Molecular Basis of Genetic Instability of Triplet Repeats*

Robert D. Wells

From the Institute of Biosciences and Technology, Center for Genome Research, Department of Biochemistry and Biophysics, Texas A&M University, Houston, Texas 77030-3303

Background and Overview

A veritable explosion is taking place in our understanding of the human genetics, biochemistry, and DNA structural issues related to human hereditary neuromuscular and neurodegenerative diseases. Also, the non-Mendelian expansion process that elicits these disease manifestations (anticipation) is under intense investigation. Within the last 3 years, the molecular basis of 10 human genetic disorders (including fragile X syndrome (FRAXA and FRAXE), myotonic dystrophy (DM),1 Kennedy's disease, Huntington's disease (HD), spinocerebellar ataxia type 1 (SCA1), and dentatorubral-pallidoluysian atrophy (DRPLA)) has been partially established (reviewed in Refs. 1–5). The diseases are characterized at the molecular level by the expansion of a simple triplet repeat (CTG and CCG) from less than 15 copies of the repeat in normal individuals to scores of copies in affected cases; thousands of copies are found in some cases of fragile X and myotonic dystrophy. These increases in size occur upon passage of an expanded repeat in the chromosome to offspring. Moreover, the symptoms of these diseases follow an unusual genetic pattern called anticipation, in which the disease becomes more severe and has an earlier age of onset with each successive generation (reviewed in Refs. 1–6). The instability of repeats in the genome has also been linked to hereditary nonpolyposis colon cancer, which may involve mutations in mismatch repair functions (4, 5, 7–9).

For example, Huntington's disease shows anticipation and has expanded CAG triplet repeats (1–5, 10). A CAG repeat of between 11 and 34 bp in the normal population encodes a polyglutamine tract in the IT15 gene. Expansion to about 90 bp occurs in HD patients. The age of onset correlates with the length of the triplet repeat with the largest changes in repeat lengths seen upon paternal transmission (11). Sperm display a heterogeneous expanded repeat length. An intermediate allele, IA, containing 30–38 (or 34–38) repeats (perhaps similar to a premutation in fragile X or DM) has been identified. Initial reports suggest that sporadic expansion of the IA allele occurs only through paternal transmission (12, 13). The function of the gene product is uncertain (14).

Considering Mendelian genetic principles, anticipation was an enigma. The discovery of expanding triplet repeats (or "mutable mutations") in diseases showing anticipation afforded a physical basis for this unusual genetic phenomenon. Expansion of the triplet repeat is responsible for the genetic defect, influencing the activity of a glutamine-containing protein (SBMA, HD, SCA1, and DRPLA) or influencing the level of expression of a gene with which the repeat is associated (fragile X and DM) (1–5). All triplet repeat genetic diseases identified to date show anticipation. Several other diseases also show anticipation including spinocerebellar ataxia type 2 (15), bipolar affective disorder (16), and hereditary spastic paraparesis (Strumpell’s disease) (17). If a correlation exists between anticipation and triplet repeats, many more diseases showing anticipation may be identified since there are more than 40 genes containing associated triplet repeats.

Expansions and Deletions

An understanding of the molecular mechanisms of triplet repeat instabilities (expansions and deletions) is important for the comprehension of anticipation. Kang et al. (18) have established a defined genetic system that shows promise for the dissection of this process. The frequency of genetic expansions or deletions in Escherichia coli depends on the direction of replication (18). Large expansions occur predominantly when the CTGs are in the leading template strand rather than the lagging strand. However, deletions are more prominent when the CTGs are in the opposite orientation (Fig. 1). Most deletions generate products of defined size classes. Strand slippage coupled with non-classical DNA structures (Fig. 2) probably accounts for these observations and relates to expansion-deletion mechanisms in eukaryotic chromosomes. To study expansions, these workers determined if a plasmid that contains (CTG)130 is completely homogenous as a cloned molecule or if deletions and expansions had occurred that gave rise to sequence heterogeneity, even in a tiny percent of the molecules. The insert containing the triplet repeat was excised from the vector and separated by gel electrophoresis. The regions of the gel either above or below the insert band were eluted and "recoloned"; recombinant plasmids were obtained that contained successively larger or smaller inserts, respectively. The family of inserts characterized by these methods contained repeat units ranging from 17 to 300. Hence, expansion and deletion occur in E. coli. This discovery lays the foundation for evaluating host cell genetic factors (replication, recombination, mismatch repair, etc.) that may elicit genetic instabilities. DNA sequence analyses showed that expansion and contraction always occurred in multiple repeats of 3 bp. Prior investigations (19) showed that deletions in dinucleotide repeat sequences occurred in multiple units of 2 bp.

Fig. 1 outlines a possible mechanism for the expansion and deletion behaviors. For expansion, a hairpin loop may form on the lagging strand nascent DNA (CTG strand). NMR investigations (20) revealed that CTG oligomers form a stable anti-parallel duplex with TT pairs, whereas the complementary CAG strand forms a metastable conformation. When the CTG is the lagging strand template (orientation II), a loop may form on the lagging strand that will be bypassed during DNA synthesis to generate deletions. Multiple slippages (6) may be promoted by an "idling polymerase" caused by a strong block such as a DNA structure or the presence of proteins (21), which causes continuous slippage (primer realignment) resulting in the expansion of larger sequences. Other workers (22) favor gene conversion events to explain germline mutations at

* This minireview will be reprinted in the 1996 Minireview Compendium, which will be available in December, 1996. This work was supported by National Institutes of Health Grant PO1 GM52982 and by a grant from the Robert A. Welch Foundation.

1 The abbreviations used are: DM, myotonic dystrophy; HD, Huntington's disease; SCA1, spinocerebellar ataxia type 1; DRPLA, dentatorubral-pallidoluysian atrophy; bp, base pair(s); SBMA, spinobulbar muscular atrophy.
human minisatellites. Evolutionary studies (23) on the cryptic FMR1 CGG repeat suggest that replication slippage and unequal crossing over have been operative for 150 million years.

CTG Is Preferentially Expanded

Ohshima et al. (24) have recently discovered that the CTG triplet repeat is the dominant genetic expansion product in E. coli. This extraordinary discovery was made possible by the successful cloning and characterization of all 10 repeating triplet sequences. The relative capacity of the 10 repeating triplet sequences to be expanded in E. coli (18) was explored with a competition study. Surprisingly, the CTG triplet repeat was expanded at least nine times more frequently than any of the other nine triplets (24). Low levels of expansion were found also for CGG, GTG, and GTC. Thus, the structure of the CTG repeat and/or its utilization by the DNA synthetic systems in vivo must be quite different from the other triplets. The surprising discovery that CTG triplet repeats are the dominant expansion products in E. coli, as found (1-5) in clinical samples from human hereditary diseases, suggests the importance of DNA structural properties (25). Other investigations have revealed that duplex CTG and CGG repeats have unorthodox properties including nucleosome assembly (26), their capacity to cause DNA polymerases to pause within the repeat sequences (27), as well as conformational features as revealed by helical repeat and polyacrylamide gel migrations (discussed below). Further elucidation of the CTG repeat structural features along with the genetic factors responsible for expansion may explain why most (8 out of 10) (1-5) triplet repeat hereditary disease genes contain CTG repeats. Although other triplet repeats are found in the human genome (29), the lengths are shorter (generally <15 repeats) than found for these disease genes. Other work (30) has shown that the CTG triplet repeat is expanded in E. coli distal to the replication origin as a single large event of ~120 bp.

In summary, these investigations (18, 24, 30) establish a genetically defined system for studying the molecular mechanisms of this non-Mendelian process. A recent report of a transgenic mouse model for SBMA (31) found no change in length with transmission. Bacterial systems may provide useful mechanistic information until a genetically defined eukaryotic system can be established. In fact, a number of similarities exist between the behaviors observed in humans and this E. coli system (reviewed in Ref. 24).

DNA Polymerase Pausing

As an accidental discovery as part of chemical probe analyses, the pausing of DNA synthesis in vitro at specific loci in double-stranded CTG and CGG triplet repeats was found (27). The DNA syntheses of CTG triplets ranging from 17 to 180 and CGG repeats from 9 to 160 repeats in length were studied in vitro. Primer extensions using the Klenow fragment of DNA polymerase I, the modified T7 DNA polymerase (Sequenase), or the human DNA polymerase β paused strongly at specific loci in the CTG repeats. The pausings were abolished by heating at 70 °C. As the length of the triplet repeats in duplex DNA, but not in single-stranded DNA, was increased, the magnitude of pausings increased. CGG triplet repeats also showed similar.

---

2 K. Ohshima and R. D. Wells, unpublished work.

---

FIG. 1. A model for orientation-dependent instability of CTG repeats during replication. Top panel, expansion in orientation I; center panel, deletion in orientation II; bottom panel, possible hairpin structures adopted by CTG and CAG single-stranded DNA. A similar model was proposed (49) to explain the greater genetic stability for long CGG-CCG tracts when CGG is in the leading template strand (orientation I) compared with CCG in the leading template strand (orientation II).

FIG. 2. DNA structures of triplet repeats. Slipped structures and toroids are not mutually excluded.
but not identical, patterns of pausings. These results indicate that appropriate lengths of the triplets adopt a non-B conformation(s) that blocks DNA polymerase progression; the resultant idling polymerase may catalyze slippages (Fig. 1) to give expanded sequences and, hence, provide the molecular basis for this non-Mendelian genetic process. Also, recent in vivo replication studies in E. coli with plasmids containing the CTG repeat revealed length-dependent pause sites. Other studies with single-stranded (CGG)$_{20}$ as a template suggest a K$^+$-dependent structure (tetraplex) that serves as a barrier to DNA synthesis in vitro.

**Mismatch Repair**

Mismatch repair-deficient E. coli (33) were studied in order to further elucidate the factors involved in genetic instabilities as well as DNA structural issues in vivo. Long CTG repeats are stabilized in ColE1-derived plasmids in E. coli containing mutations in the methyl-directed mismatch repair genes (mutS, mutL, or mutH) (34). When plasmids containing (CTG)$_{180}$ were grown for about 100 generations in mutS, mutL, or mutH strains, 60–85% of the plasmids contained a full-length repeat, whereas in the parent strain only about 20% of the plasmids contained the full-length repeat. The deletions occur only in the (CTG)$_{180}$ insert, not in DNA flanking the repeat. While many products of the deletions are heterogeneous in length, preferential deletion products of about 140, 100, 60, and 20 repeats were observed. The E. coli mismatch repair proteins apparently recognize three-base loops formed during replication and then generate long single-stranded gaps where stable hairpin structures may form, which can be bypassed by DNA polymerase during the resynthesis of duplex DNA. Similar studies were conducted with plasmids containing CGG repeats; no stabilization of these triplets was found in the mismatch repair mutants. Since prokaryotic and human mismatch repair proteins are similar (33, 35) and since several carcinoma cell lines, which are defective in mismatch repair, show instability of simple DNA microsatellites (7–9), these mechanistic investigations in a bacterial cell may provide insights into the molecular basis for some human genetic diseases.

**DNA Structure**

Simple repeat sequences in plasmids adopt non-B conformations under appropriate conditions (such as negative supercoil density, ionic strength, etc.) in vitro (reviewed in Refs. 6 and 36). For example, mirror repeat purine-pyrimidine sequences form triplexes (H-DNA) and (in certain cases) nodule DNA, alternating purine-pyrimidine sequences adopt left-handed Z-DNA, inverted repeats form cruciforms, and repeating Atracts exist in bent (curved) conformations. Some unusual structures were proven to exist in vivo in plasmids (6, 36) and in chromosomes (37). Several recent biophysical studies were reported (38–43) on short (generally <24 bp) synthetic oligonucleotides with CTG or CGG triplets, which, in general, support the concept of hairpin loops (Figs. 1 and 2) and other ordered conformations.

Long CGG and CTG triplet repeat duplex sequences adopt intrinsic structures best explained as toroids (32) since the toroidal conformations (Fig. 2) might provide a suitable homing site. Nucleosomes are the basic structural elements of chromosomes and consist of 146 bp of DNA coiled about an octamer of histone proteins that mediate general transcriptional repression. Plasmids containing lengths of CTG from 0 to 250 repeats were investigated (26). The efficiency of nucleosome formation increased with expanded triplet blocks suggesting that such blocks may repress transcription through the creation of stable nucleosomes (Fig. 3). In fact, the expanded CTG triplet repeats are the strongest known nucleosome positioning element (46), even compared to the Xenopus borealis somatic 5 S RNA gene, one of the strongest known natural nucleosome positioning sequences.

In summary, we believe that three types of non-B DNA conformations are important for triplet repeats (Fig. 2). The toroidal structure formed with duplex CGG and CTG sequences is dictated solely by these triplet repeat sequences. We assume that the toroid is a suitable homing site for histone octamer binding. Slipped structures are the only reasonable explanation for the observed mismatch repair results (34). Hence, this may be the first case where a non-B structure has been detected in vivo prior to its in vitro characterization. Also, hairpin loops may be formed by single-stranded regions during DNA replication (Fig. 1).

**Genetic Instabilities**

Several factors influence the stability of the triplet repeat inserts. First, the type of sequence plays a major role with CGG

---

4 S. M. Mirkin, personal communication.
being the most difficult to stably maintain in E. coli (49). Second, the length of the repeats is very important since longer tracts, especially for CGG, show a greater degree of instability compared with shorter inserts (30 or less). This behavior in E. coli is consistent with the mechanism of genetic anticipation for the fragile X syndrome (47). Third, the presence of interruptions greatly enhances the stability of triplet repeats especially for CGG. Alleles derived from human patients show the presence of stable and unstable CGG triplets of similar size, suggesting that a feature other than length, but intrinsic to the repeat, was responsible for stability. Eichler et al. (48) found that lengths of >33 uninterrupted CGGs 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.

Fifth, the strains of E. coli (49). Third, the presence of interruptions of up to 160 repeats in pUC-derived plasmids (compared with HB101, STBL2, and RS2). Inserts containing longer than 160 CGG repeats were extremely unstable in pUC19 and were prone to delete to smaller sized plasmids. Hence, the vector of choice is significant also. Sixth, the location of the insert in the vector is important may and may relate to the pairing observed in the DNA polymerase I/II switch site (27). Seventh, the copy number of the insert is a major factor as well as all physical determinations. The goal is to understand the molecular mechanisms responsible for genetic instabilities and to eventually eradicate these devastating human neuromuscular and neurodegenerative diseases.

Acknowledgments—I thank my collaborators for diligent and talented efforts that made this review possible and R. Iyer for assistance in preparation of figures.

REFERENCES
1. Davies, K., and Warren, S. (eds) (1993) Genome Analysis, Vol. 7, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1957–1981.
2. Bates, G., and Lehrach, H. (1994) BioEssays 15, 277–284.
3. Sutherland, G. R., and Richards, R. I. (1995) Proc. Natl. Acad. Sci. U. S. A. 92, 3636–3641.
4. Panzer, S., Kuhl, D. P. A., and Caskey, C. T. (1995) Stem Cells 14, 156–157.
5. Krahe, R., and Ashizawa, T. (1995) in Hypervariable Genetic Markers (Wetherill, J., and Groth, D., eds) pp. 29–60, CRC Press, Inc., Boca Raton, FL.
6. Wells, R. D., and Sinden, R. R. (1993) in Genome Analysis (Davies, K., and Warren, S., eds) Vol. 7, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
7. Thibodeau, S. N., Bren G., and Schaid, D. (1993) Science 260, 816–819.
8. Shibata, D., Peinado, M. A., Ionov, Y., Mikkola, S., and PerUCH, M. (1994) Nat. Genet. 6, 273–281.
9. Baker, J. S., Münkner, C. E., Zhang, L., Plug, A. W., Rotzack, M., Warren, G., Elliott, E. A., Yu, J., Ashley, T., Arniune, N., Flavell, R. A., and Liskay, R. M. (1995) Cell 82, 309–319.
10. Huntington’s Disease Collaborative Research Group (1993) Cell 72, 971–983.
11.swick, A. B., Pizzuti, A., Pfeiffer, M., Sutcliffe, J., Richards, S., Verkerk, A. M. H., Hoogen, R. J., Warren, S. T., and Nelson, D. L. (1995) Cell 82, 309–319.
12. Shibata, D., Peinado, M. A., Ionov, Y., Mikkola, S., and Peruch, M. (1994) Nat. Genet. 6, 273–281.
13. Baker, S. M., Münkner, C. E., Zhang, L., Plug, A. W., Rotzack, M., Warren, G., Elliott, E. A., Yu, J., Ashley, T., Arniune, N., Flavell, R. A., and Liskay, R. M. (1995) Cell 82, 309–319.