X syndrome, a mental retardation disease inherited by non-Mendelian transmission, contains a genetically unstable CGG region in the 5’ non-translated region. The severity of the disease is correlated with the length of the CGG tract. The cloning of 28 stable plasmids containing (CGG), inserts (where n = 6 to 240) with different extents and types of sequence interruptions (polymorphisms), and in different orientations was accomplished by three strategies in Escherichia coli. Some shorter tracts were prepared by the direct cloning of synthetic oligonucleotides, and longer runs were clones of multimers of (CGG),, (CGG),AGG(CGG),CAG(CGG), from a cDNA from a fragile-X patient or from expansions or deletions of these sequences in E. coli. The genetic stability of the inserts, especially for the longer tracts, was dependent on the sequence length, the presence of polymorphisms, the host cell genotypes, the orientation of the inserts in the vector and the position of cloning in a vector. Two-dimensional agarose gel electrophoresis studies on fully methylated and on non-methylated plasmids as well as chemical probe studies revealed the absence of underwound structures or accessible base-pairs. These DNAs enable a range of genetic and biochemical investigations into the molecular basis of the fragile-X syndrome.

© 1996 Academic Press Limited

Keywords: triplet repeat; fragile-X syndrome; CGG; expansion; DNA structure

Introduction

Several human hereditary neuromuscular and neurodegenerative diseases have been linked to variations in the length of CTG and CGG triplet repeats (TR) within their genes (reviewed by articles in Davies & Warren, 1993). The fragile-X syndrome, an inherited mental retardation disease with several physical abnormalities, affects approximately one in 2000 males and is X-linked and carried by normal carrier (or transmitting) males whose daughters can have affected offspring (reviewed by Nelson, 1993). The penetrance of the disease increases with each generation, suggesting a non-Mendelian mode of transmission (anticipation). The genetically unstable region contains multiple CGG repeats (Kremer et al., 1991) within the mRNA of a fragile-X-associated gene (FMR-1; Verkerk et al., 1991; Fu et al., 1991). Normal individuals show heterogeneity at this site containing between six and 60 copies of the CGG triplet with an average of 29 repeats. The normal repeat contains the sequence (CGG),AGG(CGG),AGG-(CGG), (Verkerk et al., 1991). Carrier (asymptomatic) males have a premutation between 60 and 200 copies of the repeat. Individuals with a full mutation, greater than 200 copies of the repeat, typically show symptoms of the fragile-X syndrome. Expansion may occur during embryonic development, between five and 20 days post conception (Reyniers et al., 1993; Wohrle et al., 1993).
The TR is located within the non-translated region of the FMR-1 gene. In affected males with greater than 200 copies of the repeat, there is no expression of the FMR-1 gene, which may be the result of methylation of a CpG island about 200 bp 5′ to the CGG repeat (Yu et al., 1992; Sutcliffe et al., 1992; Hornstra et al., 1993). Methylation of this FMR-1 CpG island has been observed in fragile-X patients (Bell et al., 1991). The CGG repeat is completely methylated when expanded and the TR loses at least one AGG triplet, forming a longer perfect CGG repeat region. The FMR-1 gene is an RNA binding protein (Ashley et al., 1993; Sioini et al., 1993), although its biological role remains to be established. Expansion of CGG TR has been associated also with the FRA11B locus (Knight et al., 1993; Jones et al., 1995).

The right-handed B conformation of DNA in the Watson–Crick double helix (10.4 bp/right-handed turn) is believed to be the principle structure in living cells. However, under appropriate conditions (sequence, negative supercoiling density, ionic conditions, etc.), several types of non-B conformations have been well characterized. These structures exist in short segments at defined loci. For example, cruciform structures form in negatively supercoiled DNA within regions of inverted repeat symmetry (Sinden, 1994; Sinden & Wells, 1992). Z-DNA is formed at regions of alternating purine and pyrimidine sequences, such as (GT)<sub>n</sub> or (GC)<sub>n</sub>, under physiological conditions in negatively supercoiled DNA. Also, alternating AG sequences and other polypurine-polypyrimidine regions that contain mirror repeat symmetry form various intramolecular triple-helical isomers in which one half of the tract unpairs and one strand hydrogen bonds utilizing Hoogsteen base-pairing in the major groove of the DNA duplex. Other alternative conformations are known to exist with simple repeating sequences including nodule DNA, parallel DNA, tetrastranded telomere DNA structures, bent DNA, and slipped structures (Sinden, 1994; Sinden & Wells, 1992; Wells, 1996). All of these alternative conformations were well characterized in vitro and several were shown to exist readily in living bacterial cells and in eucaryotes (Sinden & Wells, 1992; Lukomska & Wells, 1995).

Whereas these structural studies were conducted on a variety of simple sequences, few investigations have been conducted on TR sequences cloned into plasmids. Inserts containing long CGG repeats have been especially difficult to clone since they are prone to form deletion products (Gastier et al., 1995). Here we describe the cloning and characterization of a variety of lengths of CGG repeats containing different types of polymorphisms which may be used for future investigations including DNA structure, mismatch repair, mechanisms of replication and recombination related to expansion and deletion, and gene expression in prokaryotic and transgenic mouse systems.

Results

Cloning and characterization of plasmids containing pure CGG inserts

A family of plasmids containing pure (CGG)<sub>n</sub> repeats, where <i>n</i> ranged from 6 to 32, was prepared by a combination of standard cloning methodologies with synthetic oligonucleotides along with the procedures (Kang et al., 1995a) for generating expansions and deletions in vivo. Figure 1(a) outlines the method used for generating pRW3024 which contains (CGG)<sub>24</sub>. The oligonucleotides were designed to produce a recombinant plasmid with CGG being on the leading strand with respect to the origin of replication of the pUC19-NotI.

pRW3024 was isolated and characterized by double digestion with SacI and HindIII by PAGE (see Materials and Methods). The analysis (Figure 1(b)) reveals intense bands (arrowheads) in lanes 2 and 3 corresponding to the expected full-length restriction fragments containing ten and 24 CGG repeats, respectively. Two interesting features are apparent. Firstly, the longer sequence, pRW3024, is more unstable (more deleted) in <i>E. coli</i> DH5α compared to the shorter pRW3010; furthermore, the deletion products observed for pRW3024 are not a smear but rather resolve as distinct bands on the gel each differing from its neighbor by one CGG triplet. Secondly, in addition to deletions, expansion products are also observed which differ in length by one repeating unit. On the other hand, pRW3010 does not show the instability, i.e. expansion/deletion, observed for pRW3024, indicating that length is an important factor in the stability of the CGG TR.

To test whether the expanded/deleted bands are due to instability occurring within the TR, we eluted DNA from regions of the gel corresponding to the expansions and deletions and cloned the fragments into the SacI/HindIII region of the polylinker of pUC19-NotI. Shorter as well as longer clones containing six (pRW3006), eight (pRW3008), ten (pRW3010), 17 (pRW3017), 24 (pRW3024), 29 (pRW3029), 32 (pRW3032), 34 (pRW3675), 40 (pRW3681), 43 (pRW3679), 46 (pRW3677) and 49 (pRW3683) CGG repeats were characterized by sequencing. This is the first observation of expansion/deletion products occurring in integral multiples of the triplet repeats. Hence, the mechanism for this process must be complementary strand slippage as proposed previously (reviewed by Wells & Sinden, 1993).

Stability of (CGG)<sub>24</sub> in different host cells

The stability of a pure tract of (CGG)<sub>24</sub> in different host cells posed a challenge for obtaining a homogeneous insert preparation. The propagation of CGG TR sequences in <i>E. coli</i> DH5α provided an effective venue to create expansion as well as deletion products, but did not allow for the
stabilization needed to generate a homogeneous length of repeat. However the best host cells for the stabilization of pure \((\text{CGG})_{24}\) repeats is \(E.\ coli\) SURE (defining stability as the amount of full length insert as determined by PAGE). After 12 hours of growth (~20 generations) in \(E.\ coli\) SURE the stability was 100%, and in \(E.\ coli\) DH5\(\alpha\), 65 to 70%. Hence, DH5\(\alpha\) is optimum for creating different lengths of derivative inserts whereas SURE is the preferred cell line for insert stabilization.

### Cloning strategy for long CGG repeats in plasmids

The direct cloning of cDNA sequences from human patients has served as a source of investigative tools (Verkerk et al., 1991) but the majority of the triplet tracts are <30 repeats. Two other strategies may be exploited, the direct cloning of a synthetic DNA fragment (described above) and the \textit{in vivo} expansion method in \(E.\ coli\) (Kang et al., 1995a). For structural studies, we wished to clone more than 200 CGG repeats. The synthesis and purification of such long tracts of CGG-CCG by chemical methods is intrinsically difficult. Application of the \(E.\ coli\) expansion methodology (Kang et al., 1995a,b) to CGG repeat sequences from the fragile-X gene failed because of either less-frequent expansion of this particular TR in \(E.\ coli\) (Ohshima et al., 1996) or the difficulty in obviating deletions of the long tracts of non-interrupted CGG repeats, or both. In general, GC-rich sequences from microsatellites (Klysik et al., 1982), including CGG repeats (Gastier et al., 1995), are known to be difficult to clone. The instability of CGG repeat sequences in the human fragile-X syndrome are caused by the loss of interruption (polymorphic) sequences (AGG; Eichler et al., 1994, 1995). We observed a similar instability of cloned non-interrupted CGG sequences in \(E.\ coli\), especially for tracts longer than 30 triplets.

### Preparation of plasmids containing long repeats

In fragile-X patients, there is a correlation between the repeat length and the severity of the symptoms. For complete inactivation of the FMR-1 gene, it is known that the multiples of CGG should exceed more than 200. Thus, we wished to clone long CGG repeats (>200) in a vector for structural studies which may lead to a better understanding of the molecular basis for the genetic instability of the repeats.

The DNA fragment containing a CGG repeat was isolated from RN2, a derivative of pT7Blue(R)
with a cDNA sequence from a fragile-X patient containing (CGG)$_{11}$, ((CGG)$_{11}$AGG(CGG)$_{60}$CAG-(CGG)$_{8}$), in the EcoRV site of the polylinker with only a small amount of flanking sequence (5 bp on the HaeIII side and no DNA on the other end; Figure 2). The resulting blunt-ended DNA fragment digested by HaeIII and BstUI was multimerized by T4 DNA ligase. Successful cloning was achieved by isolating the dimer from a 5% (w/v) polyacrylamide gel and ligating it into dephosphorylated HinclII linearized pUC19 followed by transformation into E. coli HB101 or SURE by electrotransformation. Restriction analysis of DNA from a single clone out of the nine analyzed revealed the presence of the dimer molecule of (CGG)$_{81}$ as the major component (pRW3306) along with several smaller sized (deleted) CGG repeat sequences, including pRW3338 and pRW3339. Although E. coli SURE was the optimum strain used for stabilizing pRW3306, a large number of cell divisions caused the eventual instability (deletion) of the plasmid. Hence, the CGG-containing insert from pRW3306 was transferred into pBR322 (described below).

Recloning (CGG)$_{160}$ into a pBR322 derivative

Prior studies (Jaworski et al., 1989) showed that sequences that were inherently unstable genetically due to their simple repeating tracts that could form non-B structures were more stable when cloned into the promoter region of the tet gene of pBR322. Hence, we recloned the dimer insert from pRW3306 into pRW1560 to assess the stability of the insert in this vector. pRW1560 is a pBR322 derivative with a BgII site, which is compatible with BamHI ends, at the filled-in unique EcoRI site. pRW3306 was digested with BamHI and HindIII at the polylinker site of the vector. The isolated fragment then was cloned into BgII/HindIII digested pRW1560 (dephosphorylated) in HB101. Agarose gel analyses of isolated plasmids showed that ten out of 16 clones contained a homogeneous (CGG)$_{160}$ sequence. Thus, these data show that another unstable DNA sequence, a long CGG repeat, was stably cloned at an insertion-mutated tetracycline gene promoter of pBR322 as in the previous cases for Z-DNA and intramolecular triplexes (Jaworski et al., 1989).

Cloning of a (CGG)$_{240}$ insert

Since we found that a long CGG repeat can be stably cloned into pBR322 as described above, we attempted to clone the (CGG)$_{240}$ insert. A DNA fragment containing a trimer of (CGG)$_{81}$ as the major component (pRW3306) along with several smaller sized (deleted) CGG repeat sequences, containing pRW3338 and pRW3339. Although E. coli SURE was the optimum strain used for stabilizing pRW3306, a large number of cell divisions caused the eventual instability (deletion) of the plasmid. Hence, the CGG-containing insert from pRW3306 was transferred into pBR322 (described below).

Recloning (CGG)$_{160}$ into a pBR322 derivative

Prior studies (Jaworski et al., 1989) showed that sequences that were inherently unstable genetically due to their simple repeating tracts that could form non-B structures were more stable when cloned into the promoter region of the tet gene of pBR322.
CGG Triplet Repeats from the FMR-1 Gene

Figure 3. (a) Analysis of pRW3309 clones in E. coli SURE cells. Five transformants were grown in LB broth containing 70 µg/ml ampicillin at 37°C for ten hours. Lanes 1 to 5, DNA was extracted, linearized with SacI, and analyzed on a 1.5% agarose gel; lane 6, SacI-digested pRW3306; lane M, size marker DNA. (b) Stability of CGG-containing pBR322 derivatives in E. coli HB101.

Densitometric analysis of the gel showed that the expected-sized inserts were present in the following proportions; 100% for pRW3316 (lane 1), 77% for pRW3308 (lane 2) and 53% for pRW3310 (lane 3). On the other hand, these inserts are relatively unstable in pUC19; 100% for pRW3311, 46% for pRW3306 and 27% for pRW3309.

In summary, the stability of the long (CGG) inserts was greatly enhanced by cloning into the promoter region of the tetracycline-resistance gene in pBR322 via subcloning through pUC19. Thus, repeats up to 240 units were stabilized in E. coli. Hence, it may be possible to clone even longer CGG repeat sequences (more than 240 repeats) into this region from a DNA fragment derived from human disease genes.

Stability of inserts is determined by orientation and the presence of interruptions

The frequency of expansion and deletion of CTG triplets is influenced by the direction of replication in E. coli (Kang et al., 1995a). The vectors pUC19 and pBR322, which were used for cloning in this study, have a unidirectional replication origin (ColE1). In the construction of pRW3311, pRW3306 and pRW3309, the ligation of CGG inserts into a HindIII site of pUC19 could occur in either orientation because of the blunt ends. Interestingly, all inserts were found in only one orientation, namely, the CGG triplets were always in the leading strand for replication (orientation I) as revealed by dideoxy sequencing of the DNAs. For example, when pRW3311 was constructed, ten clones were analyzed for the orientation of the (CGG)81 insert in pUC19; all ten inserts were in orientation I (data not shown). Hence, the stability of the (CGG)n repeats (where n = 81, 160 and 240) is orientation-dependent in E. coli as found for CTG (Kang et al., 1995a). Further studies described below confirm this conclusion.

To study the effect of cell division on the instability of CGG repeats, we constructed plasmids in which (CGG)160 was inserted in both orientations in a modified pBR322 vector. E. coli HB101 was used for transformation. The desired constructs were obtained by a two-step method as described in Materials and Methods. A high level of instability was observed for the plasmid (pRW3348) in which (CGG)160 was inserted in orientation II (Figure 4); almost no fragments of the expected size were when analyzed on an agarose gel after digestion with HindIII and XbaI. pRW3310 was transformed again into HB101 to get a stable clone.

Since the inherent stabilization of larger CGG tracts in a pBR322 derivative (pRW1560) was established, we recloned the insert containing (CGG)81 into the same region of the vector as for pRW3310 and 3306. Figure 3(b) shows the stability of pRW3316, 3308 and 3310 which contain 81, 160 and 240 repeats of CGG, respectively, in pRW1560. At least twofold higher stability was obtained by recloning these CGG inserts into pBR322. Densitometric analysis of the gel showed that the expected-sized inserts were present in the following proportions; 100% for pRW3316 (lane 1), 77% for pRW3308 (lane 2) and 53% for pRW3310 (lane 3). On the other hand, these inserts are relatively unstable in pUC19; 100% for pRW3311, 46% for pRW3306 and 27% for pRW3309.

In summary, the stability of the long (CGG) inserts was greatly enhanced by cloning into the promoter region of the tetracycline-resistance gene in pBR322 via subcloning through pUC19. Thus, repeats up to 240 units were stabilized in E. coli. Hence, it may be possible to clone even longer CGG repeat sequences (more than 240 repeats) into this region from a DNA fragment derived from human disease genes.
found (lanes 8 to 12). Alternatively, when \((\text{CGG})_{160}\) was in orientation I (pRW3347), the insert was stably maintained (lanes 2 to 6). Further analyses of the \text{XbaI}/\text{HindIII} digests of these plasmids on 5% PAGE showed the presence of homogeneous \((\text{CGG})_{160}\) in pRW3347. Since these plasmids were prepared from a ten hour culture of a single colony on ampicillin plates (no more than 20 generations of \textit{E. coli}), the orientation of the insert is crucial for maintaining the \text{CGG} triplet repeats in a plasmid vector.

As in the case of pRW3348 and 3347, we constructed a plasmid where the \((\text{CGG})_{36}\) insert was cloned in pUC-derived plasmids in both orientations I and II giving pRW3012 and pRW3021, respectively. The recombinant plasmids were then grown in the \textit{E. coli} \text{SURE} strain for \(\sim 20\) generations at \(37^\circ\text{C}\). After isolation of the DNA and double digestion with \text{SacI} and \text{HindIII}, the products were analyzed on a 1.5% agarose gel and the bands quantitated. For pRW3012 (orientation I), the insert was 100% stable whereas in the case of pRW3021 (orientation II), only 43% of the full-length insert remained and the deletion products ranged from \(\sim 10\) to 70 \text{CGG} repeats.

The presence of non-\text{CGG} triplets (interruptions) in the tandem \text{CGG} repeat sequences also influenced the stability. As described above, all plasmids harboring non-interrupted \text{CGGs} were cloned into pUC18-\text{NotI} in orientation II. To test if the instability of pure \text{CGG} repeats is also orientation dependent, we recloned the insert from pRW3032 into pUC18-\text{NotI} to change the direction of the insert to orientation I. Several deletion products were present in the clone in which a pUC18-\text{NotI} derivative contained non-interrupted \text{CGGs} in orientation I. Alternatively, 35 repeats containing one \text{CAG} interruption and 44 repeats with two interruptions were stably cloned in orientation I, as expected (data not shown).

**Two-dimensional gel electrophoresis**

Certain types of simple repeat sequences in plasmids adopt non-\(B\) conformations (left-handed Z-DNA, cruciforms, triplexes, etc.) under the influence of negative supercoiling, and the transitions have been investigated by two-dimensional (2-D) agarose gel electrophoresis (Frank-Kamenetski & Mirkin, 1995; Panutin & Wells, 1992). These structures are stabilized by an underwinding of the primary duplex.

pRW3308, which contains 160 copies of \text{CGG}, was analyzed using 2-D gels for both non-methylated and methylated plasmids. Topoisomers were electrophoresed through 1% (w/v) agarose in TBE (pH 8.3; Figure 5). Slightly faster-moving spots which are observed in pRW3308 indicate topoisomers in which some \text{CGG} repeats were deleted, and are represented as small dots in the lower panel. No insert-specific transition was observed for either unmethylated (left panel) or methylated (right panel) plasmids compared to the vector DNA. Analysis of pRW3310, containing 240 repeats, gave similar results (data not shown).

In addition, plasmids with shorter inserts, 30, 74 and 81 repeats (including interruptions), were investigated by 2-D gel methods under various conditions to explore the potential capacity of the triplet tracts to adopt a non-B structure to obtain higher resolution. Several metal ions including 2 mM MgCl\(_2\), 400 \(\mu\text{M} \text{ZnCl}_2\), 40 \(\text{mM KCl}\), and 150 \(\mu\text{M} \text{Co(NH}_3)_6\text{Cl}_3\) were investigated. Acidic pH (4.5) was tested also to determine the effect of protonation of cytosine bases in \text{CGGs}. None of the gel patterns showed the presence of insert-specific transitions (data not shown).

**Discussion**

A series of inserts containing six to 240 copies of \text{CGG} were stably cloned in plasmids. Several factors influence the stability (deletions and expansions) of the inserts; repeat length, the presence of interruptions, the orientation of the insert relative to the unidirectional replication origin, \textit{E. coli} host strains, the location of the insert and the copy number of the host vector. The stability varies strongly with the length of the insert; longer tracts of \text{CGG} repeats show a greater degree of instability compared to shorter inserts. This behavior in \textit{E. coli} is consistent with the mechanism of genetic anticipation for the fragile-X syndrome (Sherman \textit{et al.}, 1985; Fu \textit{et al.}, 1991). Furthermore, the effect of length on DNA polymerase pausing was also observed during synthesis of the repeat \textit{in vitro} when the Klenow
fragment of DNA polymerase I was used; lengths of greater than 61 repeats showed stronger pausing sites, occurring at repeat number 30 (away from the CGG start site), when CGG was the template strand. This phenomenon was also observed with CTG triplet repeats (Kang et al., 1995b). These results suggest that, at a critical length, the CGG sequence adopts a non-B conformation(s) which blocks DNA polymerase progression, leading to the idling and subsequent slippage to give expanded products and hence provide the molecular basis for this non-Mendelian genetic process.

The canonical human FMR-1 repeat carries 30 CGG triplets interrupted by two AGG triplets at the tenth and 20th repeat. Fragile-X carriers carry longer repeats (50 to 200) that contain long stretches of uninterrupted CGG triplets which predispose this sequence to hyperexpansion in successive generations. Affected individuals have longer, methylated repeats (230 to 2000; Warren & Nelson, 1994). Our results indicate that the presence of interruptions greatly enhances the stability of the CGG tract in E. coli. Other studies on alleles derived from human patients show the presence of stable

![Figure 5. Two-dimensional agarose gel electrophoretic analysis of both unmethylated and methylated pRW3308 and vector (pBR322). Upper panels, gel photographs; lower panels, schematic representations of the gel photographs. Both the vector and the test DNAs were run simultaneously by loading the samples at different positions on the gel. The locations of nicked (N) and linear (L) DNAs are indicated. Open circles represent non-deleted topoisomers, whereas dotted regions for pRW3308 indicate topoisomers in which some CGG repeats have been deleted (see the text).](image)
and unstable CGG triplets of similar size, suggesting that a feature other than length, but intrinsic to the repeat, was responsible for stability. This supports the observations of Eichler et al. (1994) who found that lengths of >33 uninterrupted CGG triplets showed marked instability, regardless of total repeat length, suggesting that the loss of the AGG interruptions is an important mutational event in the generation of alleles predisposed to the fragile-X syndrome.

As mentioned above, another important factor dictating stability is the orientation of the CGG-containing insert. Our results indicate that the triplet repeat was stably maintained in vectors if the CGG strand is in the leading template strand (orientation I) with respect to the origin of replication. However, if CGG falls in the leading template strand (orientation II), the insert was highly destabilized (depending on the length), undergoing deletions and expansions. As in the case of CTG-repeating sequences, the frequency of expansion and deletion of the CGG triplet repeats is influenced by the direction of replication (Kang et al., 1995a) which involves an asymmetric DNA polymerase complex that simultaneously replicates the leading as well as the lagging strands (Wells & Sinden, 1993). Replication-dependent deletion between direct repeats occurs preferentially in the lagging strand due to the unequal probability of forming hairpins (Trinh & Sinden, 1991). Therefore, the deletion of the insert (in orientation II), can be explained by the propensity of the CGG template strand to form a stable hairpin (Mitas et al., 1995; Chen et al., 1995; Gacy et al., 1995; Mitchell et al., 1995; Fry & Loeb, 1994) which is bypassed by the replication machinery during resynthesis of the DNA. On the other hand, expansions within the tract are likely due to strand realignment through slippage of the complementary strands during pausing (described above) to generate a folded and elongated nascent DNA on the leading strand (Kang et al., 1995a). Deletions were the most abundant species detected, but expansions were also visible when pRW3024 (i.e. (CGG)24 in orientation II) was propagated in E. coli DH5α; the bands differed from each other by one repeating CGG unit suggesting the involvement of slipped structures during replication. This method allowed the cloning of the expanded and deleted products (six to 49 repeats) in orientation I and their propagation in E. coli SURE to give a stable DNA preparation. Studies conducted on the other nine triplet repeats revealed that the CTG sequence has the highest propensity to expand (Ohshima et al., 1996).

Among the strains which were tested (HB101, STB12, and RS2), E. coli SURE was the best choice for stably maintaining the CGG triplet repeats of up to 160 repeats, in pUC-derived plasmids. Inserts, containing more than 160 repeats were extremely unstable in pUC19 and were prone to delete to smaller-sized plasmids. The key for successful cloning for long CGG tracts (>160 repeats) was the two-step method described in the Results. High copy number plasmids offer many advantages as vectors for the cloning and expression of foreign genes but the maintenance of the unstable insert was difficult and deletions were common as described here. Interestingly, in the shuttle vector pGS100 (Sargent et al., 1995) a 12 kb, high copy number, pUC-derived plasmid, (CGG)240 was cloned ~1.2 kb away from the origin of replication and the insert was maintained with a higher stability (65 to 75%) than when cloned into pUC19, where stability dropped to 27% (data not shown). This result indicates that there are other factors, such as the location of the insert as well as the size of the vector, that influence the stability of long CGG inserts.

Instability of CGG repeats in E. coli is clearly sequence specific, and probably involves two different non-B DNA structures; one is the single-stranded DNA hairpins formed during replication which may be involved in the deletion and expansion events. The other structure is the toroidal conformation present in duplex DNA and is supported by several structural studies including chemical probe analyses (Gellibolian et al., unpublished results). This double-stranded conformation is likely the cause of DNA polymerase pausing within the CGG and CTG repeat tracts (Kang et al., 1995b) and the basis of preferential nucleosome positioning at long CTG repeats (Wang et al., 1994).

Our 2-D gel study revealed no supercoil-induced transitions with the methylated and non-methylated plasmids. Methylation plays an important role in the complete inactivation of the FMR-1 gene in fragile-X patients (Bell et al., 1991; Hornstra et al., 1993). Despite these profound biological differences between the methylated and non-methylated CGG tracts, no structural changes were observed by 2-D gel analyses and by other structural studies such as PAGE (Gellibolian et al., unpublished results). Hence, we conclude that the role of methylation is not at the DNA structural level but must be related to the interactions with the DNA-metabolizing apparatus in vivo.

In summary, the preparation and characterization of a family of plasmids containing (CGG)ₙ, where n = six to 240, enables a wide range of investigations on the etiology of fragile-X. Despite many studies, few data are available about the mechanism of repeat expansion. The study of repeat (in)stability in vitro or in bacteria is an important tool in understanding the underlying mechanism of repeat instability in humans. Also, cloning in bacteria is the first step towards the introduction of larger repeats in an animal model. In humans, the instability from a normal allele to a disease-causing allele takes many generations (Chakravarti, 1992) and the final step can only take place from a mother to her child. Understanding this mechanism might provide tools for intervention in this step, preventing a premutation from growing to a full mutation which results in disease.
Materials and Methods

Enzymes, reagents and strains

T4 DNA ligase, polynucleotide kinase and Sequenase Version 2.0 were purchased from United States Biochemicals. All restriction endonucleases and SsII methylase were from New England Biolabs, Inc. Chicken erythrocyte topoisomerase was prepared in this laboratory (gift of J. E. Larson).

Deoxynucleoside 5'-triphosphates were purchased from Boehringer-Mannheim and United States Biochemicals. Chloroquine diphosphate, ethidium bromide and ampicillin sodium salt were from Sigma. Gel electrophoresis was done using 15% DNA agarose from FMC BioProduct and Long Ranger from J. T. Baker for sequencing. The 1 kb and 123 bp DNA ladders were purchased from GIBCO BRL. X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside), IPTG (isopropyl-β-D-thiogalactopyranoside), acrylamide and N,N'-methylene-bisacrylamide were purchased from United States Biochemicals. [γ-32P]ATP was from Amersham.

Oligonucleotides were synthesized at the Institute of Biosciences and Technology.

DH5α (Gibco BRL): φ80lac ZAM15 Δ(lac-32Y0-argF)U169 deoR recA1 endA1 phoA hisdR17(ks, mi) supE44 λ- thi-1 gyrA96 relA1 SURE (Stratagene); e14 (McrA-) Δ (mcrCB-hsdSMK-rrI)171 endA1 supE44 thi-1 gyrA96 relA1 lac recB resc bscC umuC::Tn5 (Kanr) supC (F′ pro AB lacIq ZA M15 Tn10 (Tetr)). HB101 (Gibco BRL): mcrB mrr hisdS20 (r6, mi) recA13 supE44 ara14 galK2 lacY1 proA2 rpsL20 (Strr) xyl5 λ- leu2 m1 1 STBL2 (Gibco BRL): mcrA Δ (mcrBC-hsd Rsm-rrI) recA1 endA1 gyrA96 thi supE44 relA1 λ- Δ (lac-pro AB).

Plasmid construction; construction of plasmids with stretches of pure CGG (method 1)

pUC19-NotI (Kang et al., 1995a) was used for cloning homogeneous stretches of CGG triplet repeats. A 5′ng sample of this plasmid was linearized by digestion with BamHI and the DNA was purified as described above. Two duplex oligonucleotides were cloned into this plasmid, giving a total of 24 CGG units. A 2 μg (each) sample of oligo 1 (5′ GATCCGCTCTCTTGCCCC 3′) and oligo 2 (3′ GGAGAGACGGCCGCGG 5′) were mixed in one tube in the presence of 25 mM NaCl and 3 mM MgCl2 in a total volume of 20 μl. The procedure was repeated for oligo 3 (5′ GGGCCGCTGCTGGGGCCG 3′) and oligo 4 (3′ GGAGAGACGGCGCCGG 5′) at the 8th repeat (CAG) as shown in Table 1. The inserts were ligated to generate multimers using T4 DNA ligase. Longer CGG repeats, (CGG)10 and (CGG)10, are head-to-tail dimers or trimers, respectively, of (CGG)6 and were also subcloned into pUC19. pRW3306 and pRW3309 also contain non-CGG repeat sequences (CTGGG) at the junction of the two blocks of (CGG)6.

Plasmid construction; construction of plasmids with long CGG repeats (method 2)

pRW3311, pRW3306 and pRW3309 are pUC19-based plasmids and were constructed as follows. The HaeIII-BstUI fragment of RN2, which contains (CGG)6, was isolated from a polyacrylamide gel by electroelution and directly subcloned into the HinDIII site in pUC19. The CGG repeat sequences in RN2, which were derived from the cDNA of fragile-X patients (Verkerk et al., 1991), contain mutations of the perfect repeat at the 12th repeat (AGG) and at the 73rd repeat (CAG) as shown in Table 1. The insert was ligated to generate multimers using T4 DNA ligase. Longer CGG repeats, (CGG)10 and (CGG)10, are head-to-tail dimers or trimers, respectively, of (CGG)6 and were also subcloned into pUC19. pRW3306 and pRW3309 also contain non-CGG repeat sequences (CTGGG) at the junction of the two blocks of (CGG)6.

For the study of the orientation-dependent stability of CGG inserts in pBR322, the plasmids which contain (CGG)6 in both orientations were constructed. A 322 bp PvuII fragment from pUC19 containing the polylinker was inserted into the filled-in EcoRI and HindIII sites of pBR322 in both orientations to produce pRW3342 (orientation I) and pRW3343 (orientation II); then, the XhoI and HindIII fragment from pRW3306 containing ampicillin and DNA was isolated by the alkali lysis method (Maniatis et al., 1982). The DNA was characterized by double digestion with SacI and HindIII and the digested products were analyzed on an 8% polyacrylamide gel in TBE buffer (89 mM Tris-borate (pH 8.3), 2 mM EDTA). Clones containing the appropriate-sized insert were selected for sequencing.
Table 1. Cloned CGG repeat sequences

<table>
<thead>
<tr>
<th>Plasmids</th>
<th>Total number of repeats</th>
<th>Vector</th>
<th>Composition of repeat</th>
<th>Cloning methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>pRW3006</td>
<td>6</td>
<td>A</td>
<td>(CCG)$_k$</td>
<td>D (pRW3024)</td>
</tr>
<tr>
<td>pRW3008</td>
<td>8</td>
<td>A</td>
<td>(CCG)$_k$</td>
<td>D (pRW3024)</td>
</tr>
<tr>
<td>pRW3330</td>
<td>9</td>
<td>B</td>
<td>(CCG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3010</td>
<td>10</td>
<td>A</td>
<td>(CCG)$_h$</td>
<td>D (pRW3024)</td>
</tr>
<tr>
<td>pRW3017</td>
<td>17</td>
<td>A</td>
<td>(CCG)$_h$</td>
<td>D (pRW3024)</td>
</tr>
<tr>
<td>pRW3331</td>
<td>20</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3024</td>
<td>24</td>
<td>A</td>
<td>(CCG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3332</td>
<td>26</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3029</td>
<td>29</td>
<td>A</td>
<td>(CCG)$_k$</td>
<td>E (pRW3024)</td>
</tr>
<tr>
<td>pRW3032</td>
<td>32</td>
<td>A</td>
<td>(CCG)$_k$</td>
<td>E (pRW3024)</td>
</tr>
<tr>
<td>pRW3341</td>
<td>35</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3328</td>
<td>44</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3329</td>
<td>46</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3337</td>
<td>54</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3335</td>
<td>61</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3334</td>
<td>71</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>D (pRW3311)</td>
</tr>
<tr>
<td>pRW3338</td>
<td>81</td>
<td>A</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3336</td>
<td>81</td>
<td>C</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3335</td>
<td>98</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3339</td>
<td>105</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3327</td>
<td>127</td>
<td>A</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3326</td>
<td>130</td>
<td>A</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3306</td>
<td>160</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3321</td>
<td>160</td>
<td>A</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3308</td>
<td>160</td>
<td>C</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3309</td>
<td>240</td>
<td>B</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
<tr>
<td>pRW3310</td>
<td>240</td>
<td>C</td>
<td>(CCG)$_h$ CAG(CGG)$_k$</td>
<td>(pRW3311)</td>
</tr>
</tbody>
</table>

* Inserts cloned into pUC19NotI; A; pUC19, B; and pBR322, C.

b Sequences are shown based on the leading strand of the plasmid.

c pRW3024, pRW3311, pRW3306 and pRW3309 were prepared by the direct cloning methods as shown in Figure 1(a) and Figure 2. For the other plasmid constructions, the inserts were isolated from the plasmid indicated in the parenthesis and recloned into the appropriate vector. A series of CGG inserts were obtained by isolating either an expanded (E) or deleted (D) CGG repeat sequence.
(CGG)$_{10}$ was ligated into the same restriction site of pRW3342 and pRW3343 to give pRW3347 (orientation I) and pRW3348 (orientation II), respectively.

The integrity of the derivative plasmids and the lengths of their inserts were verified by restriction endonuclease digestion followed by agarose and polyacrylamide gel electrophoresis. Since the CGG containing restriction fragments migrate abnormally fast compared to random sequence DNA (size markers) on non-denaturing polyacrylamide gels (Gellibolian et al., unpublished results), the lengths of the inserts from the results on agarose gels. Complete or partial sequencing was accomplished using the dideoxy chain termination method (see below).

**Sequencing of CGG-containing plasmids**

All primers were purchased from New England Biolabs Inc. (M13/pUC sequencing primer (−20), M13/pUC reverse sequencing primer (−24), pBR322 EcoRI site primer (clockwise), and pBR322 HindIII site primer (counter-clockwise)). The 5' terminus of each primer was radiolabeled using [γ-32P]ATP and T4 polynucleotide kinase. Oligomers were desalted and separated from unincorporated radioactive nucleotides using NENSORB 20 (DuPont) or Sephadex G-50 (Pharmacia). We used the modified method for dideoxy sequencing using Sequenase Version 2.0 to read longer CGG inserts. Denatured plasmid must be freshly made for sequencing; storage of denatured DNA at −20°C caused polymerase pausing during primer extension of CGG-containing plasmids. Plasmids (3 to 5 µg) were resuspended in 18 µl of water and 2 µl of 10 M NaOH and incubated for 15 minutes at 37°C. Strong alkaline conditions (1 M NaOH) were used for complete denaturation. The mixtures were placed on ice and diluted with 132 µl of cold water. A 19 µl volume of 3 M sodium acetate (pH 5.2) was added to neutralize the solution. Denatured plasmids were precipitated with 2.5 volumes of ethanol (at −20°C for several hours). DNA was collected by centrifugation (16,000 × g) for ten minutes at 4°C and the pellet was rinsed with 70% (v/v) ethanol (500 µl) and dried. Use of a primer proximal to the CGG insert is crucial for preventing polymerase pausing during the primer extension reactions. Note that we observed enhanced polymerase pausing inside both the CTG and CGG triplet repeats when a distal primer was used (Kang et al., 1995b). On ice, 5 µl (2 to 3 ng) of labeled primer was added to the denatured plasmid and the solution was mixed with 6 µl of water and 2 µl of 5× Sequenase reaction buffer (200 mM Tris-HCl (pH 7.5), 100 mM MgCl$_2$, 250 mM NaCl). The primer was annealed to the denatured DNA by heating the solution at 65°C for five minutes, and then slowly cooling to 30 to 37°C over a period of 30 minutes (not to exceed one hour). The tube was chilled on ice, and 1 µl of 0.1 M DTT and 6.5 units of Sequenase were added. Increasing the deoxynucleotide concentration with respect to the dideoxynucleotide concentration is necessary to extend the sequence further from the primer and to obtain equal intensities of sequencing ladders on all four (A, G, C and T) lanes. The solution (3.5 µl) was transferred to 2.5 µl of each ddNTP termination mixture (160 mM ddNTPs, and 7.3 mM ddNTP). After ten minutes incubation at 37°C, 4 µl of stop solution (95% (w/v) formamide, 20 mM EDTA, 0.05% (w/v) bromophenol blue, and 0.05% (w/v) xylene cyanole FF) was added, the DNA was heat denatured at 90°C for six minutes, and analyzed by electrophoresis on 10% (or 8%) Long Ranger gels containing 30% (or 40%) formamide and 7 M urea. Autoradiography of the gel was done using either Fuji RX film or Hyperfilm-MP (Amersham). Using this modified method, we could read 100 CGG repeats on the CGG strand and 40 CGG repeats on the CCG strand. The sequences of both complementary strands were determined for all plasmids in Table 1.

**Methylation of CGG-containing plasmids**

A 5 µg sample of pRW3308 was methylated with 20 units of SssI methylase in the presence of 160 mM S-adenosylmethionine in 10 mM Tris-HCl (pH 7.9), 10 mM MgCl$_2$, 50 mM NaCl, 1 mM DTT at 37°C for 12 hours. If necessary, additional S-adenosylmethionine was added to the reaction mixture to obtain a complete reaction. Complete methylation was assessed by digestion with ActI (CCGC/GCGG) which is a methylation-sensitive restriction enzyme which recognizes every CGG repeat sequence. Digestion was analyzed on a 1.0% agarose gel and the DNA was found to be completely methylated (fully resistant to ActI digestion). Methylated pRW3308 was extracted with phenol and precipitated with ethanol.

**Two-dimensional agarose gel electrophoresis**

Topoisomer populations were generated as described previously (Singleton & Wells, 1982). Both methylated and non-methylated pRW3308 (6.2 µg) were incubated in 150 µl solution containing 10 mM Tris-HCl (pH 7.6), 50 mM KCl, 1 mM EDTA, 10 mM 2-mercaptoethanol, and 0 to 12.5 µM ethidium bromide. Chicken erythrocyte topoisomerase was added and the reaction was carried out for 12 hours at 37°C. DNAs were purified by two extractions with phenol and three extractions with ether, and precipitated with ethanol. Topoisomers (1.5 µg/lane) were electrophoresed through 1% agarose at 3.3 V/cm in TBE for 21 hours at room temperature. For the second dimension, gels were run for 18 hours at 3.3 V/cm in TBE containing 1 µM chloroquine. Gels were stained with ethidium bromide and photographed.

**Acknowledgements**

We thank E. Klysik and J. Harris for technical assistance, Dr A. Bacolla for the preparation of several plasmids, and Des J. Klysik, A. Jaworski, S. Amirhaeri, S. Kang and D. L. Nelson for helpful discussions. This work was supported by the N.I.H. (GM52982) and the Robert A. Welch Foundation.

**References**


CGG Triplet Repeats from the FMR-1 Gene

FMR1, has characteristics of an RNA-binding protein. *Cell*, 74, 291–298.


Edited by J. Karn

(Received 27 October 1995; received in revised form 9 February 1996; accepted 13 February 1996)