united nations educational, scientific and cultural organization

international atomic energy agency

the
abdus salam
international centre for theoretical physics

SMR: 1098/5

WORKSHOP ON THE STRUCTURE OF BIOLOGICAL MACROMOLECULES

(16 - 27 March 1998)

"The Human Genome and Unusual DNA Structures"

presented by:

E. Morton BRADBURY
Life Sciences Division, LS-2, M88
Los Alamos National laboratory
Los Alamos, New Mexico 87545
U.S.A.

NUCLEOSOME AND CHROMATIN STRUCTURES AND FUNCTIONS

Sari PENNINGS
Department of Biochemistry
Hugh Robson Building
George Square
University of Edinburgh
Edinburgh EH8 9XD
Scotland, U.K.

E. Morton BRADBURY
Los Alamos National Laboratory
MS M888
Los Alamos, New Mexico 87545
Department of Biological Chemistry
School of Medicine
UC-Davis
Davis, California 95616

The diploid human genome contains 6x10^9 bp of DNA of total length 2.04 m packaged into cell nuclei 6-8 µm in diameter. Despite decades of intensive research we are still far from understanding the rules that govern the packaging of these enormously long eukaryotic DNA molecules into chromosomes and cell nuclei. Some answers will come from the sequence data generated by the Human Genome Project, particularly the identification of sequence motifs involved in both the long range organization of chromosomes and in nuclear architecture. Such sequence motifs probably bind to proteins in the chromosomal scaffold, the nuclear matrix and nuclear membrane. How chromosome organization and nuclear architecture are involved in chromosome function is not well understood. In an attractive working model for long-range order in metaphase chromosomes the DNA is constrained by scaffold proteins into loops of average size 50 kbps [1]. These loops are packaged by the histones H1, H2A, H2B, H3 and H4 into nucleosomes and higher order chromatin structures.

1. Histones

Histones H3 and H4 are among the most rigidly conserved proteins in nature which implies that every residue in these proteins is essential for their functions. Histones H2A and H2B are more variable and each comprises a family of proteins. The syntheses of some members of the H2A and H2B families are cell cycle dependent e.g., H2A1, H2A2 and others are not, e.g., H2AX, H2AZ. The very lysine rich histones are the most variable of the histones and also comprise a protein family, some members of

C. Nicolini (ed.), Genome Structure and Function, 111–126.

which are cell specific. These different histone subtypes provide considerable potential for variability in nucleosome structures and functions.

Histones are the major structural proteins found in chromosomes. The highly conserved histones H3 and H4 are involved in essential interactions in generating the structural framework of the nucleosome which is then completed by the binding of the more variable histones H2A and H2B and the most variable very lysine rich H1 histones [see 2,3]. Histones are multidomain proteins [3]: i) each of histones H3 and H4 has a flexible basic N-terminal domain and a structural apolar central and C-terminal domain which is involved in interactions between H3 and H4 as shown by nuclear magnetic resonance (NMR) spectroscopy [4] and x-ray crystallography [5]; ii) histones H2A and H2B have flexible basic N-terminal domains and C-terminal tails [6]. Their conserved apolar central domains are structured and involved in interactions between H2A and H2B [5,6]; and iii) all H1 subtypes and H1° and H5 have three well-defined domains; a flexible basic N-terminal domain, a central globular domain and a flexible basic C-terminal half of the molecule [7-9]. Histones have been shown to form specific complexes [see 2]. These are the (H2A, H2B) dimer, the $(H3_2,H4_2)$ tetramer and the histone octamer $[(H2A,H2B)_2\ (H3,H4_2)]$ that forms the protein core of the nucleosome. The nucleosome is completed by the binding of the fifth histone H1.

1.1 HISTONE MODIFICATIONS

Histones are subjected to reversible chemical modifications that change the chemical nature of the modified residues; acetylations of lysines in the N-terminal domains of H2A, H2B, H3 and H4 and phosphorylations of serines and threonines in the basic N and C-terminal domains of H1, H3 and H2A [see 3, 10-12]. In addition H2A and H2B are modified by the reversible covalent attachment of ubiquitin to lysines in the C-terminal tails of H2A and H2B [see 3,13]. Acetylations and ubiquitinations modify only about 5% of the core histones and thus can affect only small subcomponents of chromatin. In contrast, all of the H1 and H3 protein molecules are phosphorylated at metaphase and these phosphorylations appear to be required for general chromosome functions at mitosis.

The acetylations of the core histones have been associated with chromatin replication [10-12], transcriptionally active and potentially active genes [see 3,10,11] and the replacement of histones by protamines during spermiogenesis [14], i.e., all aspects of DNA processing. Ubiquitinations of histones have been associated with potentially active chromatin [13,15] and with nucleosomes containing heat shock genes in the non-induced state [16]. Ubiquitinated H2A is absent in metaphase chromosomes [17] and we have shown that uH2A and uH2B are deubiquitinated shortly before metaphase and are reubiquitinated in anaphase leading to the suggestion that ubiquitin labels an important subset of genes, and has to be removed prior to metaphase to allow the correct packaging of nucleosomes into metaphase chromosomes [18]. The phosphorylations of the very lysine rich histones have been strongly implicated in the initiation and control of chromosome condensation [3, 19-21], a process which also requires, as a later event, the phosphorylation of histone H3 [21]. In relating chromatin structure and function it is significant that all of the reversible chemical modifications are located in the basic, flexible N and C-terminal domains of the histones [see 3]. Reversible histone modifications most probably provide the mechanisms for modulating chromatin structure in response to cell functions. However, the effects of reversible chemical modifications on chromatin structure, stability and accessibility are largely unknown. Also unknown are the locations of the basic flexible N- and C-terminal regions of histones in nucleosomes and higher order chromatin structures.

2. Nucleosome Structure

Since its discovery in 1973, the nucleosome has been the focus of studies directed towards an understanding of chromatin structure and function [see 2]. For most somatic cells, the nucleosome contains 195 ± 10 bp DNA, the histone octamer and histone H1. Nucleosomes from some specialized cells contain different DNA repeats; chicken erythrocytes (212 bp), sea urchin sperm 241 bp, rabbit neuronal cells 165 bp [see 22-24]. The ends of the DNA can be further trimmed by micrococcal nuclease digestion to give well-defined sub-nucleosome particles. These are the chromatosome with 168 bp DNA and the full histone complement [25] and the core particle with 146 bp DNA and the histone octamer [see 3]. The well-defined core particle has been subjected to intensive structural studies. Neutron scatter studies of these particles in aqueous solution proved that DNA was coiled around the histone octamer core [26-31]. From the neutron scatter curves and pair distance distribution functions, it could be deduced that in solution the core particle was a disc 11.0 nm in diameter by 5.5 to 6.0 nm thick with 1.7 ± 0.2 turns of DNA of mean radius 4.5 nm coiled with a pitch of 3.0 nm on the outside of the particle [27-31]. Low resolution x-ray (2.0 nm) and neutron (1.6 nm) diffraction of core particle crystals gave a model of a wedge shaped disc 11.0 x 5.7 nm with 1.8 turns of DNA of mean radius 4.4 nm and pitch of 2.8 nm [32-35] showing that at low resolution the solution and crystal structures are very similar. The resolution of the core particle crystal structure was extended by x-ray diffraction to 0.7 nm [30]. In the 0.7 nm structure the DNA is not uniformly bent around the octamer but follows a more irregular path with bends. The calculated radius of gyration, Rg, for the observed histone octamer electron density in the core particle crystal structure of 2.97 nm is substantially lower than that determined by neutron scatter contrast matching of 3.3 nm [26,28-31]. However, not all of the electron density is accounted for in the core particle crystal structure indicating the presence of disordered regions in the histones. This difference between the neutron scatter octamer Rg of 3.3 nm and the calculated Rg of 2.97 nm for the observed histone electron density in the crystal structure, has been attributed to disorder in the N- and C-terminal flexible, basic domains of the histones in the core particle [31]. Further support for this proposal comes from our findings [37] that controlled proteolytic removal of the N- and C-terminal domains from the histone octamer reduces its Rg from 3.35 nm to 2.98 nm. More recently the crystal structure of the histone octamer has been solved to 0.33 nm resolution [5]. This shows the modes of interactions between the structured, apolar, central regions of the core histones in the (H2A,H2B) dimer, ($H3_2$,$H4_2$) tetramer and the histone octamer. However, no electron density was observed for the flexible N- and C-terminal domains of the core histones [5] most probably because of static or dynamic disorder of these regions. These disordered regions correspond exactly with the disordered, mobile regions identified by NMR spectroscopic studies of the (H2A,H2B) dimer and the ($H3_2$,$H4_2$) tetamer [4,6]. An NMR study of the histone octamer and trypsin-trimmed histone octamer showed clearly that the N- and C-terminal tails are mobile [38]. These regions correspond exactly with the disordered regions in the 0.33 nm solution crystal structure of the histone octamer [5]. An understanding of chromatin structure/function relationships will require details of the molecular interaction of these flexible basic N-terminal domains and C-terminal tails in chromatin and the effects of the reversible chemical modifications on these interactions. Previously, we have shown that histone hyperacetylation has little effect on the shape of the 146 bp core particle [39]. Based on these observations it was predicted that histone hyperacetylation would exert its effects on the DNA entering and leaving the nucleosome i.e., on the DNA regions outside of the core particle 146 bp DNA. Recent x-ray scatter studies of fully defined 195 bp acetylated nucleosome particles support this prediction [manuscript in preparation]. Understandings of the modes of interactions and the functions of the flexible basic N-terminal domains of the core histones and the C-terminal tails of histones H2A and H2B are central to understanding nucleosome structures and function. Recently, S.I. Usachenko in our group has mapped the histone DNA contacts in nucleosome core particles from which the C- and N-terminal regions of histone H2A were selectively trimmed by trypsin or clostripain [40]. It was found that the flexible trypsin sensitive C-terminal "tail" of H2A contacted the DNA at the dyad axis, whereas its globular domain contacts the end of the 146 bp DNA in the core particle. The appearance of the H2A contact at the dyad axis was found only in the absence of linker DNA and did not depend on the absence of linker histones. In the absence of linker histones no contact of H2A with the DNA at the dyad axis was observed. It was presumed under these conditions in native H1 depleted chromatin that the C-terminal "tail" bound to the linker DNAs entering and leaving the nucleosome. These results demonstrate the ability of the histone H2A C-terminal tails to rearrange. This rearrangement may play a role in nucleosome disassembly and reassembly and the retention of the H2A/H2B dimer or octamer during the passage of polymerases through the nucleosome.

The current models for the chromatosome and the nucleosome are based on the crystal and solution structures of the 146 bp core particle. The model for the chromatosome contains two full turns of DNA that are coiled around the histone octamer and complexed with the fifth histone H1 [24,41] most probably through the binding of the H1 globular domain [42,43]. The nucleosome model contains in addition the linker DNAs which join adjacent chromatosomes. The paths of these linker DNAs in the different order of chromatin structures are not known. Major outstanding questions are: i) the locations and modes of binding of the N-terminal domains which through reversible chemical modification are involved in chromatin functions; ii) the mode(s) of binding of the very lysine rich histones, and iii) the DNA paths entering and leaving the nucleosome.

3. Chromatin Structure

Chromatin is made up of repeating subunits, the nucleosome, joined by the continuity of the DNA molecule. At low ionic strength, chromatin is in an extended form first described as the 10 nm fibril. Neutron scatter studies of extended chromatin gave a mass per unit length equivalent to one nucleosome/10 ± 2 nm i.e., a DNA packing ratio of about 6 to 7 [44]. The measured cross-section radii of gyration of the DNA and histone components required an arrangement of flat discs with their faces roughly parallel to the axis of the fibril, i.e., edge-to-edge. An edge-to-edge zig-zag arrangement of nucleosome discs has also been proposed from E.M. studies [41].

On increasing the ionic strength, the extended form of chromatin undergoes a transition to the 30 nm fibril. Neutron scatter studies of this transition suggest a family of supercoils undergoing increasing compaction with increasing ionic strength. In its most compact form, the hydrated supercoil has a mass per unit length equivalent to 6-7 nucleosomes per turn of a coil of pitch 11.0 nm and outer diameter of 34 nm [44]. It is described by the previously proposed supercoil [45] or solenoid [46] of nucleosomes. In the fiber diffraction pattern of chromatin, the small number of diffuse diffraction features and their orientation are explained at low resolution by this supercoil of nucleosomes [45,47]. This supercoil or solenoid is a one start helix. An alternative proposal is based on the E.M. observations of an intermediate unfolded state of chromatin in which adjacent nucleosomes form a close-packed zig-zag [48,49]. Recent cryo E.M. studies of oligonucleosomes support an irregular zig-zag conformation of chromatin in solution [50]. In one model this zig-zag ribbon is coiled into a supercoil, i.e., a two-start helix [49]. Additional to being a one-start or a two-start helix these models differ markedly in the locations of histone H1 and linker DNA. In the former model, linker DNA and H1 are located in the hole along the axis of the supercoil of nucleosomes [46,47] whereas in the latter model H1 and linker DNA are located between adjacent nucleosomes in the coiled ribbon, i.e., in the "wall" of the nucleosome coil. A third type of model [51] is a two-start left-handed helix with linker DNA crossing from one side of the solenoid to the opposite side, similar to an earlier proposal [52]. Neutron scatter studies of long chromatin reconstitated with deuterated H1 have shown that the bulk of H1 is located in the hole along the axis of the 34 nm supercoil of nucleosomes [53]. This provides strong support for the supercoil [45] or solenoid [46,47] models for the "30 nm" filaments. More recently atomic probe microscopy studies of chromatin structure by van Holde and colleagues [54-57] have raised questions concerning the degree of order in the "30 nm fibril". One problem with the published models for the "30 nm fibril" [44-49] is that they are depicted as very regular structures extending over large distances. However, the scanning force microscopy studies [54-58] show that the regularity of the "30 nm filament" extends over short distances only. This accords with E.M. data [50] and with previous x-ray [47] and neutron [45] diffraction that give only a small number of diffuse diffraction peaks consistent with a low degree of order in the structure.

4. DNA Packing Ratio of Active Chromatin

In the E.M. active transcriptional units have been observed for the ribosomal RNA genes in the embryo of Oncopeltus fusciatus [59] and the Balbiani rings of the salivary glands of Chironomus tentaus [60,61]. For Balbiani rings a comparison of the length of the transcription product i.e., the 7SRNA with the length of the transcribing chromatin fiber gave a DNA packing ratio of 3 to 4:1. From E.M. tomographic studies, packing ratios of 4 to 8:1 have been obtained for different regions of the Balbiani ring transcription unit [62,63]. The latter value is comparable with the neutron scatter determined packing ratio of the extended 10 nm filament of 6 to 7:1 [44]. Thus, the unfolding of the 34 nm supercoil of nucleosomes to the extended form and beyond is probably a major step in the structural transition from inactive to active chromatin.

5. Factors Involved in Active Chromatin

The changes involved in chromatin structure and stability which precede the passage of RNA polymerase remain poorly defined. Correlations have been found of histone composition and subtypes, histone modifications and non-histone proteins with active chromatin. These include: i) full or partial depletion of histone H1 [64,65], ii) hyper acetylation of the core histones, particularly H3 and H4 [see 10-12]; iii) ubiquitination of histones H2A and H2B [14,16,17]; iv) "active" core particles which selectively bind RNA polymerase II are depleted in one (H2A,H2B) dimer [66] and v) the binding of high mobility group (HMG) proteins [67]. We have shown, however, that full or partial dissociation of the histone octamer is not required for transcript elongation although arrays of nucleosome cores by phage T7 RNA polymerase (see later) [68]. Concerning chromatin structure/function relationships we have shown that histone acetylations, in particular the acetylation of histones H3 and H4 [69,70] cause a reduction in the nucleosome DNA linking number change, ΔL_k, from -1.04 ± 0.08 to -0.82 ± 0.05 thus releasing negative DNA supercoiling from acetylated nucleosomes into a constrained chromatin domain which would facilitate chromatin domain unfolding. Very recently those studies have been extended to the effects of the fully acetylated forms of either histones H3 or H4. It was found that the full acetylation of H4 but not H3 caused the change in the linking number [71]. The sites of reversible ubiquitinations of H2A and H2B are located in their basic C-terminal tails [14,17,18,72]. It has been proposed that ubiquitin labels potentially active chromatin containing heat shock [14] and stress genes [17,18]. Thus, knowing the location of ubiquition and how ubiquitination of H2A and H2B modifies chromatin structure and stability are essential to understanding the functions of histone ubiquitination.

6. Nucleosome Positioning

Virtually all of the DNA in eukaryotic genomes is packaged by histones into nucleosomes and higher order chromatin structures. Some of these nucleosomes have been shown to be precisely positioned on the underlying DNA sequence, most probably for functional requirements [see 72,73]. The factors involved in the precise positioning of nucleosomes are not well-understood. Additional to the identification of nucleosome positioning sequences [see 73] statistical analyses of DNA sequences contained in native core particles suggest the involvement of more general sequence properties. Following an earlier proposal that DNA flexibility or bendability might be a factor in nucleosome positioning [74], sequence constraints as determinants of nucleosome core particle positioning have been identified [75,76]. Based on these findings, earlier observations that long stretches of some homo-nucleotide sequences cannot be assembled into nucleosomes [77,78] can be explained by their increased stiffness relative to native DNA sequences. However, from sequence engineering experiments [79,80] it has been concluded that although DNA bendability should be considered as a general property of DNA sequences, other more specific factors in nucleosome positioning cannot be excluded. Very lysine rich (VLR) histones that are required for the stability of the 30 nm supercoil or solenoid of nucleosomes [41], are thought also to be involved in nucleosome positioning [81].

To study the protein factors involved in nucleosome positioning my laboratory has used DNA substrates containing tandem repeats of nucleosome length DNA from the sea urchin lytechins 5S RNA gene. This DNA repeat has been shown to contain a unique nucleosome positioning sequence [82,83]. Head to tail tandem repeats of nucleosome length DNA have been constructed by Simpson's group [83]; these are 18 repeats of 207 bp DNA $(207)_{18}$ and 45 repeats of 172 bp DNA $(172)_{45}$. The 172 bp sequence contains 7 unique restriction sites and the longer 207 bp sequence contains 8 unique restriction sites. To determine the precise positions of nucleosome cores formed by the histone octamer, the assembled 172_{45} and 207_{18} chromatins were trimmed back to 146 bp core particles by micrococcal nuclease and 5' end-labeled. The 146±2 bp of DNA extracted from these core particles was digested with up to eight restriction enzymes. Analysis on denaturing polyacrylamide gels allowed the core particle boundaries to be mapped relative to the unique restriction sites. This analysis showed that most but not all of the histone octamers assemble on one dominant but not strictly unique position from nucleotide 6 to 153 bp on both the 172_{45} and 207_{18} DNA head-to-tail multimers. The unique position reported for the histone octamer assembled on the 260 bp monomer from 5SrDNA [82] lies 10 to 15 bp downstream from the above major site identified for the 172 bp and 207 bp DNA multimers. This could imply strongly that regions of DNA external to the 172 bp and 207 bp DNA influence the final position of the histone octamer on the 260 bp DNA monomer.

In addition to the dominant position from 6 to 153 bp occupied by a large proportion of the histone octamers, minor positions were identified that flanked the major position. Relative to this major position one of the minor positions in the 207 bp DNA was 10 bp upstream; two minor positions were 10 and 20 bp downstream and two other minor positions were further away, 40 bp downstream and 50 bp upstream. The 172 bp multimer gave the same octamer positions except that the more distant locations of 40 bp downstream and 50 bp upstream were absent. It is to be noted that all of the minor positions on the 172 bp and 207 bp DNA multimers are located in multiples of ±10 bp away from the dominant position. This is significant because 10 bp is the helical repeat of B-form DNA coiled around the histone octamer in the nucleosome core particle [36].

These observations, for nucleosome cores assembled on the tandemly repeated nucleosome positioning DNA sequences, of a major position flanked by minor positions spaced by multiples of 10 bp DNA most probably result from the dynamic nature of the primarily electrostatic interactions between the basic histones and the DNA coil. They raise the possibility that nucleosome cores have the ability to move between the major and minor positions depending on solution ionic conditions and temperature.

7. Effects of VLR Histones on Nucleosome Positions

Mirococcal nuclease digestion of $(207)_{18}$ and $(172)_{45}$ chromatins assembled with histone H5 showed regularly spaced nucleosomes. Both the core particle 146 bp and, importantly, the chromatosome 168 bp nuclease digestion stops were well-defined suggesting that H5 (and H1) provide protection similar to that of native chromatin. Of considerable interest was the finding that the complexity of the 207 bp DNA nucleosome bands discussed above was reduced by H5. The addition of histone H5 appeared therefore to reposition many of the nucleosome cores between the minor position and dominant positions. DNA was extracted from the 208 bp and 172 bp chromatosome bands and digested with up to eight restriction enzymes as described above for the 146 bp DNA from the core particle. The digested DNA from the 207 bp chromatosome corresponded to two major positions with comparable populations 10 bp apart. Surprisingly, the 172 bp chromatosome was more complex and gave digestion bands corresponding to four major positions of similar abundance all spaced by differences of 10 bp. The boundaries of these 172 bp and 207 bp chromatosomes were in the same "10 bp phase" as found for the 146 bp core particles suggesting identical rotational settings of the DNA for both types of particles. Clearly, the interactions of the outer regions of the chromatosome DNA with the globular domain of H1 and H5 influences the probability of some nucleosome core locations. Thus on the 5S rDNA 207 bp and 172 bp nucleosome positioning sequences the chromatosome is a real positioning entity defined by both the octamer/DNA interactions and chromatosome/H1 or H5 interactions.

8. Nucleosome Mobility

The observation of a cluster of nucleosome core positions flanking a dominant position on a strongly positioning DNA sequence suggests that the nucleosome cores may be able to exchange between positions in the cluster depending on ionic solution conditions and temperature. Two-dimensional gel electrophoresis has been used to investigate the effects of buffer and temperature on the positions of the nucleosome cores [85]. Mononucleosomes excised from the long 207_{18} chromatin by the restriction enzyme AvaI migrate as three bands in a nondenaturing nucleoprotein particle polyacrylamide gel. This AvaI digest was divided into two aliquots for two nucleoprotein polyacrylamide gels and run in parallel at 4°C. One gel was incubated at 37°C for one hour and the other gel kept at 4°C. Both gels were run in parallel in the second dimensions under the same conditions as the first dimension. For the gel which was incubated at 4°C between the first and second dimensions the three nucleosome core bands run on the diagonal as expected. For the gel that was incubated at 37°C between the first and second dimension each of the three bands from the first dimension redistributed into three bands in the second dimension. The two-dimensional gel assay experiment was repeated with an AvaI digest that had been incubated at 37°C for one hour in buffer prior to gel loading. When this gel was incubated between the first and second dimensions, the 207 bp nucleosomes migrated as a square of spots, indicating that each of the three original bands had again redistributed into three bands. Taken together these results show that 207 bp nucleosome cores excised from $(207)_{18}$ chromatin assembled with histone octamers have the ability to redistribute at 37°C. Nucleosome positioning therefore appears to have a dynamic character. The mobility was observed in low salt at 37°C but not at 4°C. At 4°C the mobility may be too slow to detect. The positions in the cluster have the same coiling of DNA around the nucleosome, but the boundaries of the nucleosome cores lie at 10 bp increments, i.e., the B-form DNA helical repeat, along the path of the DNA coil. Because the dominant position of the octamer on the repetitive sequence is flanked by weaker positions spaced by 10 bp intervals, it would appear that the positioning signal has at least some rotationally defined character to it, such as bendability. A purely translational signal (requiring alignment at defined points in the nucleosome) would not allow this type of fluctuation around an energetically favored nucleosome binding site. The mobility of nucleosome cores on the strongly positioning 5S gene DNA sequence suggests that mobility is a general property of H1 depleted chromatin. This was tested for long H1 depleted chromatin digested with restriction enzymes to avoid the trimming of overhanging DNA ends [86]. This was necessary because a range of mononucleosome lengths greater than 170 bp is required to distinguish between differently positioned nucleosomes in gel electrophoresis. The slowest migrating nucleosomes contained DNA lengths ranging from about 220 bp to greater than 300 bp.

The two dimensional gel electrophoresis was carried out as described above with gel stripes containing the bands of interest from the first dimension gel. Both first and second dimensions were at 4°C. Between the two runs the gel strips were incubated for 1 hour at 4°C for the control and for 37°C for the mobility experiment. The control incubation of mononucleosomes at 4°C shows the diagonal line expected of immobile nucleosomes whereas for the incubation at 37°C many but not all the mononucleosomes have become mobile on the underlying DNA sequences as shown by an off-diagonal "fan" of DNA intensity. There is a marked bias for an increase in electrophoretic velocity indicating a preference for nucleosome cores to occupy end positions. Thus at low ionic strengths the mobility of histone octamers is potentially a general behavior of a large proportion of native nucleosomes [86].

An important determinant of nucleosome positioning is the DNA anisotropy of flexibility required to accomodate the DNA tight bending around the histone octamer [76]. Because the binding affinity is the cummulative effect of many small bends positioning is often rotationally unique but translationally degenerate [87]. There are several examples of multiple positions spaced by 10 bp [88,89]. The mobility of a nucleosome core appears to depend on the sequences flanking its position. The histone octamer would be mobile if the DNA coil continued beyond the immediate location of the nucleosome core. Within this extended DNA coil the histone octamer would be able to jump through units of a B-form DNA helical repeat. This mobility would be limited by the same elements that act as boundaries to nucleosome positioning [reviewed in 73]. The significance of this nucleosome core mobility is that unlike the nucleosome sliding observed at higher non-physiological ionic strengths, which suppresses histone DNA interaction, all of the histone/DNA contacts are maintained at the low ionic strengths used in these experiments. It would appear that chromatin is a more dynamic structure than is widely assumed [86].

In vivo, local ionic conditions differ and a number of other factors such as histone modifications, binding of very lysine rich histones, DNA binding proteins or interactions with adjacent nucleosomes may suppress or enhance nucleosome mobility. Nucleosome cores may be fixed or free to move depending on functional requirements. Binding sites for transacting factors could become exposed to the factors involved in the control of gene expression.

9. Very Lysine Rich Histones Suppress Nucleosome Mobility

The hypothesis that nucleosome core mobility may be suppressed for functional requirements was tested by the binding of the very lysine rich histones H1 and H5. Very lysine rich histones have been identified as general repressors of transcription

[90,91]. Both H1 and H5 can be faithfully reconstituted into 5SrDNA chromatin [84]. Using the two dimensional gel electrophoresis, the mobility of histone octamers positioned on constructs of sea urchin 5SrDNA was shown to be efficiently suppressed by the binding of H1 or H5 to nucleosomes [92]. This implies that if nucleosome mobility is required for access to the underlying DNA sequences then the very lysine rich histones could function as general gene repressors through the immobilization of nucleosome cores. This function would be additional to the role of very lysine rich histones in stabilizing the "30 nm" supercoil of nucleosomes. Histone H5 was found to be a stronger inhibitor of nucleosome core mobility which correlates with its stronger binding to chromatin [92]. These results have been reproduced using the Xenopus 5SrDNA sequence repeat. Very lysine rich histones were found to inhibit nucleosome core mobility and repress transcription, demonstrating that stable states of gene repression can be established even at the nucleosome level [93]. All the above studies raise the possibility that during development the redistribution of very lysine rich histone subtypes provide another mechanism for suppressing the activities of genes not required at particular stages of development.

In vivo mechanisms for controlling nucleosome organization and functions have recently been identified that involve complex cofactors. Nucleosome rearrangements have been reported for H1 containing chromatin assembled in a cell free Drosophila embryo extract. On the addition of transcription factors it was found that nucleosomal arrays at the promoters of hsp70 and hsp26 genes were disrupted but only in the presence of ATP [94-96]. Thus this nucleosome rearrangement was dependent on ATP hydrolysis. A nucleosome remodeling cofactor (NURF), a 500 kDa protein complex, was identified in these extracts [97]. The NURF is thought to function with transcription factors in an ATP dependent manner to rearrange H1 containing nucleosomes positioned on the promoter regions prior to transcription. NURF is distinct from the previously reported SWI/SWF transcription activator comples [reviewed in 98]. The SWI/SNF complex is a 2 MDa protein complex that relieves the constraints of nucleosomal packaging of DNA prior to transcription through a mechanism that is different from that of NURF. These *in vivo* mechanisms for the disruption or reorganization of nucleosomal arrays are presumably to allow access of transacting factors to their specific DNA binding sites. They are more complex than nucleosome core mobility in the absence of histone H1 that has also been shown to allow access of transcription factors to their DNA binding sites [93].

10. Transcription Through Nucleosomes

Chromatin structure presents a barrier to the efficient transcription of active genes. Chromatin changes associated with active genes have been listed above and include the hyperacetylation of the histone octamer, particularly histones H3 and H4. Presumably the nucleosome remodeling cofactor [93-97] binds to the active or accessible promoter regions of specific genes to disrupt nucleosomes for transacting factors to bind and initiate gene expression. Nucleosome mobility [85,86,92] also provides a mechanism for transacting factors to bind to gene regulatory sequences [93] and initiate gene expression. Following gene activation RNA polymerases are faced with transcribing the DNA packaged into "active" nucleosomes. The bacteriophage SP6 RNA polymerase and eukaryotic RNA polymerases II and III have been shown to transcribe through one nucleosome [99,100] or short stretches of nucleosomes [101,102]. The problems with these studies were that the nucleosome templates were not fully defined as regards protein composition, nucleosome spacing and positioning or they contained only one or a few nucleosomes.

Our studies of transcription through nucleosomes [68,103-104] have used the well-characterized tandemly repeated nucleosome positioning sequence $(207)_{18}$ described above for the nucleosome positioning and mobility studies. The DNA construct $(207)_{18}$ was inserted between the T7 and SP6 transcription promotors of pGEM-32. Nucleosome cores were assembled on supercoiled, closed circular $pT(207)_{18}$ and double label experiments were performed to determine the effect of nucleosome cores on both the initiation and elongation of transcripts by T7 RNA polymerase. Both transcript initiation and elongation were inhibited, the extent of the inhibition being directly proportional to the number of nucleosome cores assembled on the $pT(207)_{18}$ DNA template. Continuous regularly spared linear arrays of nucleosome cores were obtained by digesting the assembled $pT(207)_{18}$ chromatin with DraI, for which a unique restriction site lies within the nucleosome positioning sequence of the 207 bp repeat. This site is protected from DraI by the formation of nucleosome cores. DraI will cut only naked DNA repeats and not the DNA repeats assembled into nucleosome cores. Thus the digestion of partially assembled $pT(207)_{18}$ with DraI gives the T7 promotor followed by continuous runs of assembled nucleosome cores of different lengths. *In vitro* transcription with T7 RNA polymerase gave an RNA ladder with a 207 nucleotide spacing demonstrating that transcription had proceeded through continuous arrays of positioned nucleosome cores. It was shown that nucleosome cores partially inhibit the elongation of transcripts by T7 RNA polymerase, while allowing passage of the polymerase through each nucleosome core at an upper efficiency of 85%. Hence, complete transcripts are produced with high efficiency from short nucleosomal templates, whereas the production of full length transcripts from large nucleosomal arrays is relatively ineffective. These results indicate that nucleosome cores have significant inhibitory effects *in vitro* not only in transcription initiation but also on transcription elongation and that special mechanisms probably exist to overcome those inhibitory effects *in vivo*. The question of whether histone octamer disociation was required for the process of transcription was addressed by the extensive cross-linking of the histone octamer prior to its assembly into $pT(207)_{18}$ chromatin [68]. Transcription studies of this heavily cross-linked $(207)_{18}$ assembled chromatin lacking H1 have demonstrated that there is no need for the disassociation of the histone octamer during elongation because transcription was not affected by the irreversible crosslinking of histones within the histone octamers.

As discussed above, histone H1 has a profound effect on the stability of nucleosomes and is required for the generation and stability of the 34 nm supercoil of nucleosomes. H1 has been strongly implicated in the formation of transcriptionally silent chromatin and has been demonstrated to be a general repressor of transcription initiation *in vitro* [106,107]. The effects of histone H1 on transcription of the $pT(207)_{18}$ assembled chromatin by T7 RNA polymerase have been investigated and both transcription initiation and elongation were found to be fully suppressed by histone H1. One interpretation of these results is that very lysine rich histones may inhibit transcription by stabilizing nucleosomal structures. This would suppress the mobility of nucleosome cores located on promotors, and thus reduce the accessibility of promoter regions to transacting factors and RNA polymerases. Further, stabilization of nucleosome cores may provide resistance to the progress of RNA polymerase during transcription elongation. This raises the possibility that histone H1 provides another mechanism for regulation of transcription of nucleosomal templates by the suppressor of nucleosome mobilities.

Summary

All aspects of DNA processing are proving to be dauntingly complex and involve not only polymerases and many regulatory factors but also the functions of nucleosomes. Our understanding of chromatin structure/function relationships has advanced very slowly. It is now clear, however, that an understanding of histone and nucleosome functions are integral to understanding DNA processing. Histones are no longer thought of as passive structural components of chromosomes but through the reversible chemical modifications of histones are clearly involved in chromosome functions. Our view of nucleosome cores as static structures has changed drastically. Nucleosome cores are capable of short-range mobility and can exchange between a cluster of positions spaced by intervals of 10 bp DNA. This nucleosome mobility clearly has functional significance because the potential now exists for sequence specific DNA binding proteins to influence the position of nucleosome cores and allow access to their DNA binding sites. The ability of the very lysine rich histones to suppress nucleosome core mobility implies that they can function at the individual nucleosome level in addition to their role in generating and stabilizing higher order chromatin structures.

Acknowledgements

Research support is acknowledged to EMB from the Department of Energy (DE-FG03-88ER 60673), and the National Institutes of Health (PHS GM26901), and to SP from the Wellcome Trust .

References

1. Saitoh, Y., Laemmli, U.K. (1993) From the Chromosomal Loops and the Scaffold to the Classic Bands of Metaphase Chromosomes, *Cold Spring Harbor Symposia on Quantitative Biology* **58**, 755-765.
2. van Holde, K.E. (1988) *Chromatin*, (A. Rich, ed.) Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo.
3. Bradbury, E.M. (1992) Reversible Histone Modifications and the Chromosome Cell Cycle, *Bioassays* **14**, 9-16.
4. Moss, T., Cary, P.D., Abercrombie, B.D., Crane-Robinson, C., and Bradbury, E.M. (1976) A pH-Dependent Interaction Between Histones H2A and H2B Involving Secondary and Tertiary Folding, *Eur. J. Biochem.*, **71**, 337-350.
5. Arents, G., Burlingame, R.W., Wang, B.C., Love, W.E., and Moudrianakis, E.N. (1991) The Nucleosomal Core Histone Octamer at 3.1 A resolution: A Tripartite Protein Assembly and a Left-Handed Superhelix, *Proc. Natl. Acad. Sci. USA* **88**, 10148-10152.
6. Moss, T., Cary, P.D., Crane-Robinson, C., and Bradbury, E.M. (1976) Physical Studies on the H3/H4 Histone Tetramer, *Biochemistry* **15**, 2261-2267.
7. Bradbury, E.M., Chapman, G.E., Danby, S.E., Hartman, P.G., and Riches, P.L. (1976) Studies on the Role and Mode of Operation of the Very-Lysine-Rich Histone H1 (F1) in Eukaryote Chromatin. The Properties of the N-Terminal and C-Terminal Halves of Histone H1, *Eur. J. Biochem.* **57**, 521-528.
8. Hartman, P.G., Chapman, G.E., Moss, T., and Bradbury, E.M. (1977) Studies on the Role and Mode of Operation of the Very-Lysine-Rich Histone H1 in Eukaryote Chromatin, *Eur. J. Biochem.*, **77**, 456.

9. Chapman, G.F., Hartman, P.G., Cary, P.D., Bradbury, E.M., and Lee, D.R. (1978) A Nuclear Magnetic Resonance Study of the Globular Structure of the H1 Histone, *Eur. J. Biochem.*, **86**, 35.
10. Johnson, E.M., and Allfrey, V.G. (1978) In: *Biochemistry Actions of Hormones*", Vol. 5 (G. Litwac. ed.) Academic Press, New York.
11. Matthews, H.R. and Waterborg, J. (1985) in "The Enzymology of Post Translational Modifications of Proteins", Vol. 2, pp. 125-185.
12. Yasuda, H., Mueller, R.D., and Bradbury, E.M. (1986) Molecular Regulation of Nuclear Events in "Mitosis and Meiosis", (Schlegel, R.A., Halleck, M.S., and Rao, P.N., eds.), Academic Press, New York pp. 391-361.
13. Busch, H. and Goldknoph, I.L. (1981) Ubiquitin-Protein Conjugates, *Mol. Cell Biol.* **40**, 173-187.
14. Christensen, M.E. and Dixon, G.H. (1982) Hyperacetylation of Histone-H4 Correlates with the Terminal, Transcriptionally Inactive Stages of Spermatogenesis in Rainbow Trout, *Dev. Biol.* **93**, 404-415.
15. Goldknoph, I.L., Wilson, G., Ballard, N.R., and Busch, H. (1980) Chromatin Conjugate Protein A24 is Cleaved and Ubiquitin is Lost During Chicken Erythropoiesis, *J. Biol. Chem.* **255**, 10555-10558.
16. Levinger, L. and Varshavsky, A. (1982) Selective Arrangement of Ubiquitinated and D1 Protein-Containing Nucleosomes Within the Drosophila Genome, *Cell* **28**, 375-385.
17. Matsui, S.L., Seon, B.K., and Sandberg, A.A. (1979) Disappearance of a Structural Chromatin Protein A24 in Mitosis: Implications for Molecular Basis of Chromatin Condensation, *Proc. Natl. Acad. Sci., USA* **76**, 6386-6390.
18. Mueller, R.D., Yasuda, H., Hatch, C.L., Bonner, W.M., and Bradbury, E.M. (1985) Phosphorylation of Histone H1 Through the Cell Cycle of *Physarum polycephalum*, *J. Biol. Chem.* **260**, 5147-5153.
19. Bradbury, E.M., Inglis, R.J., and Matthews, H.R. (1974) Control of Cell Division by Very Lysine-Rich Histone F1 Phosphorylation, *Nature* **241**, 257-261.
20. Bradbury, E.M., Inglis, R.J., Matthews, H.R., and Langan, T.A. (1974) Molecular Basis of Mitotic Cell Division in Eukaryotes, *Nature* **249**, 553.
21. Gurley, L.R., D'Anna, J.A., Halleck, M.S., Barham, S.S., Walters, R.A., Jett, J.H., and Tobey, R.A. (1981) In: *Cold Spring Harbor Conferences on Cell Proliferation* **8**, 1073-1093.
22. McGhee, J.D. and Felsenfeld, G. (1980) Nucleosome Structure, *Ann. Rev. Biochem.* **40**, 1115-1156.
23. Bradbury, E.M. and Matthews, H.R. (1982) Chromatin Structure, Histone Modifications in the Cell Cycle, in "Cell Growth" (Nicolini, C. ed.) Plenum Press, NY, pp. 411-454.
24. Korberg, R.D. and Klug, A. (1981) The Nucleosome, *Sci. American* **244**, 48-60.
25. Simpson, R.T. (1978) Structure of the Chromosome, a Chromatin Particle Containing 160 base pairs of DNA and all the Histones, *Biochemistry* 17, 5524-5531.
26. Bradbury, E.M., Baldwin, J.P., Carpenter, B.G., Hjelm, R.P., Hancock, R., and Ibel, K. (1975) Neutron-Scattering Studies of Chromatin, *Brookhaven Symp. Biol. 27, IV* (Schoenborn, B.P., ed.) pp. 97-116.
27. Pardon, J.F., Worcester, D.C., Wooley, J.C., Tatchell, K., van Holde, K.E., and Richards, B.M. (1975) Low-Angle Neutron Scattering from Chromatin Subunit Particles, *Nucl. Acids Res.* **2**, 2163-2176.
28. Suau, P., Kneale, G.G., Braddock, G.W., Baldwin, J.P., and Bradbury, E.M. (1977) A Low Resolution Model for the Chromatin Core Particle by Neutron Scattering, *Nucleic Acids Research* 4, 3769-3786.
29. Hjelm, R.P., Kneale, G.G., Suau, P., Baldwin, J.P., and Bradbury, E.M. (1977) Small Angle Neutron Scattering Studies of Chromatin Subunits in Solution, *Cell* **10**, 139-151.
30. Richards, B.M., Pardon, J., Lilley, D.M.J., Cotter, P., and Wooley, J. (1977) Sub-Structure of Nucleosomes, *Cell Biol. Int. Rep.* **1**, 107-116.
31. Braddock, G.W., Baldwin, J.P., and Bradbury, E.M. (1981) Neutron Scattering Studies of the Structure of the Chromatin Core Particle, *Biopolymers* **20**, 327-343.
32. Finch, J.T., Lutter, L.C., Rhodes, D., Brown, R.S., Rushton, B., Levitt, M., and Klug, A. (1977) Structure of Nucleosome Core Particles of Chromatin, *Nature* **269**, 29-36.
33. Finch, J.T., Brown, R.S., Rhodes, D., Richmond, T., Rushton, B., Lutter, L.C., and Klug, A. (1981) X-Ray-Diffraction Study of a New Crystal Form of the Nucleosome Core Showing Higher Resolution, *J. Mol. Biol.* **145**, 757-769.
34. Bentley, C.A., Finch, J.T., and Lewit-Bentley, A. (1981) Neutron Diffraction Studies on Crystals of Nucleosome Cores Using Contrast Variation, *J. Mol. Biol.* **145**, 771-784.
35. Richmond, T.J., Klug, A., Finch, J.T., and Lutter, L.C. (1981) In: *Proc. 2nd SUNY Conversation in Biomolecular Stereodynamics* (Sarma, R.H. ed.) Vol. II., Adenine Press, NY, pp. 109-123.
36. Richmond, T.J., Finch, J.T., Rushton, B., Rhodes, D., and Klug, A. (1984) Structure of the Nucleosome Core Particle at 7A Resolution, *Nature* **311**, 532-537.
37. Wood, M.J., Yau, P.M., Imai, B.S., Goldberg, M.W., Lambert, S.J., Fowler, A.L., Baldwin, J.P., Godfrey, J., Moudrianakis, E.N., Ibel, K., May, R.P., Koch, M., and Bradbury (1991) Neutron and X-Ray Scatter Studies of the Histone Octamer and Amino and Carboxyl Domain Trimmed Octamers, *J. Biol. Chem.* **266**, 5696-5702.
38. Schroth, G.P., Yau, P.M., Imai, B.S., Gatewood, J.M., and Bradbury, E.M. (1990) A NMR Study of Mobility in the Histone Octamer, *FEBS Lett.* **268**, 117-120.

39. Imai, B.S., Yau, P.M., Baldwin, J.P., Ibel, K., May, R.P., and Bradbury, E.M. (1986) Hyperacetylation of Core Histones Does not Cause Unfolding of Nucleosomes: Neutron Scatter Data Accords with Disc Structure of the Nucleosome, *J. Biol. Chem.* **261**, 8784-8792.

40. Usachenko, S.I., Barykin, S.G., Gavin, I.M., and Bradbury, E.M. (1994) Rearrangement of the Histone H2A C-Terminal Domain in the Nucleosome, *Proc. Natl. Acad. Sci. USA*, **91**, 6845-6849.

41. Thoma, F., Koller, T.H., and Klug, A. (1979) Involvement of Histone H1 in the Organization of the Nucleosome and of the Salt-Dependent Superstructures of Chromatin, *J. Cell Biol.* **83**, 403-427.

42. Crane-Robinson, C., Bohm, L., Puigdomenech, P., Cary, P.D., Hartman, P.G., and Bradbury, E.M. (1980) Structural Domains in Histones, *FEBS DANA-Recombination Interactions and Repair*, Pergamon Press, Oxford and New York.

43. Allan, J., Hartman, P.G., Crane-Robinson, C., and Aviles, F.X. (1980) The Structure of Histone H1 and its Location in Chromatin, *Nature* **288**, 675-679.

44. Suau, P., Bradbury, E.M., and Baldwin, J.P. (1979) Higher-Order Structures of Chromatin in Solution, *Eur. J. Biochem.* **97**, 593-602.

45. Carpenter, B.G., Baldwin, J.P., Bradbury, E.M., and Ibel, K. (1976) Organization of Subunits in Chromatin, *Nucl. Acids Res.* **3**, 1739-1746.

46. Finch, J.T. and Klug, A. (1976) Solenoidal Model for Superstructure in Chromatin, *Proc. Natl. Acad. Sci. USA* 73, 1897.

47. Widom, J. and Klug, A. (1985) Structure of the 300A Chromatin Filament: X-Ray Diffraction from Oriented Samples, *Cell* **43**, 207-213.

48. Worcel, A., Strongatz, S., and Riley, D. (1981) Structure of the 300A Chromatin Filament: X-Ray-Diffraction from Oriented Samples, *Proc. Natl. Acad. Sci. USA* **78**, 1461-1465.

49. Woodcock, C.L.F., Frado, L.L.Y., and Rattner, J.B. (1984) The Higher-Order Structure of Chromatin: Evidence for a Helical Ribbon Arrangement, *J. Cell Biol.* **99**, 42-52.

50. Bednar, J. Horowitz, R.A., Dubochet, J., and Woodcock, C.L. (1995) Compaction: 3-Dimensional Structural Information from Cryoelectron Microscopy, *J. Cell Biol.* **131**, 1365-1376.

51. Williams, S.P., Athey, B.D., Muglia, L.J., Scheppe, R.S., Gough, A.H., and Langmore, J.P. (1986) Chromatin Fibers are Left-Handed Double Helices with Diameter and Mass per Unit Length that Depend on Linker Length, *Biophys. J.* **49**, 233-248.

52. Staynov, D.Z. (1983) Possible nucleosome arrangements in the higher-order structure of chromatin. *Int. J. Biol. Macromol.* **5**, 3-9.

53. Graziano, V., Gerchman, V., Schneider, D.K., and Ramakrishnan, V. (1994) Histone H1 is Located in the Interior of the Chromatin 30-nm Filament, *Nature* **368**, 351-354.

54. Yang, G., Leuba, S.H., Bustamante, C., Zlatanova, J., and van Holde, K. (1994) Linker DNA Accessibility in Chromatin Fibers of Different Conformations: A Re-Evaluation, *SPIE Proc.* **2384**, 13-21.

55. Leuba, S.H., Yang, G., Robert, C., Samori, B., van Holde, K., Zlatanova, J., and Bustamante, C. (1994) Three-Dimensional Structure of Extended Chromatin Fibers as Revealed by Tapping-Mode Scanning Force Microscopy, *Proc. Natl. Acad. Sci.* **91**, 11621-11625.

56. Zlatanova, J., Leuba, S.H., Yang, G., Bustamante, C., and van Holde, K. (1994) Linker DNA Accessibility in Chromatin Fibers of Different Conformations: A re-Evaluation, *Proc. Natl. Acad. Sci.* **91**, 5277-5280.

57. Leuba, S.H., Zlatanova, J., and van Holde, K. (1994) On the Location of Linker DNA in the Chromatin Fiber. Studies with Immobilized and Soluble Micrococcal Nuclease, *J. Mol. Biol.* **235**, 871-880.

58. Allen, M.J. (1995) Ph.D. Dissertation "Application of Atomic Force Microscopy to In Vitro and In Situ Investigations of Somatic and Sperm Chromaton Structure", University of California-Davis.

59. Foe, V.E. (1977) Modulation of Ribosomal RNA Synthesis in Oncopeltus Fasciatus: An Electron Microscopic Study of the Relationship Between Changes in Chromatin Structure and Transcriptional Activity, *Cold Spring Harbor Symp. Quant. Biol.* **42**, 723-740.

60. Andersson, K., Mahr, R., Bjorkroth, B., and Daneholt, B. (1982) Rapid Reformation of the Thick Chromosome Fiber Upon Completion of RNA Synthesis at the Balbiani Ring Genes in Chironomus Tentans, *Chromosoma* **87**, 33-84.

61. Daneholt, B. (1982) In: Insect Ultrastructure 1 (King and Akai, eds.) Plenum Publishing Corporation, pp. 382-401.

62. Olins, A.L., Olins, D.E., and Lezzi, M. (1982) Ultrastructural studies of Chironomus salivary-gland cells in different states of Balbiani ring activity. *Eur. J. Cell Biol..* **27**, 161-169.

63. Olins, D.E., Olins, A.L., Levy, H.A., Durfee, R.C., Margie, S.M., Timnel, E.P., and Dover, S.D. (1983) Electron-Microscope Tomography: Transcription in 3-Dimensions, *Science* **220**, 498-500.

64. Levy-Wilson, B. and Dixon, G.H. (1979) Limited Action of Micrococcal Nuclease on Trout Testis Nucleo Generates Two Mononucleosome Subsets Enriched in Transcribed DNA Sequences, *Proc. Natl. Acad. Sci. USA* **76**, 1682-1686.

65. Yasuda, H., Mueller, R.D., Logan, K.A., and Bradbury, E.M. (1986) Histone H1 in Physarum polycephalum; Its High Level in the Plasmodial State Increases in Amount and Phosphorylation in the Sclerotial Stage, *J. Biol. Chem.* **261**, 2349-2354.

66. Baer, B.W. and Rhodes, D. (1983) Eukaryotic RNA Polymerase-II Binds to Nucleosome Cores from Transcribed Genes, *Nature* **301**, 482-488.

67. Weisbrod, S. and Weintraub, H. (1981) Isolation of Actively Transcribed Nucleosomes Using Immobilized HMG 14 and 17 and an Analysis of Alpha-Globin Chromatin, *Cell* **23**, 391-400.

68. O'Neill, T.E., Smith, J.G., and Bradbury, E.M. (1993) Histone Octamer Dissociation is not Required for Transcript Elongation Through Arrays of Nucleosome Cores by Phage T7 RNA Polymerase in vitro, *Proc. Natl. Acad. Sci. USA* **90**, 6203-6207.

69. Norton, V.G., Imai, B.S., Yau, P.M., and Bradbury, E.M. (1989) Histone Acetylation Reduces Nucleosome Core Particle Linking Number Change, *Cell* **57**, 449-457.

70. Norton, V.G., Marvin, K.W., Yau, P.M., and Bradbury, E.M. (1990) Nucleosome Linking Number Change Controlled by Acetylation of Histones H3 and H4, *J. Biol. Chem.* **265**, 19,848-19,852.

71. Cao, Y., Yau, P., and Bradbury, E.M., manuscript in preparation.

72. West, M.H.P. and Bonner, W.M. (1980) Histone 2B can be Modified by the Attachment of Ubiquitin, *Nucl. Acids Res.* **8**, 4671.

73. Simpson, R.T. (1986) Nucleosome Positioning in vivo and in vitro, *Bioessays* **4**, 172-176.

74. Trifonov, E.N. (1980) Helical Model of Nucleosome Core, *Nucl. Acids Res.* **8**, 4041-4053.

75. Drew, H.R. and Travers, A.A. (1985) DNA Bending and its Relation to Nucleosome Positioning, *J. Mol. Biol.* **186**, 773-790.

76. Satchwell, S.C., Drew, H.R. and Travers, A.A. (1986) Sequence Periodicities in Chicken Nucleosome Core DNA, *J. Mol. Biol.* **191**, 659-675.

77. Rhodes, D. (1979) Nucleosome Cores Reconstituted from poly(dA-dT) and the Octamer of Histones, *Nucl. Acids Res.* **6**, 1805-1816.

78. Prunell, A. (1982) Nucleosome reconstitution on Plasmid-Inserted Poly(dA)•poly(dT), *EMBO J.* **1**, 173-179.

79. Neubauer, B., Linxweiler, W., and Horz, W. (1986) DNA Engineering shows that Nucleosome Phasing on the African-Green Monkey Alpha-Satellite is the Result of Multiple Additive Histon, *J. Mol. Biol.* **190**, 639-645.

80. Thoma, F. and Zatchei, M. (1988) Chromatin Folding Modulates Nucleosome Positioning in Yeast Minichromosomes, *Cell* **55**, 945-953.

81. Stein, A. and Mitchell, M. (1988) Generation of Different Nucleosome Spacing Periodicities in vitro: Possible Origin of Cell Type Specifficity, *J. Mol. Biol.* **203**, 1029-1043.

82. Simpson, R.T. and Stafford, D.W. (1983) Structural Features of a Phased Nucleosome Core Particle, *Proc. Natl. Acad. Sci., USA* **50**, 51-55.

83. Simpson, R.T., Thoma, F., and Brubaker, J.M. (1985) Chromatin Reconstituted from Tandemly Repeated Cloned DNA Fragments and Core Histones: A Model System for Study of Higher-Order St, *Cell* **42**, 799-808.

84. Meersseman, G., Pennings, S., and Bradbury, E.M. (1991) Chromatosome Positioning on Assembled Long Chromatin: Linker Histones Affect Nucleosome Placement on 5S rDNA, *J. Mol. Biol.* **220**, 89-100.

85. Pennings, S., Meersseman, G., and Bradbury, E.M. (1991) Mobility of Positioned Nucleosomes on 5S rDNA, *J. Mol. Biol.* **220**, 101-110.

86. Meersseman, G., Pennings, S., and Bradbury, E.M. (1992) Mobile Nucleosomes - A General Behavior, *EMBO J.* **11**, 2951-2959.

87. Shrader, T.E. and Crothers, D.M. (1989) Artificial Nucleosome Positioning Sequences, *Proc. Natl. Acad. Sci., USA*, **86**, 7418-7422.

88. Lowman, H. and Bina, M. (1990) Correlation Between Dinucleotide Periodicities and Nucleosome Positioning on Mouse Satellite DNA, *Biopolymers*, **30**, 861-876.

89. Simpson, R.T. (1991) Nucleosome Positioning: Occurrence, Mechanisms, and Functional Consequences, *Prog. Nucl. Acids Res. Mol. Biol.*, **40**, 143-184.

90. Wolffe, A.P. (1989) Dominant and Specific Repression of Xenopus Oocyte 5S RNA Genes and Satellite I DNA by Histone H1, *EMBO J.*, **8**, 527-537.

91. Laybourn, P.J. and Kadonaga, J.T. (1991) Role of Nucleosomal Cores and Histone H1 in the Regulation of Transcription by RNA Polymerase II, *Science*, **254**, 238-245.

92. Pennings, S., Meersseman, G., and Bradbury, E.M. (1994) Linker Histones H1 and H5 Prevent the Mobility of Positioned Nucleosomes, *Proc. Natl. Acad. Sci. USA*, **91**, 10275-10279.

93. Ura, K., Hayes, J.J., and Wolffe, A.P. (1995) A Positive Role for Nucleosome Mobility in the Transcriptional Activity of Chromatin Templates: Restriction by Linker Histones, *EMBO J.*, **14**, 3725-3765.

94. Tsukiyama, T., Becker, P.B., and Wu, C. (1994) ATP-Dependent Nucleosome Disruption at a Heat-Shock Promoter Mediated by Binding of GAGA Transcription Factor, *Nature*, **367**, 525-532.

95. Wall, G., Varga-Weisz, zp.D., Sandaltzopoulos, R., and Becker, P.B. (1995) Chromatin Remodeling by GAGA Factor and Heat Shock Factor at the Hypersensitive Drosophila hsp26 Promoter in vitro, *EMBO J.*, 14, 1727-1736.

96. Pazin, M.J., Kamakaka, R.T., and Kadonaga, J.T. (1994) ATP-Dependent Nucleosome Reconfiguration and Transcriptional Activation from Preassembled Chromatin Templates, *Science*, **266**, 2007-2011.

97. Tsukiyama, T. and Wu, C. (1995) Purification and Properties of an ATP-Dependent Nucleosome Remodeling Factor, *Cell*, **83**, 1011-1020.

98. Peterson, C.L. and Tamkun, J.W. (1995) The SWI-SNF Complex: A Chromatin Remodeling Machine, *Trends. Biochem. Sci.*, **20**, 143-146.

99. Lorch, Y., LaPointe, J.W., and Kornberg, R.D. (1987) Nucleosomes Inhibit the Initiation of Transcription but Allow Chain Elongation with the Displacement of Histones, *Cell*, **49**, 203-210.

100. Losa, R. and Brown, D.D. (1987) A Bacteriophage RNA Polymerase Transcribes *in vitro* Through a Nucleosome Core Without Displacing it, *Cell*, **50**, 801-808.

101. Morse, R.H. (1989) Nucleosomes Inhibit Both Transcriptional Initiation and Elongation by RNA Polymerase III in vitro, *EMBO J.*, **8**, 2343-2351.

102. Felts, S.J., Weil, P.A., and Chalkley, R. (1990) Transcription Factor Requirements for in vitro Formation of Transcriptionally Competent 5S rDNA Gene Chromatin, Mol. Cell. Biol., **10**, 2390-2401.

103. O'Neill, T.E., Roberge, M., and Bradbury, E.M. (1992) Nucleosome Arrays Inhibit both Initiation and Elongation of Transcripts by T7 RNA Polymerase, *J. Mol. Biol.*, **223**, 67-78.

104. O'Neill, T.E., Pennings, S., Meersseman, G., and Bradbury, E.M. (1995) Deposition of Histone H1 Onto Reconstituted Nucleosome Arrays Inhibits Both Initiation and Elongation of Transcripts by T7 RNA Polymerase, *Nucleic Acids Res.*, **23**, 1075-1082.

CHAPTER 41

Structural Studies on the Unstable Triplet Repeats

S. V. SANTHANA MARIAPPAN, XIAN CHEN, AND PAOLO CATASTI

Life Sciences Division, LS-2, MS 880, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

E. MORTON BRADBURY

Life Sciences Division, LS-2, MS 880, Los Alamos National Laboratory, Los Alamos, New Mexico 87545; and Department of Biological Chemistry, School of Medicine, University of California at Davis, Davis, California 95616

GOUTAM GUPTA

Theoretical Biology and Biophysics, T-10, MS-K710, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

6

I. INTRODUCTION

The expansions of triplet DNA repeats define a new type of mutation in the human genome [1, 2]. The genetic instability due to triplet expansion is associated with many genetically inherited neurological disorders [3–6]. Figures 41-1 and 41-2 show different triplet repeats in different genetic disorders and their associated genes. These triplet repeats are located most frequently inside the noncoding regions (upstream, downstream, intron) of genes and less frequently inside the coding regions. All of the unstable triplet repeats identified so far are GC-rich of the form (CXG)/(CX'G), where X and X' are complementary to each other: for example, GCC/GCC in the fragile X syndrome (FraX) (Fig. 41-1A, [3]), CTG in myotonic dystrophy (DM) (Fig. 41-1B, [4]), and CAG in Huntington's disease (HD) (Fig. 41-1C, [5]) belong to this category. The only exception is the GAA/TTC repeat associated in Friedreich's ataxia (FRDA) (Fig. 41-2, [6]).

Expansions of DNA triplet repeats (or any repeat) may involve one of the three mechanisms [7–9]: (i) unequal crossover, i.e., crossover between tracts misaligned by an integral number of repeats; (ii) slippage during DNA replication, i.e., during replication, the primer and template strand transiently dissociate and the slippage of the strands can then result in either expansion or deletion; and (iii) misalignment followed by excision repair, i.e., a mutagenic alternative DNA secondary structure may be formed during or after replication, which is excised and repaired leading to either deletion or expansion. Although these mechanisms have the potential to explain genetic instabilities associated with disease-related triplets, the exact mechanism has yet to be proven. However, it is generally accepted that the key step in triplet expansion is the formation of non-Watson–Crick DNA structures during replication or crossover recombination [10, 11]. Hence, the identification and characterization of these unusual DNA structures formed by triplet repeats are of crucial importance in understanding the mechanism of expansion.

Unusual DNA structures include hairpins, cruciforms and junctions, and intramolecular triplexes and tetraplexes. Once they are identified and completely characterized, it is necessary to determine whether these unusual structures are preferentially stabilized in longer repeats. Finally, it has to be determined by *in vitro* and *in vivo* assays whether these structures can, indeed, cause DNA slippage structures during replication.

In this chapter, we provide experimental evidence in support of the hypothesis that unusual DNA structures are responsible for the expansions of the disease-related triplets and their associated genetic instabilities. For this, we have performed the following experiments: (i) we have characterized the unusual DNA structures by nondenaturing gel electrophoresis of short and long

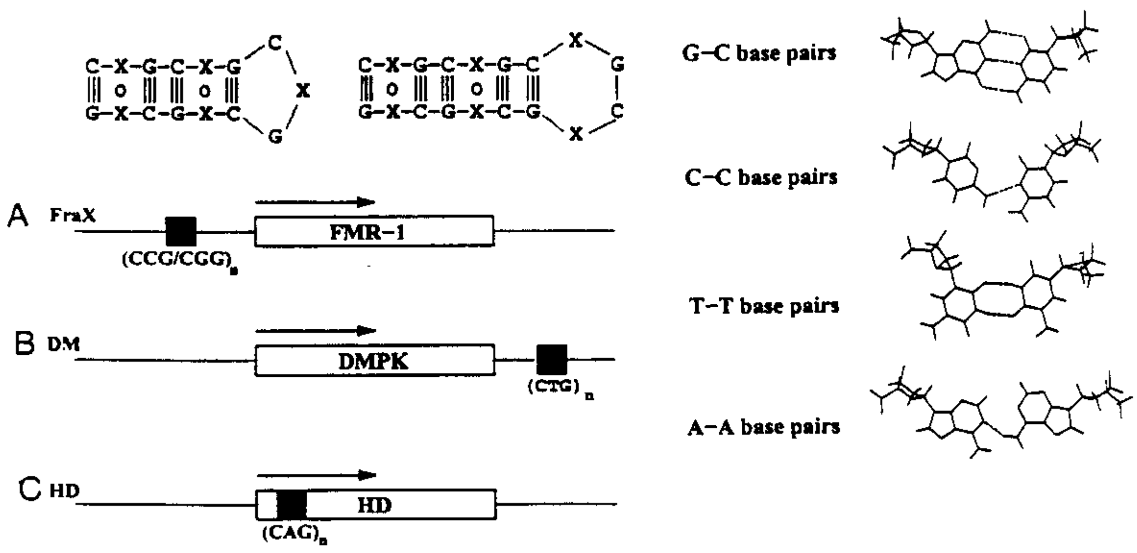

FIGURE 41-1 GC-rich triplet repeats and their locations with respect to their associated genes: (A) CCG upstream of the FMR1 gene, (B) CTG downstream of the DMPK gene, and (C) CAG inside the exon of the HD gene. Hairpin structures formed by the $(CXG)_n$ triplet repeats have three-nucleotide loops for odd repeat numbers and four-nucleotide loops for even repeat numbers. Note that mismatches have orientations similar to the flanking G-C pairs and therefore can be easily embedded into the structure.

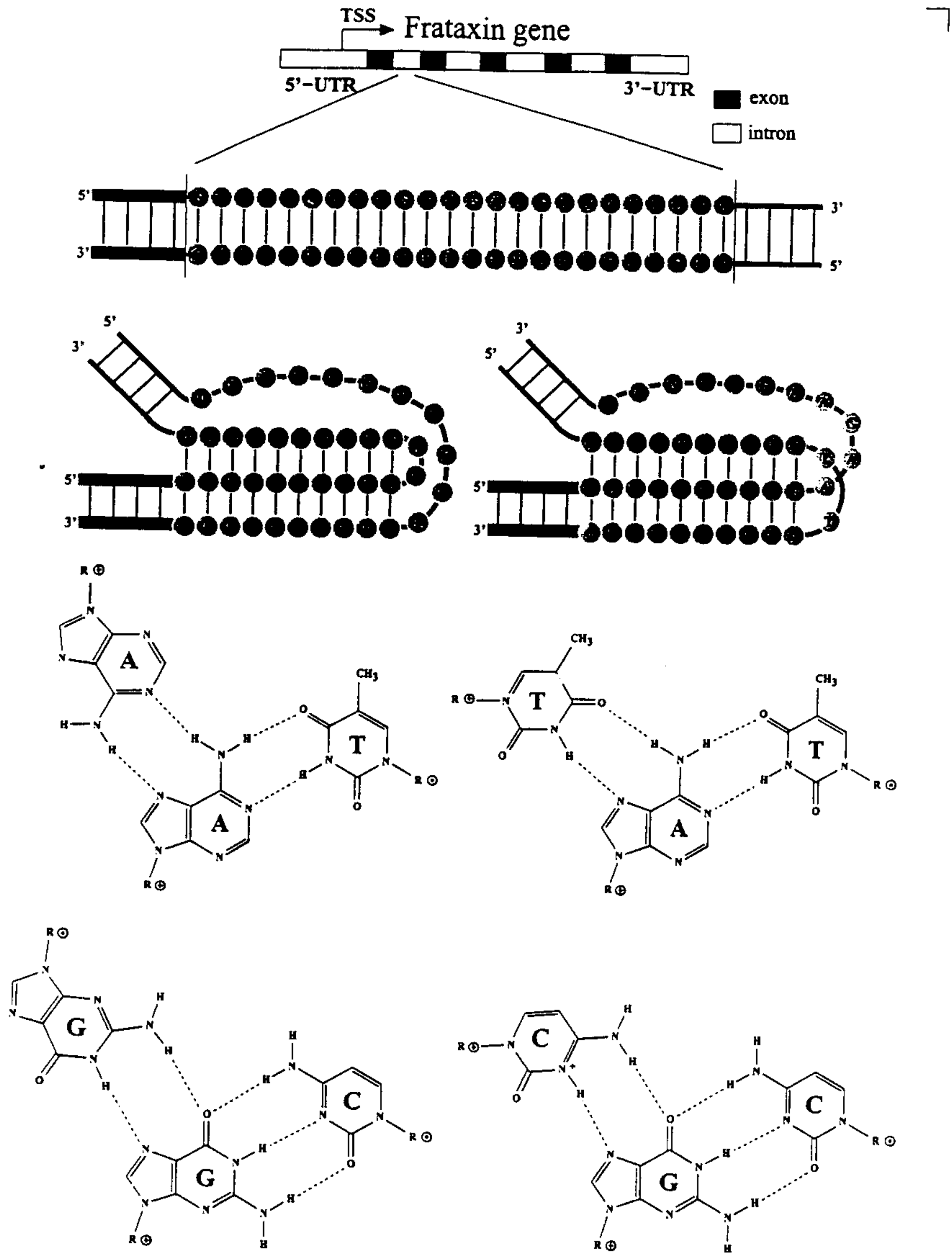

FIGURE 41-2 Possible triplexes formed by the GAA/TTC repeats inside the first intron of the frataxin gene. The folding of the TTC strand leads to a triplex with C⁺ · G · C and T · A · T triads and, therefore, this structure is more stable at acidic pH. The folding of the GAA strand leads to a triplex with G · G · C and A · A · T triads.

triplet repeats, (ii) we have determined the high-resolution structures of short triplet repeats by homonuclear (^{1}H-^{1}H) and hetero (^{15}N-^{1}H) NMR spectroscopy, and (iii) we have detected these unusual DNA structures by an *in vitro* replication assay using various DNA polymerases and their accessory proteins. Two classes of unusual DNA structures are discussed, i.e., the (CXG) triplet repeats that tend to form hairpin structures (Figs. 41-1A–41-1C) and the GAA/TTC triplet repeats that tend to form triplexes (Fig. 41-2). For clarity, the structural studies are discussed separately for each triplet repeat.

Many structural investigations on triplet repeats have been carried out by us and other laboratories [12–48]. The work by other laboratories essentially falls into five distinct categories: (i) detailed thermodynamic analyses that correlate the stabilities of various (CXG) hairpins with their repeat lengths [34], (ii) a combination of low-resolution NMR and thermodynamic and genetic analyses that provides an elegant explanation of how the genetic instability in the triplet repeats results from unusual DNA structures and not from a deficiency in the mismatch repair system as found for the dinucleotide instability in colon cancers [11, 24–27], (iii) gel mobility studies on long tracts of triplet repeats that show the presence of multiple slipped structures [18, 37], (iv) reconstitution experiments that show differential abilities of different triplet tracts to form nucleosomes [e.g., $(CTG)_n$ forms better positioned and more stable nucleosomes than $(GCC)_n$] [29, 30, 40, 41, 106], and (v) *in vitro* and *in vivo* replication and mismatch repair assays with long tracts of triplet repeats that demonstrate the presence of slippage structures [28, 31, 39, 42–45, 47, 48]. These data are extremely important since they provide clear indication that the disease-related triplet repeats can form stable unusual DNA structures which may also be present during their replication. As explained below, our efforts [13, 32, 33, 46, 105] complement the literature by providing high-resolution NMR structural details of the hairpins or triplexes formed by the triplet repeats (see Figs. 41-1 and 41-2). Also our replication assay clearly distinguishes the hairpin-induced slippage structures from the triplex-induced slippage structures.

Homonuclear (^{1}H-^{1}H) and (^{15}N-^{1}H) heteronuclear NMR spectroscopy give the following structural details of the hairpins and triplexes: (i) exact base pairing schemes, (ii) precise chainfolding, and (iii) interactions involving nucleotides in the loop and in the stem. These structural details enable us to explain how single interruptions in the GCC (or CAG) repeat or in the GAA/TTC repeat confer genetic stability by lowering the stability of the hairpin or the triplex, respectively. In addition, the local structure of the CpG sites in the stem of the $(GCC)_n$ hairpin helps explain why this hairpin is a better substrate for methylation by the human methyltransferase than either the Watson–Crick duplex, $(GCC)_n \cdot (GGC)_n$, or the $(GGC)_n$ hairpin.

For our *in vitro* replication assay we have selected repeat lengths, $n < 40$, such that only one copy or a few copies of the alternative structures are formed in the template. Thus, the nature of the replication product in the *in vitro* assay truly reflects the nature of the alternative structure. For example, the formation of a hairpin in the template causes a replication bypass and a reduction in the length of the replication product that corresponds to length of the hairpin. On the other hand, if a triplex is formed in the template replication arrest occurs in the middle of the repeat and the point of arrest indicates the length of the triplex. Note that the CXG repeats tend to form hairpin-induced slippage structures whereas the GAA/TTC repeats tend to form triplex-induced slippage structures. However, for $n \gg 40$, the presence of multiple copies of the alternative structures may lead to a higher order template structure resulting in replication arrests irrespective of whether the individual units are hairpins or triplexes. Apart from distinguishing a hairpin from a triplex, we can also determine the stability of a hairpin or a triplex formed by triplet repeats in the template as a function of its length. Note that in our replication assay the presence of the growing complementary strand, DNA polymerase, and structure-destabilizing proteins such as single-strand binding proteins and other ATP-dependent accessory proteins tend to destabilize a hairpin or a triplex. Although not proven directly, it is reasonable to predict that the same slippage structures will be present either in the template or in the growing chain during replication of these triplet repeats *in vivo*. Hairpin- or triplex-induced DNA slippage structures in the template should lead to deletion whereas the slippage in the growing chain should cause expansion.

II. FRAGILE X TRIPLET REPEAT, $(GCC)_n/(GGC)_n$: STRUCTURAL BASIS FOR EXPANSION AND CpG METHYLATION

The fragile X syndrome is the most common X-linked mental disorder, accounting for 50% of all reported cases [49]. The fragile X syndrome was originally identified by the presence of a microscopic gap or constriction, termed a fragile site, in the long arm of the X chromosome at Xq27.3 in affected individuals by culturing these cells under conditions of folate deficiency [50]. Recently, the gene associated with fragile X syndrome has been isolated and is called FMR1 (fragile X mental retardation-1). The FMR1 gene shows three important features in individuals affected with fragile X syndrome [49–53]: (i) the 5′ untranslated region of the gene contains the triplet repeats of (GGC/GCC) which are massively expanded, (ii) the CpG islands inside the triplet repeat are hypermethylated, and (iii) the expression of the FMR1 gene is either considerably reduced or completely suppressed. The expansion of GGC/GCC triplet repeat and the associated hypermethylation are probably the cause of the suppression of the FMR1 gene and the fragile sites in the X chromosomes. The number of GGC/GCC repeats in normal phenotypes varies between 6 and 53 with 29 occurring most frequently. Premutation alleles have between 54 and 200 repeats, whereas full mutation alleles have more than 200 repeats (Fig. 41-1A). The risk of expansion to the full mutation is depéndent on the size of the premutation allele. If the repeat number is small (50–70 copies) then the risk is low, and if the number of copies is high (>90) the risk is close to 100%. The risk of expansion depends also on the purity of the repeat; a single base interruption [e.g., $(GCC)_9 \cdot TCC \cdot (GCC)_9$] in the original repeat sequence reduces the risk [54].

In this section, we first show that the individual single strands of the fragile X repeat, i.e., $(GCC)_n$ and $(GGC)_n$, can form hairpin structures. We then describe the results of an *in vitro* replication assay that demonstrates the presence of hairpin-induced slippage structures. We also show by a methylation assay why the $(GCC)_n$ hairpin-induced slippage structure is an excellent substrate for the human methyltransferase, the enzyme that methylates the Cs at the CpG sites. Finally, we propose a structure-based mechanism of how expansion and hypermethylation can cause suppression of the FMR1 gene and the onset and progression of the fragile X syndrome.

A. Structural Characterization of $(GCC)_n$ by Gel Electrophoresis

Theoretically, at neutral pH, the two individual strands of the fragile X repeat can form either a mismatched homoduplex or a monomeric hairpin. The homoduplex and the stem of the hairpin of the $(GCC)_n$ strands involve Watson–Crick G · C pairs and mismatched C · C pairs. Note that the hairpin of $(GCC)_n$ should have half the length but approximately the same cross-section as the homoduplex (i.e., $[(GCC)_n]_2$) or the Watson–Crick duplex (i.e., $(GCC)_n \cdot (GGC)_n]$. Therefore, the duplex is expected to show about half the gel mobility of the corresponding hairpin. The electrophoretic mobilities of $(GCC)_n$ in a nondenaturing (15%) polyacrylamide gel reveal the presence of only hairpins for repeat lengths, $n > 5$ at both 5 and 200 mM NaCl concentration [32].

B. Structural Characterization of $(GCC)_n$ by NMR

The imino proton spectra of $(GCC)_5$ and $(GCC)_6$ at 5°C and at pH 6.3 show the presence of G-imino protons within 13.4–13.1 ppm that correspond to Watson–Crick G · C pairs as well as a broad envelop around 11.0 ppm that corresponds to loop G-imino protons. The temperature-dependent imino proton profile of $(GCC)_{5,6}$ reveals that the loop G-imino signals disappear above 5°C. Deconvolution of the areas under the imino signals indicates the presence of four G · C pairs in $(GCC)_5$ and five G · C pairs in $(GCC)_6$ which are consistent with either a blunt hairpin or a slipped hairpin. For example, a blunt hairpin of $(GCC)_5$ should have the G1 · C15 pair whereas the slipped hairpin of $(GCC)_5$ should have the unpaired C15. See Fig. 41-3A for descriptions of slipped and blunt hairpins of $(GCC)_5$. In order to distinguish between the slipped and blunt hairpin, imino proton spectra have been recorded at 5°C for the analogs, $(GCC)_4GC$ (Fig. 41-3B) and $G(GCC)_5$ (Fig. 41-3C). If $(GCC)_5$ formed a blunt hairpin, the removal of G15 should show the loss of G1 · C15 pair in the imino spectrum of $(GCC)_4GC$ whereas if $(GCC)_5$ formed a slipped hairpin, the removal of G15 should leave the imino spectrum of $(GCC)_4GC$ unaltered which is exactly what we have observed [32]. Again the addition of a 5′ G (i.e., G0 in $G(GCC)_5$ of Fig. 41-3C) should lead to an increase in the total number of G · C pairs for a slipped hairpin whereas the same modification for a blunt hairpin should have no change in the imino spectrum. In fact, $G(GCC)_5$ shows an increase in the number of imino protons corresponding to G · C pairs [32]. Hence, the imino spectrum of $G(GCC)_5$ is also consistent with a slipped hairpin structure of $(GCC)_5$. Note that the same base pairing pattern is preserved in the stems of slipped hairpins formed by $(GCC)_5$ and $(GCC)_6$. However, the number of nucleotides in the loop is different in the two cases: as shown in Fig. 41-3, four nucleotides are present in the loop of the slipped $(GCC)_5$ hairpin while three nucleotides are present in the loop of the slipped $(GCC)_6$.

More direct evidence for a blunt or a slipped $(GCC)_n$ hairpin is obtained by monitoring the pairing of the C at the CpG step in this triplet repeat, i.e., in a blunt hairpin this C should be G · C-paired whereas in a slipped hairpin it should be C · C-paired. We have identified the pairing of the C at CpG site of the $(GCC)_n$ hairpin by performing ^{15}N-^{1}H HSQC (heteronuclear sin-

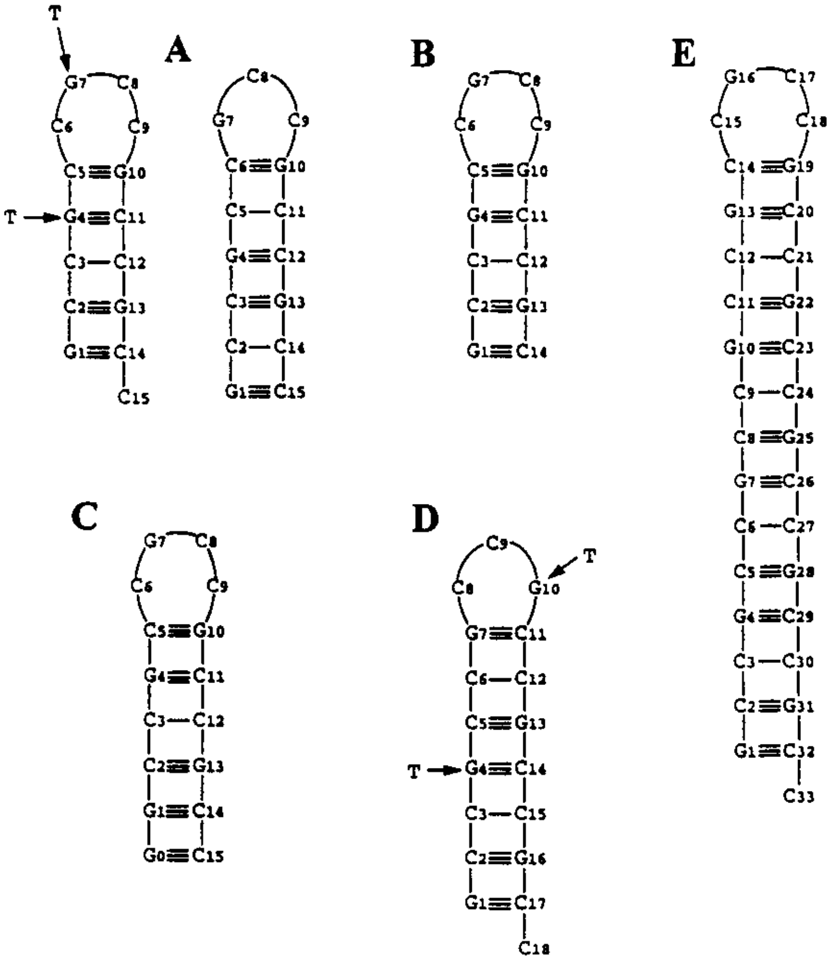

FIGURE 41-3 The hairpin structures of (A) $(GCC)_5$, (B) $(GCC)_4GC$ (C) $G(GCC)_5$, (D) $(GCC)_6$, and (E) $(GCC)_{11}$. The blunt and the slipped hairpins are shown for $(GCC)_5$. NMR data are only consistent with the slipped hairpin structures of $(GCC)_{5,6,11}$. Loop-G signals around 11.0 ppm are observed in all three systems. With increasing temperature loop-G resonance is the first to disappear, e.g., above 5°C in $(GCC)_5$. Subsequently the imino resonances from the terminal G · C pairs in the stem disappear at 35°C. The effect of single site G→T substitutions are also studied. The H-bonding and open-closure of the Cs in the C · G and C · C pairs in stem and the Cs in the loop are studied by specific ^{15}N4-labeling of the Cs.

gle quantum coherence) spectroscopy on three oligomers [105]. Two of them are $(GCC)_5$ sequences and both are ^{15}N4 (amino)-labeled at single sites (one at C11 and the other at C3). The third is a 7-base-pair-long duplex, $(C1\underline{G2C3C4G5C6}G7)_2$ with ^{15}N4-labeling at C3 and C4. In the duplex, the central 5 (underlined) base pairs mimic the building block of the stem of a $(GCC)_n$ hairpin. Also, in the duplex C3 is G · C-paired whereas C4 is C · C-paired and this allows unambiguous identifications of the ^{15}N4/^{1}H signals of Cs in the G · C and C · C pairs. The ^{15}N-^{1}H HSQC spectrum of $(GCC)_5$ with ^{15}N-labeling at C11 shows a pair of crosspeaks as expected from a C in a G · C pair. This is only consistent with a slipped $(GCC)_5$ hairpin (and not with a blunt hairpin). Again, the ^{15}N-^{1}H HSQC spectrum of $(GCC)_5$ with ^{15}N-labeling at C3 shows a single crosspeak as expected from a C in a C · C pair which is only consistent with a slipped $(GCC)_5$ hairpin.

We have also studied the interaction and exchange properties of the C · C pair in a slipped $(GCC)_5$ hairpin [105]. For this we have incorporated ^{15}N4-labels at C2 and C11 (both involved in G · C pairs), at C3 and C12 (both involved in C · C pairs), and at C6 (in the loop). The loop amino signal of C is upfield-shifted with respect to the amino signal of C from the C · C pair. The ^{15}N-^{1}H-^{1}H HMQC-NOESY experiments reveal NOEs between the amino protons of C in the C · C pair and the imino protons from the neighboring G · C pairs in the stem. This proves that the C · C pair is internally stacked in the stem of the $(GCC)_5$ hairpin. The pH- and temperature-dependent ^{15}N-^1H HSQC experiments reveal that the amino protons of the C · C pair exchanges more rapidly than those of the G · C pair but slower than those that belong to the C in the loop. The ^{15}N-^1H HSQC spectrum of $(GCC)_{11}$ in which two consecutive Cs in the stem are $^{15}N4$-labeled also confirms that the Cs at the CpG steps of the stem are C · C-paired whereas the Cs at the GpC steps are G · C-paired.

Detailed analyses of the NOESY at 25, 50, 75, 100, 125, 200, and 500 ms of mixing and the DQF-COSY data of the slipped $(GCC)_{5,6}$ hairpins reveal that all the nucleotides adopt (*C2′-endo, anti*) conformation [32]. The presence of continuous sequential interactions involving both exchangeable and nonexchangeable protons reconfirms that the C · C pairs in these hairpins are internally stacked. The Cs in the C · C pairs are not protonated since in $(GCC)_{5,6,11}$ we have observed no imino signal from protonated Cs within the pH range 6–7 [32]. The C · C pairs probably involves a single H bond between amino (N4) donor and imino or carbonyl (N3 or O2) acceptor. This leads to two possibilities in which either of the two Cs can act as a proton donor or an acceptor. As previously shown, the C · C pairs in these hairpins are more susceptible to open-closure than the G · C pairs. In addition, weaker intra- and internucleotide NOESY cross-peaks at the C · C pairs of the $(GCC)_5$ and $(GCC)_6$ hairpins indicate the presence of local flexibility. In 400-ps unrestrained molecular dynamics, the C3 · C12 pair in the $(GCC)_5$ hairpin can undergo local periodic sliding motions between the two degenerate H-bonding states without violating local or distant NOE constraints. Such a sliding motion makes Cs in the C · C pairs intrinsically more flexible than Cs in the G · C pairs. As discussed later, the flexibility of the C · C pair at the CpG step of the $(GCC)_n$ hairpins imparts an exceptional substrate specificity for the human methyltransferase. Figures 41-4A and 41-4B show the ensemble-averaged slipped hairpin structures of $(GCC)_5$ and $(GCC)_6$ as derived from the NMR data.

Gao and co-workers [20] and Mitas and co-workers [55] proposed an "E-motif" for $(CCG)_n$, in which the Cs at the CpG sites are C · G-paired. Gao and co-workers have based their prediction on the slipped duplex structure of $(CCG)_2$ whereas Mitas and co-workers based their prediction on gel electrophoresis, P1 digestion, and chemical modification studies, 5′a$(CCG)_{15}$a′3′, where a and a′ are complementary to each other. We are concerned that pronounced end-effects significantly distort the structures in the short $(CCG)_n$ (n = 2 or 3) duplexes studied by Gao and co-workers whereas the overinterpretation of the experimental data of the $(CCG)_{15}$ sequence with sticky (a and a') flanks mars the structural conclusions derived by Mitas and co-workers.

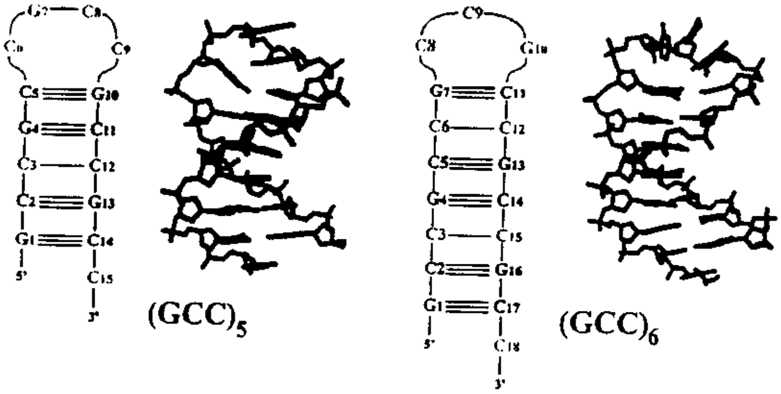

FIGURE 41-4 Schematic and three-dimensional structures of the C-rich strands of the fragile X triplet repeats: $(GCC)_5$ and $(GCC)_6$. The structures are derived using 1D/2D proton NMR spectroscopy and averaged over 100 sampled structures. Both $(GCC)_5$ and $(GCC)_6$ form slipped hairpins with a 3′ overhanging C. Both structures fold so as to maximize G · C Watson–Crick base pairs. Single hydrogen bonded C · C mismatches are present at the CpG steps in the stem. All nucleotides are in (*C2′-endo, anti*) conformations. $(GCC)_5$ has four nucleotides in the loop, whereas $(GCC)_6$ has only three. Analyses of the 2D NMR COSY and NOESY data lead to about 200 NOEs for each structure; 100 structures compatible with the distance constraints are extracted from the 400-ps restrained MD trajectory and energy minimized. All the structures belong to the same cluster with an average Mean Square Deviations (MSD) of 0.6 Å^2 for $(GCC)_5$ and 0.7 Å^2 for $(GCC)_6$.

Therefore, we have studied $(GCC)_{5,6,7,\&11}$ which all form stable hairpins under physiological salt concentrations. In all these hairpins, the Cs at the CpG sites of the stem are C · C-paired and this rules out the possibility of an "E-motif" for $(CCG)_n$.

C. Structural Characterization of $(GGC)_n$ by Gel Electrophoresis

We have carried out gel mobility studies of $(GGC)_n$ for n = 5, 6, 7, and 11. For short repeat lengths, i.e., $(GGC)_{5,6,7}$, two populations have been observed: a homoduplex (the major population) and a hairpin (the minor population). Higher DNA and salt concentrations favor the duplex population. However, for longer repeat lengths, i.e., $(GGC)_{>11}$, the hairpin is the predominant population at all DNA and salt concentrations. Therefore, the gel data [32] unambiguously demonstrate that the $(GGC)_n$ strands of the fragile X repeat are capable of forming hairpin structures when the repeat number is large (i.e., $n > 11$).

D. Structural Characterization of $(GGC)_n$ by NMR

Although gel electrophoresis indicates the presence of both hairpin and duplex structures for $(GGC)_n$, only duplex structures are predominantly present for n = 4–11 under NMR solution conditions (DNA concentrations being two orders of magnitude higher in NMR experiments). We have determined high-resolution structures of $(GGC)_{4,5,6}$ by NMR [32] since the duplex structure adequately models the stem of the hairpin formed by longer $(GGC)_n$ sequences. Figure 41-5 schematically describes the $(GGC)_4$ duplex and its analogs that we have studied to determine the pairing scheme in the duplex.

A detailed analysis of NMR data reveals that $(GGC)_4$, $(GGC)_5$, and $(GGC)_6$ all form duplexes with a 6-base-pair-long structural repeat,

$$G1^{anti}\text{-}G2^{anti}\text{-}C3^{anti}\text{-}G4^{anti}\text{-}G5^{syn}\text{-}C6^{anti}\text{-}G7^{anti}$$

$$C1^{anti}\text{-}G2^{syn}\text{-}G3^{anti}\text{-}C4^{anti}\text{-}G5^{anti}\text{-}G6^{anti}\text{-}C7^{anti}$$

Two symmetric O6—H-N1 H bonds are present in the G^{anti} · G^{syn} pairing [32, 56]. Figure 41-6A shows the ensemble-averaged structure of the $[(GGC)_4]_2$ duplex that is consistent with the NMR data. We used a molecular modeling approach to construct the hairpin structures of the G-rich strands. The stem of the hairpin is constructed on the basis of the NMR data of the duplex and then the two arms of the stem are connected by an energetically stable loop segment. Figure 41-7B shows the proposed energy-minimized hairpin model of $(GGC)_9$ in which the stem structure is consistent with the NMR data of $[(GGC)_4]_2$ duplex.

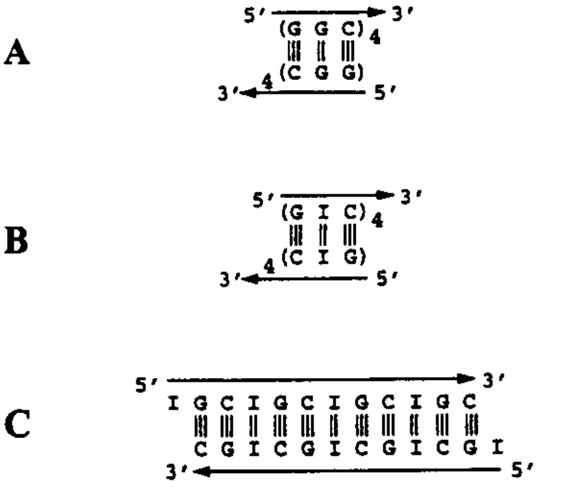

FIGURE 41-5 The H-bonding schemes in the (A) $(GGC)_4$, (B) $(GIC)_4$, and (C) $(IGC)_4$ duplexes. NMR studies on $(GGC)_4$ and its analogs with G → I substitutions help us to prove that the G · G base pairing in $(GGC)_4$ is through the imino protons.

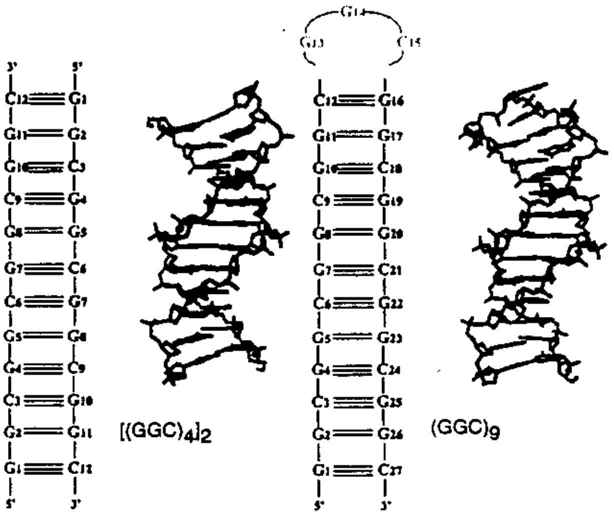

FIGURE 41-6 (left) Three-dimensional structure of $[(GGC)_4]_2$ averaged over 100 structures derived using 1D/2D NMR spectroscopy and restrained molecular dynamics simulations; 200 distance constraints for the restrained MD simulations are estimated from the mixing-time-dependent NOESY data using full-relaxation matrix analysis. (B, right) Three-dimensional average structure of $(GGC)_9$ hairpin from the 100 hairpin structures determined by restrained MD simulations. The distance constraints for the stem used in the molecular dynamics simulations were the same as those of $[(GGC)_4]_2$ duplex and no constraints are imposed on the loop.

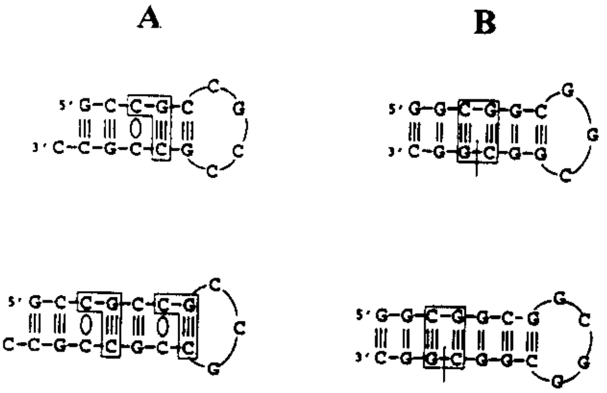

FIGURE 41-7 The nature of CpG sites in (A) $(GCC)_{5,6}$ and (B) $(GGC)_{5,6}$ hairpins. Note that the Cs at the CpG sites in the stem are C · C-paired in the case of $(GCC)_n$ hairpins, whereas in the $(GGC)_n$ hairpins the same Cs are G · C-paired.

Mitas and co-workers [21] have also suggested a $G^{syn} \cdot G^{anti}$ pairing for the 5′a$(CGG)_{15}$a3′ hairpin. However, Gao and co-workers [19] have concluded that there is no G · G pairing in the short $(CGG)_{2,3}$ duplexes since they could not observe any imino signal due to a G · G pair. Probably the short length and less than optimal DNA and salt concentrations hinder the formation of a uniformly paired duplex structure.

E. Structural Differences in the Hairpins Formed by the $(GCC)_n$ and $(GGC)_n$ Strands

NMR and gel electrophoresis data show that the individual $(GCC)_n$ and $(GGC)_n$ strands of the fragile X repeat can form hairpin structures under physiological conditions [13, 32]. The $(GCC)_n$ strand can form a hairpin even for short repeats ($n > 5$) whereas the $(GGC)_n$ strand requires longer repeats ($n > 11$). Also as shown in Fig. 41-7, the CpG sites are different in these two hairpins. In the $(GCC)_n$ hairpins the Cs at the CpG sites in the stem are C · C-paired whereas in the $(GGC)_n$ hairpins the same Cs are G · C-paired. This difference in the local CpG structures in these two hairpins affects their substrate efficiencies for the human methyltransferase because in the catalytic process, the most efficient configuration of the target CpG has been postulated to be the one that involves the C in a C · C pair and the G in a G · C pair [57–59]. As shown in Fig. 41-7, the $(GCC)_n$ hairpins have exactly the same CpG configuration that is preferred by the human methyltransferase whereas the $(GGC)_n$ hairpins have both C and G of the CpG site in G · C pairs. We have carried out a methylation assay [14] with the human methyltransferase on $(GCC)_n$ and $(GGC)_n$ hairpins and the corresponding Watson–Crick duplex, $(GCC)_n \cdot (GGC)_n$ for $n = 5$, 6, 7, 10, 11, 15, 18, and 21. Our results show that for a given repeat length the substrate efficiency of the $(GCC)_n$ hairpin is about 5 times higher than the Watson–Crick duplex, $(GCC)_n \cdot (GGC)_n$. The substrate efficiency of the $(GGC)_n$ hairpin is even lower than the Watson–Crick duplex, $(GCC)_n \cdot (GGC)_n$. It is to be noted that the catalytic domains of methyltransferases are conserved through evolution from bacteria to humans [60]. Also, it has been shown by X-ray crystallography that in the activated (substrate–bacterial methyltransferase) complex, the C of CpG is in a "flipped out" conformation [61, 62]. Since they are C · C-paired, the Cs of CpG in the $(GCC)_n$ hairpin will flip out more easily than the same Cs that are G · C-paired in either the Watson–Crick duplex, $(GCC)_n \cdot (GGC)_n$, or the $(GGC)_n$ hairpin which would account for the higher substrate efficiency of the $(GCC)_n$ hairpin than either the Watson–Crick duplex, $(GCC)_n \cdot (GGC)_n$, or the $(GGC)_n$ hairpin.

F. Evidence for Hairpin-Induced Slippage Structures of the Fragile X Repeat: An *in Vitro* Replication Assay

We have performed *in vitro* replication of M13 single-stranded DNA templates by *Taq* polymerase assays for the intrinsic preference of hairpin formation by the $(GCC)_n$ or $(GGC)_n$ strands in presence of its complementary Watson–Crick partner. Figure 41-8 outlines the experimental design (for details, see [63]). Triplet repeats, $(GCC)_n$ or $(GGC)_n$ [$n = 8$ or 21], are inserted into the single-stranded M13 phage vectors [M13mp18 or M13mp19]. The replication (or primer extension in our case) is carried out by a 17-nucleotide-long primer that attaches to the template 40 nucleotides away from the insert. The replication is stopped by using dideoxy terminators. The replication product is sequenced on an Applied Biosystem Auto-Sequencer. If the insert in the DNA template forms a hairpin, the chain elongation during replication should either continue past the base of the hairpin or stop at the beginning of the hairpin; either result, as shown in Fig. 41-8, will result in a shorter replication product depending upon the size of the hairpin in the insert sequence. If the replication arrests at the insert then the replication product will contain only the initial flanking sequence and part of the insert. If the replication skips the hairpin then the replication product will contain both flanking sequence and a shortened insert sequence. In the absence of a hairpin in the insert, the replication product should correspond to the entire length of the DNA template including the $(GCC)_n$ or $(GGC)_n$ insert and the flanking sequences.

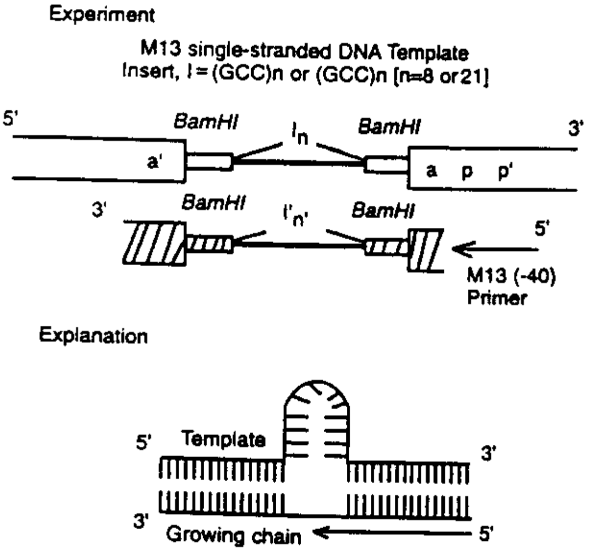

FIGURE 41-8 An *in vitro* replication assay using M13 single-stranded DNA with $(CCG)_n$ or $(CGG)_n$ inserts: the experimental protocol and the explanations of the replication bypass due to hairpin formation in the insert. In the discussion of these results, $(GCC)_n$ or $(CCG)_n$ [$(CGG)_n$ or $(GGC)_n$] are used interchangeably. For the $(GCC)_{21}$, a finite portion of the insert escapes replication at 45, 60, and 72°C since the skipped portion forms a hairpin. The length of the hairpin decreases upon increasing the reaction temperature. When $(GCC)_8$ or $(GGC)_{8,21}$ are used as inserts, there is no hairpin formation at 45, 60, and 72°C.

In all cases, the majority of the replication products belong to one size category. A complete replication product [i.e., $(GGC)_8$ and the two flanking sequences] is obtained for the $(GCC)_8$ insert in the template for a reaction temperature of 60°C which melts the $(GCC)_8$ hairpin. For $(GCC)_{21}$, a finite length of the $(GCC)_{21}$ insert is always bypassed within a temperature range of 45–72°C and the length of the bypass increases with decreasing temperature. However, at a reaction temperature of 85°C, the entire length of the $(GCC)_{21}$ insert is replicated. This observation is explained by the formation of a hairpin structure in the $(GCC)_{21}$ insert in the template. Note that the ends of the hairpin stem can still be replicated because they fray into unpaired single strands. The extent of end-fraying increases with increasing temperature and as a result, the length of the central $(GCC)_{21}$ insert bypassed by *Taq* polymerase decreases with increasing temperature. In this assay, unless the replication traverses the full length of the insert, $(GCC)_n$, complete Watson–Crick complementarity between the insert and the replicated DNA is not achieved. During replication, this strand asymmetry (i.e., lack of perfect complementarity) facilitates hairpin-folding of the $(GCC)_n$ insert. This hairpin may be bypassed if the rate of replication is much faster than the rate of decay of the $(GCC)_n$ hairpin. Therefore, within the range of reaction temperature, 45–72°C, if a residual $(GCC)_n$ hairpin is long enough to be stable, then it may escape replication. This kind of bypassing of a hairpin during replication is analogous to the deletion of a hairpin formed by extrachromosomal palindromic sequences observed during the replication of yeast DNA.

A very different result is obtained when the complementary triplet repeats, $(GGC)_{8,21}$, are inserted into the M13 single-stranded DNA template. Replication bypasses are not observed for either $(GGC)_8$ or $(GGC)_{21}$ within 45–85°C. This is consistent with our earlier observation [13] that the $(GGC)_n$ strand has a lower propensity for hairpin formation than its complementary partner, $(GCC)_n$. Fry and Loeb [12] observed a slowly migrating species in the native gel of the individual $(GGC)_n$ strand. This species, which constituted less than 40% of the total population, was assumed to be a G-quartet structure although from our NMR and gel mobility data we have found no evidence of such a structure [13, 32]. In general, a G-quartet structure in the M13 template blocks replication; for example, the formation of multiply folded G-quartet structure by the insulin-linked polymorphic region (ILPR), $(ACAG_4TGTG_4)_n$, leads to replication arrest at the beginning of the insert and this replication arrest is not normally released even in the presence of replication accessory proteins (*Escherichia coli*, SSB/human RP-A, helicase, etc.—see [93]). However, in the case of a $(GGC)_{21}$ insert in the M13 template a complete replication product, including the insert and the flanking sequences, is observed. This rules out the formation of G-quartet structure during replication. It may be pointed out that hairpin or G-quartet structure of the $(GGC)_n$ strand may be detected in the replication assay for longer lengths (n) or in the presence of KCl.

The individual $(GCC)_n$ strand can form a hairpin structure even for $n = 5$. However, in the presence of its complementary strand, $(GCC)_n$ requires a sufficiently long n (21 in our case) for the formation of a hairpin. Similarly, the individual $(GGC)_n$ strand forms a hairpin structure for $n > 11$. However, in the presence of its complementary strand, the $(GCC)_n$ strand does not form a hairpin even for $n = 21$. Both the $(GCC)_n$ and $(GGC)_n$ strands require even longer n for hairpin formation when replication protein A (RP-A), helicase, etc., are also present during replication [64]. Replication accessory proteins tend to unwind self-assembled single-stranded structures in a partially sequence-specific manner [65]. Nonetheless, since single-stranded regions are created during replication there is always a finite probability of hairpin formation by the (GCC)$_n$ or the (GGC)$_n$ strand of the FraX repeats.

G. Extremely High Methylation Efficiencies of the Hairpin-Induced Slippage Structures: A Methylation Assay

Our *in vitro* replication assay shows that the slippage structures are essentially three-way junctions in which the (GCC)$_n$ hairpin has the potential to slip and slide on the two Watson–Crick duplex arms (Fig. 41-9). Such a process may be facilitated by substrate–enzyme interactions at 37°C. In these three-way junctions, the potential for multiple locations of the (GCC)$_n$ hairpin on the Watson–Crick duplex allows the G · C-paired CpG sites in the Watson–Crick duplex to be converted into the C · C-paired CpG sites in the stem of the hairpin. Therefore, in a mobile three-way junction a larger number of C · C-paired CpG sites will be recruited for methylation than for a fixed hairpin formed by an excess of the (GCC)$_n$ strand. To test this hypothesis [63], we have constructed completely mobile (Fig. 41-9A), partially mobile (Fig. 41-9B), and immobile three-way junctions (Fig. 41-9C). In the completely mobile three-way junctions (Fig. 41-9A), the (GGC) strands of shorter lengths are annealed with the (GCC) strands of longer lengths: for example, (GGC)$_{10}$ · (GCC)$_{15}$, (GGC)$_{10}$ · (GCC)$_{18}$, (GGC)$_{10}$ · (GCC)$_{21}$, and (GGC)$_{15}$ · (GCC)$_{21}$. In the partially mobile three-way junctions (Fig. 41-10B), the free ends of the Watson–Crick duplex are covalently closed by two T$_4$ loops; therefore, in these single-stranded

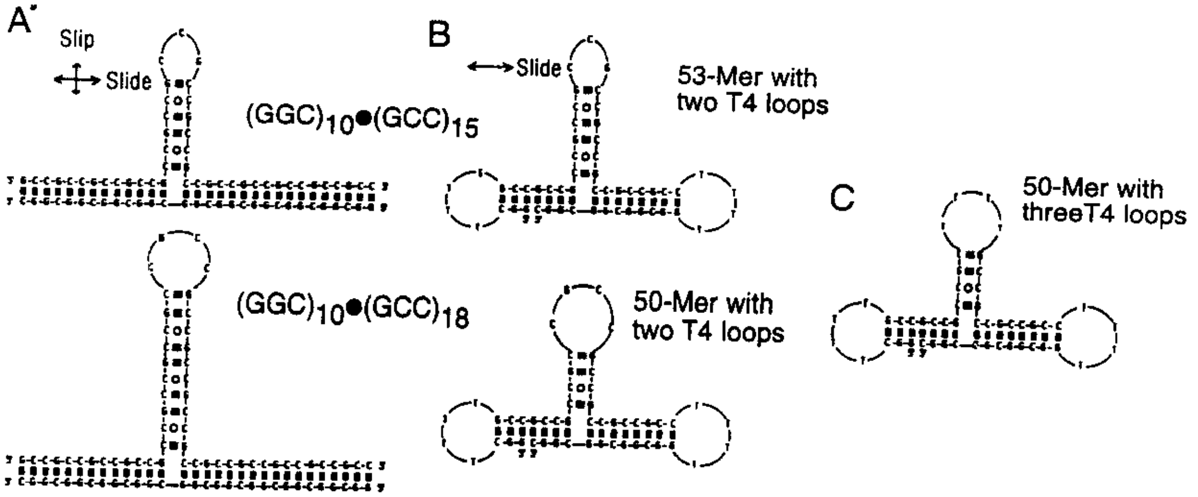

FIGURE 41-9 Three types of three-way junctions: (A) completely mobile, (B) partially mobile, and (C) immobile. (A) Completely mobile three-way junctions are formed by annealing GCC and GGC strands of unequal lengths (e.g., (GGC)$_{10}$ · (GCC)$_{15}$, (GGC)$_{10}$ · (GCC)$_{18}$, (GGC)$_{10}$ · (GCC)$_{21}$, and (GGC)$_{15}$ · (GCC)$_{21}$). Three-way junctions are created by the hairpin formation involving the excess of the longer GCC strands. The loop in the hairpin has either three or four nucleotides depending upon whether the longer GCC strand has odd (e.g., (GGC)$_{10}$ · (GCC)$_{15}$, upper panel) or even (e.g., (GGC)$_{10}$ · (GCC)$_{18}$, lower panel) number of repeats in excess. Note that the Cs at the CpG sites of the hairpin are C · C-paired (and not G · C-paired); this makes the hairpin a better substrate for methylation than the Watson–Crick duplex. Both slipping and sliding of the (GCC)$_n$ hairpin are possible in all of these three-way junctions leading to the conversions of low-affinity Watson–Crick CpG sites of methylation into high-affinity hairpin CpG sites of methylation. In addition, after methylation when the hairpin slides it may return a methylated

*CpG
GpC

to the Watson–Crick duplex which creates a hemimethylated (and high-affinity) CpG site. Sliding (and slipping) of the hairpin double methylation will cause a more efficient double methylation site,

*CpG
GpC*

*C is the 5 methyl cytosine. Note that the presence of the TCC interruption in the loop of the (GCC)$_n$ hairpin will restrict sliding while the TCC interruption in the stem of the (GCC)$_n$ hairpin will destabilize the three-way junction. (B) Two partially mobile three-way junctions: 53- and 50-mer. The ends of the Watson–Crick regions are connected by two T$_4$ loops which essentially prevent slipping of the hairpin. (C) An analog of the 50-mer in (B) in which the (CGCC) loop is replaced by a T$_4$ loop; this prevents slipping and sliding.

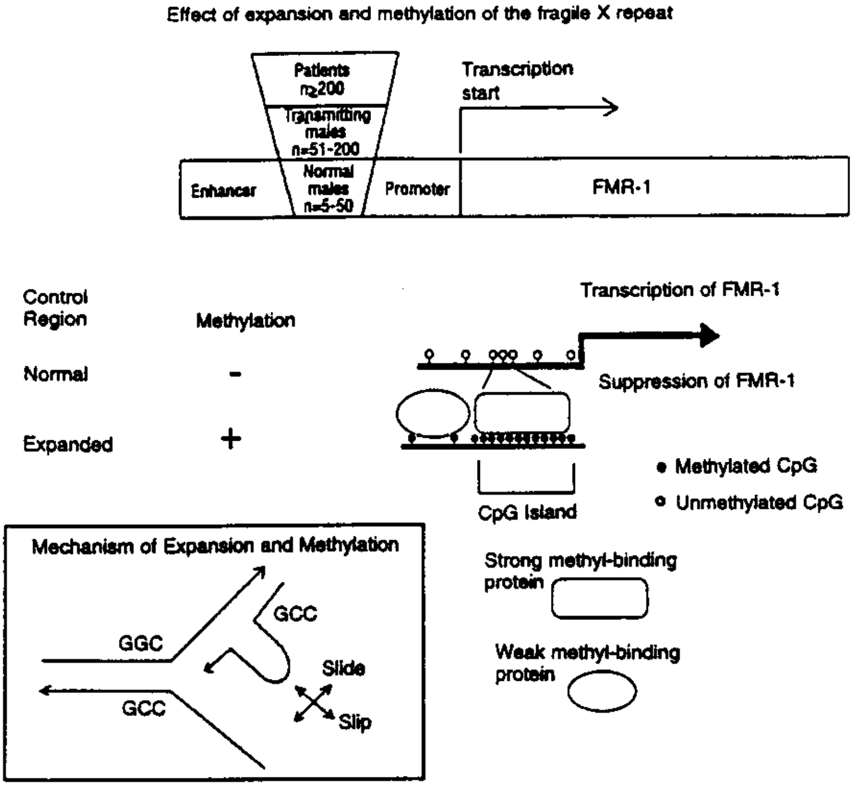

Figure 41-10 A schematic representation of the FMR1 gene. The location of the fragile X repeat, $(GCC)_n \cdot (GGC)_n$, is shown in the 5′ untranslated region although the exact location and the interrelationship of the promoter and the enhancer are not yet known. The ranges of repeat numbers, *n*, are shown for normal, transmitting, and affected males. Therefore, from normal to affected males, the control region undergoes a transition from low-density CpG to high-density CpG island due to the massive expansion and the subsequent methylation of the fragile X repeat. The high-density methylated CpG sites attract strong methyl binding proteins that cannot be dissociated by transcription activator proteins or RNA polymerase and this leads to the suppression of the FMR1 gene. (Inset) A sketch of how the presence of the transient and mobile three-way junction can cause both expansion and methylation. The intrinsic preference of hairpin formation by the GCC strand leads to a three-way junction with a Watson–Crick anchor. When the GGC strand acts as the template, such a three-way junction leads to expansion if the GCC hairpin escapes repair. As shown later, these three-way junctions are also excellent substrates for CpG methylation by the human methyltransferase.

three-way junctions the $(GCC)_n$ hairpin can only slide (and not slip). The partially mobile three-way junctions (Fig. 41-9B) are expected to be less efficient in recruiting new CpG sites in the hairpin conformation than the completely mobile three-way junctions (Fig. 41-9A). In the immobile three-way junctions (Fig. 41-9C) the loop segment of the $(GCC)_n$ hairpin is replaced by a T_4 loop in addition to closure of the free ends of the Watson–Crick duplex by two T_4 loops. The structures of the DNA substrates for methylation have been characterized by combining nondenaturing gel electrophoresis, digestion studies using single-strand-specific P1 nuclease, and NMR studies of the exchangeable imino protons [14]. The possibility of the three-way junctions for $(GGC)_{10} \cdot (GCC)_{15}$ has also been independently established by Kallenbach and coworkers [66] by nondenaturing gel electrophoresis and digestion by *Exo*VII (an enzyme that cleaves single-strand tails at the end of the duplexes). Due to the lack of a single-stranded tail (see Fig. 41-8A), the three-way junction of $(GGC)_{10} \cdot (GCC)_{15}$ was found to be resistant to digestion by *Exo*VII.

The rate of methylation by the human methyltransferase is obtained by measuring the tritium count on cytosines transferred from the tritiated cofactor, *Ado-Met* [57–59]. Table 41-1 lists the rates of methylation for different three-way junctions; two $(GCC)_n$ hairpins and a Watson–Crick duplex are included as controls.

TABLE 41-1 Rates of Methylation by the Human Methyltransferase

Substrates	Rate (fmol/min)	Number of CpG/substrate	Relative rate[a]
	Three-way junctions		
Immobile (Fig. 41-9C)			
50-mer with three T_4 loops	4.0	9	1.0
Partially mobile (Fig. 41-9B)			
50-mer with two T_4 loops	76.6	11	19.2
53-mer with two T_4 loops	92.4	12	23.1
Completely mobile (Fig. 41-9A)			
$(GGC)_{10} \cdot (GCC)_{15}$	76.3	23	25.4
$(GGC)_{10} \cdot (GCC)_{18}$	192.5	26	48.1
$(GGC)_{10} \cdot (GCC)_{21}$	364.9	29	91.2
$(GGC)_{15} \cdot (GCC)_{21}$	298.5	34	74.6
	Controls		
Watson-Crick duplex			
$(GGC)_{15} \cdot (GCC)_{15}$	29.6	28	7.4
Hairpins			
$(GCC)_{10}$	30.2	9	7.6
$(GCC)_{21}$	159.7	20	39.9

[a] Scaled with respect to the rate for the 50-mer with four T_4 loops.

The rates are scaled for the enzyme to DNA ratio. The effective rate of methylation is determined by the initial substrate–enzyme recognition, the kinetics of the transition to the activated state, and the subsequent release of the product. For the initial recognition, the methyltransferase requires the target CpG site and additional flanking base pairs [60]. The actual size and the sequence of the recognition element distinguish one methyltransferase from another although the catalytic mechanism involving the "flipped-out C" remains the same for all the enzymes [62]. Hence, once the Watson–Crick duplex or the hairpin is above a critical size and has the correct recognition element, the kinetics of transition to the activated state essentially determines the rate of methylation. In the completely mobile three-way junctions, the effective rate of methylation is governed by the following factors: (i) the number of the CpG sites in the Watson–Crick duplex, (ii) the number of CpG sites in the hairpin, and (iii) the rate of interconversion of the Watson–Crick CpG sites to the hairpin CpG sites due to the slipping and sliding of the $(GCC)_n$ hairpin. The third factor creates a greater number of high-affinity hairpin CpG methylation sites in the mobile three-way junctions than in a hairpin of fixed length. In addition, after methylation, if the $(GCC)_n$ hairpin slips or slides, it generates hemimethylated CpG sites in the flanking Watson–Crick duplexes which are again better substrates for methylation than the unmethylated CpG sites [58, 59]. Therefore, due to the presence of high-affinity hairpin CpG sites and hemimethylated Watson–Crick CpG sites, the completely mobile three-way junctions are expected to be much better methylation substrates than either the single $(GCC)_n$ hairpin or the Watson–Crick $(GCC)_n \cdot (GGC)_n$ duplex.

The importance of the mobility of the $(GCC)_n$ hairpin in the three-way junctions becomes evident from comparisons of the rates of methylation for the partially immobile three-way junctions with two T_4 loops (see Fig. 41-9B) with the rate of methylation for the completely immobile junction (see Fig. 41-9C). Although these two types of substrates have an almost equal number of CpG sites, the partially mobile three-way junction is 20 times more efficient than the immobile three-way junction (Table 41-1). This difference is attributed to the fact that the $(GCC)_n$ hairpin is able to slide in the partially mobile three-way junctions (see Fig. 41-9B) whereas it is completely locked in the immobile three-way junction (see Fig. 41-9C). Also, note that the three-way junctions with the highest probability of slipping and sliding, namely, $(GGC)_{10} \cdot (GCC)_{21}$ and $(GGC)_{15} \cdot (GCC)_{21}$, also have the highest rates of methylation.

The three-way junction, $(GGC)_{10} \cdot (GCC)_{21}$, is about 14 times more efficient as a substrate than the Watson–Crick $(GCC)_{15} \cdot (GGC)_{15}$ duplex although these two substrates have almost the same number of CpG sites. The observed difference in methylation in these two substrates results from the differences in their structure and dynamics. The rate of methylation for $(GGC)_{10} \cdot (GCC)_{21}$ is over twice that of the single $(GCC)_{21}$. This observation argues against the possibility of a reaction mechanism in which $(GGC)_{10} \cdot (GCC)_{21}$ dissociates into single $(GCC)_{21}$ and $(GGC)_{10}$ hairpins. If this were true, the rate of methylation for $(GGC)_{10} \cdot (GCC)_{21}$ should equal the rate of methylation for the single $(GCC)_{21}$ hairpin because, as previously mentioned, the $(GGC)_{10}$ hairpin does not get methylated by the human methyltransferase [13]. Therefore, the presence and mobility of the $(GCC)_n$ hairpin make the three-way junctions (see Fig. 41-9A) better substrates for methylation by the human methyltransferase than either the single $(GCC)_n$ hairpin or the Watson–Crick $(GCC)_n \cdot (GGC)_n$ duplex.

H. A Mechanism of Suppression of the FMR1 Gene in Fragile X Syndrome

Figure 41-10 describes a molecular mechanism in for the expansion and hypermethylation of the fragile X triplet repeats based upon our high-resolution NMR, *in vitro* replication, and methylation data. The intrinsic preference of hairpin formation by the GCC strand initiates mobile three-way junctions during replication that provide a molecular basis for the repeat expansion and hypermethylation of the CpG island inside the fragile X repeat. The resulting high density of methylated CpG islands provides binding sites for methyl-CpG-binding proteins [67–71] leading to the suppression of the FMR1 gene.

Since TCC/GGA interruptions confer stability to the FraX repeat [54], we have performed 1D NMR studies to examine the effect of the TCC interruptions inside the stem and the loop of the $(GCC)_n$ hairpins. The imino protons of the $(GCC)_n$ hairpins with and without TCC interruptions show different temperature-dependence profiles, thereby also suggesting (qualitatively) differences in their stabilities. The imino protons of the $(GCC)_5$ hairpin disappears at 40°C. The imino protons of the $(GCC)_5$ hairpin with a TCC interruption in the stem disappear at 20°C whereas the imino protons of thé $(GCC)_5$ hairpin with a TCC interruption in the loop disappear at 35°C. It appears that the TCC interruption in the stem causing two consecutive mismatches significantly destabilizes the $(GCC)_n$ hairpin [63]. Hence, such an interruption should weaken the possibility of slippage during replication and allow stable transmission of the fragile X repeats over generations. On the other hand, the TCC interruption in the loop causes a marginal difference in the stability of the $(GCC)_n$ hairpin [63]. However, such an interruption involving a CTCC loop (like the T_4 loop) will restrict the mobility of the $(GCC)_n$ hairpin in the three-way junction (see Figs. 41-9B and 41-9C). Even if formed during replication, the immobile three-way junction with a TCC loop will be efficiently repaired. In addition, as shown in Table 41-1, an immobile three-way junction is a poor substrate for methylation. Therefore, our studies help us visualize how the TCC/GGA interruptions protect against the expansion and hypermethylation of the fragile X repeat.

III. MYOTONIC DYSTROPHY (DM) TRIPLET REPEATS, $(CTG)_n$: ROLE OF THE HAIRPIN STRUCTURES IN EXPANSION AND ABNORMAL EXPRESSION OF THE DMPK GENE

Myotonic dystrophy (DM) is an autosomal dominant disorder characterized primarily by myotonia and progressive weakness and is the most common adult-onset muscular dystrophy [1, 2, 4]. The rare congenital form of DM is associated with profound hypotonia and mental retardation. The gene that causes myotonic dystrophy (DMPK-myotonic dystrophy protein kinase) has recently been identified [72, 73]. The DMPK gene also contains regions of strong homology to cAMP-dependent protein kinase. The 3′ untranslated region of the gene contains CTG triplet repeats (Fig. 41-1B). The length of this triplet repeat sequence is highly polymorphic with the number of copies varying from 5 to 37 in normal individuals. The carriers and affected individuals have more than 39 copies and in some cases it expands beyond 2000 copies. The degree of expansion correlates with the severity of the disease and in a few cases reverse mutations have been observed resulting in the reduction in the number of CTG triplet sequence to the normal population range with the concomitant disappearance of the DM symptoms. This suggests that DM is primarily caused by expansions of the CTG repeat at the 3′UTR of the DMPK gene.

Structural studies by gel electrophoresis and high-resolution NMR are discussed in this section (for details, see [33]). It is shown that the $(CTG)_n$ repeats form stable hairpin structures under physiological salt concentrations even for short repeat lengths (i.e., $n = 5$ or 6). High-resolution NMR spectroscopy allows detailed analyses of the structures and dynamics of these DNA hairpins. Our preliminary data also suggest that similar hairpins are also possible at the RNA level. The presence of hairpins in the 3′UTR of the DMPK mRNA may be of biological significance for various reasons. For example, it has been shown in *E. coli* that putative hairpin sequences in the 3′UTR of a gene can mediate efficient termination of mRNA transcription [74]. Similarly, it has also been shown that hairpin forming sequences, which are evolutionarily conserved (from Xenopus to humans), are present in the 3′UTR of the histone genes and these sequences are involved in mRNA processing and/or in mRNA transport from the nucleus to the cytoplasm [75]. The transport of the DMPK mRNA is particularly relevant for DM since in normal phenotypes the mRNA is efficiently transported from the nucleus whereas in disease phenotypes the mRNA remains bound to the nuclear matrix proteins [76]. Therefore, it is important to understand the structural difference between the 3′UTR $(CUG)_n$ repeats in the mRNA of normal phenotypes and those (expanded) repeats in disease phenotypes.

A. Structural Characterization of $(CTG)_n$ by Gel Electrophoresis

Figure 41-11 shows the possible structural forms of $(CTG)_{5,6}$ and their analogs. Note that $(CTG)_{5,6}$ can adopt either a monomeric hairpin or a mismatched duplex. The nondenaturing gel electrophoretic mobility data of $(CTG)_{5,6}$ distinguish these two possibilities. $(CTG)_{5,6}$ migrate faster than the 10-base-pair duplex. This suggests the presence of unimolecular hairpins. The $(CTG)_5$ hairpin is expected to migrate like a 7/8-base-

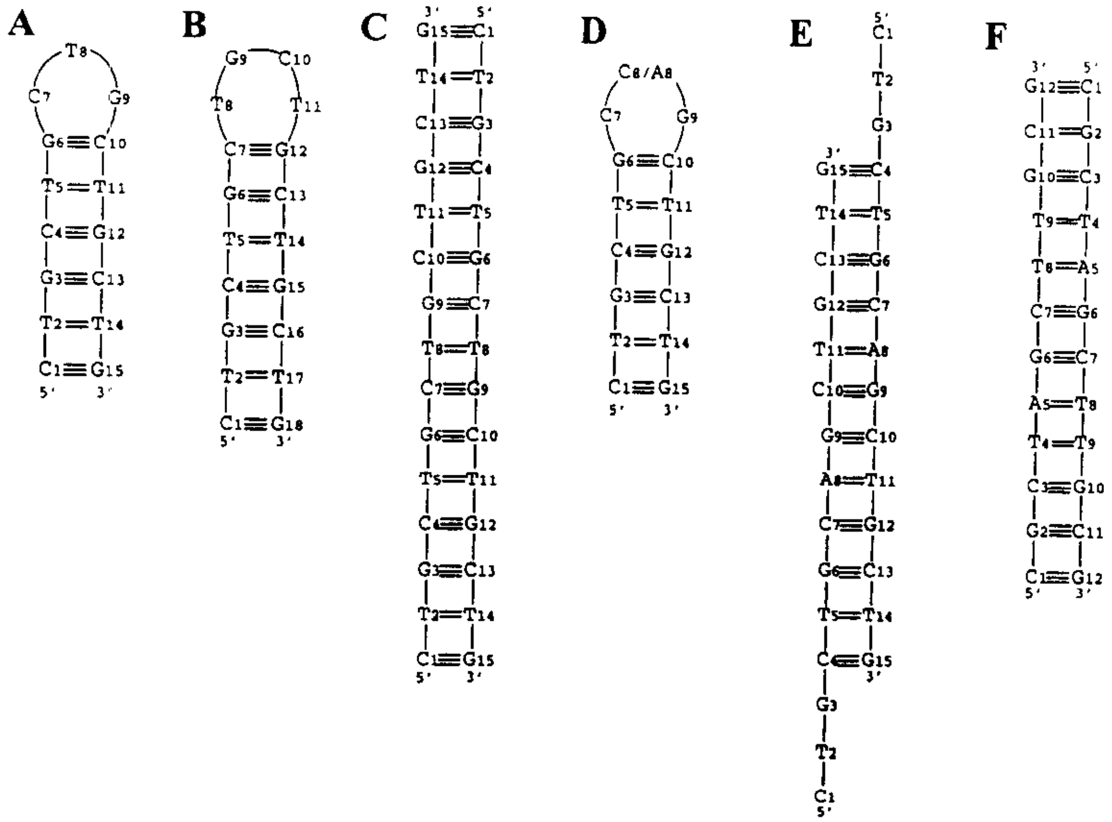

FIGURE 41-11 Secondary structures formed by various $(CTG)_n$ analogs. Hairpins: (A) $(CTG)_5$, (B) $(CTG)_6$, (D) $(CTG)_2CCG(CTG)_2$ and $(CTG)_2CAG(CTG)_2$, and (F) CGCTAGCTTGCG. Homoduplexes: (C) $(CTG)_5$, (E) $(CTG)_2$-CAG$(CTG)_2$, and (G) CGCTAGCTTGCG. The oligomer in (G) has been shown to form two T · T pairs with two H bonds. A comparative NMR study on $(CTG)_{5,6}$ also shows the presence of T · T pairs with two H bonds. The $T_8 \rightarrow C_8$ substitution in $(CTG)_5$ does not alter the hairpin structure whereas the $T_8 \rightarrow A_8$ induces a hairpin-to-duplex equilibrium.

pair-long duplex whereas the $(CTG)_6$ hairpin should migrate like a 9/10-base-pair-long duplex. Also, similar gel patterns are observed under two different (i.e., 5 and 200 mM) NaCl concentrations. In addition, hairpins still remain the predominant conformation even when the DNA concentrations of $(CTG)_{5,6}$ are raised from 0.25 to 25 mM [33].

B. Structural Characterization of $(CTG)_n$ by NMR

Analyses of the NOESY and DQF-COSY data also reveal that all the constituent nucleotides in $(CTG)_{5,6}$ hairpins adopt (*C2'-endo, anti*) conformations with two H-bonded T · T pairs in the stem [33, 77]. Observation of a few (but key) interproton distance constraints defining intraloop and loop–stem interactions in the $(CTG)_5$ hairpin allows us to distinguish among four different (CTG) loop conformations: (i) three bases in the 3′ side of the stem, (ii) one base in the 5′ with two bases in the 3′ side of the stem, (iii) two bases in the 5′ with one base in the 3′ side, and (iv) three bases in the 5′ side of the stem; 200-ps-restrained MD simulations have been done separately using each model as a starting configuration. The structures derived from these four models show difference only in the single-stranded loop segments of the hairpins. In model (i), all the three bases in the loop are stacked with the 3′ side of the stem. In model (ii), T8 and G9 are stacked with each other on the 3′ side while C7 is stacked on the 5′ side of the stem. Model (iii) has G9 stacked with the 3′ side of the stem while C7 and T8 are stacked in the 5′ side of the stem. In model (iv), C7, T8, and G9 are all stacked with the 5′ side of the stem although G9 is partially flipped out of the stacked array. Model (i) shows better agreement with the distance constraints in the loop. Figure 41-12A shows the lowest energy structure of $(CTG)_5$ belonging to the family of models (i).

Figure 41-12B shows the lowest energy structure for $(CTG)_6$ that best satisfies the NMR constraints. In this structure, the four nucleotides in the (TGCT) loop are divided equally on each side of the stem (i.e., two nucleotides on each side). This results in a T · T pair in

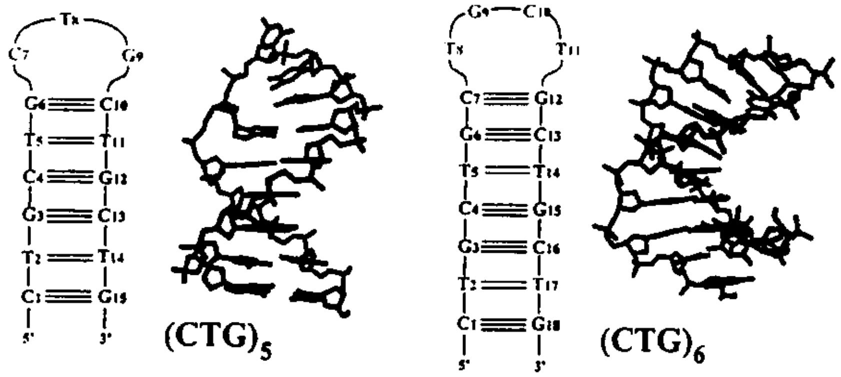

FIGURE 41-12 Three-dimensional average structures of $(CTG)_5$ and $(CTG)_6$. The structures are derived using 1D/2D proton NMR spectroscopy. Both $(CTG)_5$ and $(CTG)_6$ form blunt hairpins. Both structures fold so as to maximize G-C Watson-Crick base pairs. Two hydrogen bonded T · T mismatches are present in the stem. All nucleotides are in (C2'-*endo*, *anti*) conformations. $(CTG)_5$ has three nucleotides in the loop, whereas $(CTG)_6$ has four. Analyses of 2D NMR COSY and NOESY data lead to 200 NOEs for each structure. NOEs were converted into distances using full-relaxation matrix analysis; 100 structures compatible with the distance constraints were generated. All the structures belong to the same cluster with an average MSD lower than 1.0 Å^2 for each structure.

the loop although we do not have any experimental evidence in favor of such a pairing. It is possible that the T · T pair in the loop of $(CTG)_6$ opens and closes so fast on the NMR time scale that the imino signal is not observed.

For $(CTG)_n$, we have chosen $n = 5$ and 6 for our NMR studies since at these repeat lengths a hairpin is the only conformation under physiological salt concentrations [33]. Gao and co-workers [19] have studied the short $(CTG)_{2,3}$ duplexes, probably to more accurately determine the structures of the stem of a $(CTG)_n$ hairpin. However, they have not been able to determine whether the T · T pairs in the duplex are singly or doubly H-bonded. We have combined pH and temperature studies on various $(CTG)_n$ sequences to show that the T · T pairs are doubly H-bonded [33]. Mitas and co-workers [15] have performed gel electrophoresis, P1 digestion, and chemical modification studies on 5'a$(CTG)_{15}$a'3'. It is not surprising that the modified $(CTG)_{15}$ sequence tends to form a hairpin structure since the flanking a and a' impose a constraint for hairpin folding. Although their system of choice has been biased, Mitas and co-workers have reached qualitatively the right conclusion regarding the $(CTG)_n$ sequences. However, they have managed to overinterpret their data to arrive at several wrong conclusions about the finer details of the $(CTG)_n$ hairpins. They have suggested that the T · T pairs are singly H-bonded. They have also proposed an inaccurate loop structure of the $(CTG)_n$ hairpins.

C. Site-Specific Dynamics of the $(CTG)_n$ Hairpins

We have calculated the order parameters [33, 78–80] for different interproton vectors in the $(CTG)_{5,6}$ hairpins by MD simulations hairpins with only hydrogen bonding constraints (and with no NOE constraints). The calculated values are compared with those estimated from the experimental cross-relaxation constants. Theoretical and experimental values of the order parameters, S^2, and the apparent correlation times, t_a, are computed for interproton vectors with fixed distances such as H6-H5 in cytosines and H2'-H2" in sugars. NOE intensities for these two vectors only reflect the dynamics of the corresponding cyotosines and sugars in the $(CTG)_{5,6}$ hairpins. On a scale of 1 to 0, $S^2 = 1$ implies extreme rigidity and $S^2 = 0$ implies extreme flexibility. Similarly a small value of t_a implies extreme flexibility whereas a high value of t_a implies extreme rigidity. The cytosines in the loop, i.e., C7 in $(CTG)_5$ and C10 in $(CTG)_6$, are most flexible. In the $(CTG)_{5,6}$ hairpins, the Cs in the interior of the stems are least flexible. However, the Cs at the two termini of the stem are moderately flexible.

Analyses of S^2 and t_a of various intrasugar H2'-H2" dipolar interactions indicate that in $(CTG)_5$, G3 and G12 (both in the stem) are the least flexible (Table 42-2). Theoretical calculations reveal that within the loop of the $(CTG)_5$ hairpin, C7 and T8 are less flexible than G9. Sugars corresponding to mismatches in the $(CTG)_5$ hairpin are also more flexible than those from Watson–

3 (T)

TABLE 41-2

Interaction	Base position	σ (s^{-1})[a]	τ_{app} (ns)
H5–H6	$(CTG)_5$		
	C7	0.48	2.1
	C1/C10	0.53	2.2
	C4/C13	0.61	2.5
	$(CTG)_6$		
	C1	0.66	2.7
	C10[b]		
H2′-H2″	$(CTG)_5$		
	T8	0.86	1.0
	T5/T11/T14	0.58	0.9
	$(CTG)_6$		
	G9/G18	0.34	0.8
	C1/C16	0.86	1.0
	C4/C7	0.80	1.0
	T2/T11	0.80	1.0
	C10/T8	0.98	1.1
	T5/T14/T17	0.46	0.9

Note. The methodology of the estimation of σ and τ from the NOESY data at 0–125 ms of mixing is described in Refs. [33] and [78–80].

[a] 10% error in the estimated σs.

[b] NOE, not observed up to 125 ms of mixing.

Crick pairs. Similar features are also observed for $(CTG)_6$.

D. Possible Role of the $(CUG)_n$ Hairpins in the Processing/Transport of the DMPK mRNA

The intrinsic propensity of hairpin formation by the $(CTG)_n$ sequence may also manifest itself at the level of mRNA. The formation of RNA hairpins by the $(CUG)_n$ sequences on the 3′ untranslated side of the DMPK gene may either halt the transcription machinery or provide a specific target for protein binding in the post-transcriptional mRNA processing and/or mRNA transport. It has been reported [81] that the levels of precursor mRNAs from the normal and DM alleles show no difference. However, the posttranscriptional processing of the normal and DM alleles are quite different in that the mRNA maturation is severely impaired when $(CUG)_n$ triplets are expanded in disease phenotypes. As stated earlier, it has also been demonstrated that the precursor mRNA remains bound to the nuclear matrix in DM phenotypes [76]. These data agree with our hypothesis that a few $(CUG)_n$ hairpins enable the formation of specific RNA–protein complexes required for efficient termination of transcription and for post-transcriptional mRNA processing or transport. This specificity is impaired when the $(CUG)_n$ triplets are expanded or the formation of multiple hairpins in expanded repeats allows several single-stranded loops in different hairpins to bind to the nuclear matrix proteins which are specific for single-stranded regions.

IV. HUNTINGTON'S DISEASE (HD) REPEAT, $(CAG)_n$: UNUSUAL HAIRPIN STRUCTURES AND THEIR ROLE IN EXPANSION

Huntington's disease (HD) is an autosomal dominant syndrome linked to chromosome 4p [82, 83]. Movement disorder, emotional disorder, and dementia are the common symptoms in HD. Recently, the gene responsible for HD has been identified from chromosome 4p, which is referred to as IT15 (Interesting Transcript 15). The HD gene contains a CAG triplet repeat sequence inside the first exon (Fig. 41-1C). Normal individuals have between 11 and 34 copies with a median of 19, whereas affected individuals have 37 and 86 copies with a median of 45. Patients with longer repeats have an earlier age of onset with a high correlation between the length of the repeat and the age of onset. The onset pattern of the disease suggests a possible role of expansion of the CAG repeat in Huntington's disease.

In this section, we describe the hairpin and homoduplex structures formed by the $(CAG)_n$ repeats (for details, see [46]). For shorter repeat numbers (i.e., $n = 5$ or 6) the $(CAG)_n$ repeats form homoduplex structures whereas for longer repeats (i.e., $n = 10$ or 11) they form doubly folded hairpins, i.e., hairpins with two single-stranded loops. Single H-bonded A · A pairs are present in the homoduplexes and in the hairpins. We have also performed an *in vitro* replication assay to show the presence of a hairpin for $(CAG)_{21}$ in the template.

A. Hairpin and Homoduplex Structures of $(CAG)_5$ and $(CAG)_6$

1. GEL ELECTROPHORESIS

Figure 41-13 shows the schematic representations of hairpin and mismatched duplex structures of $(CAG)_5$ and $(CAG)_6$. Previously, we have shown for $(GCC)_{5,6}$ and $(CTG)_{5,6}$ that although the hairpin folding is different for odd (i.e., $n = 5$) and even (i.e., $n = 6$) repeat numbers the base-pairing scheme of the stem remains the same. In Fig. 41-13, the $(CAG)_{5,6}$ hairpins are assumed to have similar folding patterns for odd and even repeat numbers. The gel mobility of $(CAG)_5$ and

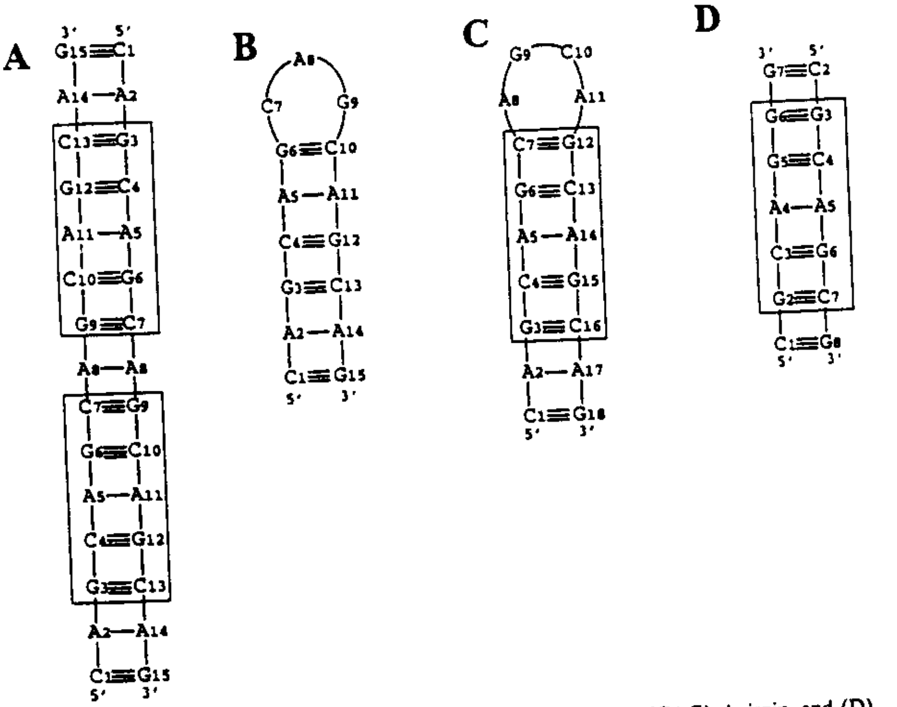

FIGURE 41-13 (A) $[(CAG)_5]_2$ duplex, (B) $(CAG)_5$ hairpin, (C) $(CAG)_6$ hairpin, and (D) 7 base pair duplex, $[(CGCAGCG)]_2$. The nucleotides are numbered from 5′ to 3′ direction. The 5-base-pair consensus structural motif is indicated by boxes in all the structures except for $(CAG)_5$ hairpin. Note that the hairpins with odd number of copies have three nucleotides, whereas hairpins with even number of copies have four nucleotides in the loop. The 7-base-pair duplex is chosen in order to characterize the detailed conformation of all the five nucleotides in the 5-base-pair structural motif. The terminal 5′ and 3′ G · C pairs are added to arrest the end-fraying of the 5-base-pair structural motif.

$(CAG)_6$ at neutral pH under different NaCl concentrations (20, 150, and 500 mM) show that $(CAG)_5$ forms both hairpin and homoduplex structures as major and minor species, respectively, whereas $(CAG)_6$ forms exclusively a hairpin structure. It appears that within the range of 20–500 mM NaCl concentrations, the $(CAG)_6$ hairpin is thermodynamically more favorable than the $(CAG)_5$ hairpin. The gel mobility data at three different DNA concentrations (5, 10, and 20 mM) for $(CAG)_5$ and $(CAG)_6$ also reveal that they exist predominantly in the hairpin form at 5 mM DNA and increasing the DNA concentration increases the relative population of the homoduplex.

2. HIGH-RESOLUTION STRUCTURE OF $[(CGCAGCG)]_2$ AND ITS RELEVANCE TO $[(CAG)_5]_2$

Within the range of DNA concentrations required for NMR experiments, we are unable to trap exclusively the hairpin conformation of $(CAG)_5$. Hence, our structural studies are restricted to the self-assembled $[(CAG)_5]_2$ duplex, which also adequately models the stem of the $(CAG)_5$ hairpin (Fig. 41-13). We have also carried out structural studies on the 7-base-pair duplex, $(CGCAGCG)_2$ which shares a common $[(GCAGC)]_2$ motif with the $[(CAG)_5]_2$ duplex or the stem of the hairpin. This enabled us to determine the exact stereochemistry of each nucleotide in the common motif and the exact nature of the A · A pairing.

Measurements of thermodynamic stabilities of various pur · pur/pyr · pyr/pyr · pur mispairs in a model duplex suggest that the A · A base pair is thermodynamically the least stable of all mispairs [86]. However, recent ^{1}H NMR and UV-melting studies indicate that A · A base pairs can be incorporated into a DNA duplex without globally distorting the structure and a without a drastic reduction in the stability [87]. Therefore, it is of interest to determine the nature of the pairing and stacking of the adenines in the $[(CAG)_5]_2$ duplex. For this, we have carried out heteronuclear (^{15}N-^{1}H) NMR experiments on the $[(CA^*G)_5]_2$ duplex where A*s represent ^{15}N6-labeled adenines. These experiments on measurements on the $[(CA^*G)_5]_2$ duplex and the

NOESY experiments on the [(CGCAGCG)]$_2$ duplex reveal single H-bonded A · A base pairing in both the duplexes.

A detailed analyses of NOESY data at 25, 50, 75, 100, 125, 200, and 500 ms of mixing show that the A · A mismatch in the middle of the (CGCAGCG)]$_2$ duplex disrupts neither the local B-DNA geometry nor the overall structure. Figure 41-14A shows the average of the 100 energy-minimized structures [(CGCAGCG)]$_2$. Gervais *et al.* [87] also made similar observations for an 11-base-pair duplex related to the K-ras gene. Note also that the distance between the hydrogen-bonded amino proton of one adenine and the H2 of the other is close in space (~2.7 Å), consistent with the observation of an NOE for N6H$_2$-H2 dipolar interaction.

Because of resonance overlap and low signal-to-noise-ratio, we have not been able to derive many independent distance constraints to determine the hairpin and duplex structures of (CAG)$_5$. Hence they have been modeled by using the same distance constraints determined for [(CGCAGCG)]$_2$. The duplex and the stem of the hairpin are divided into overlapping structural blocks on the basis of the 5-base-pair [(GCAGC)]$_2$ motif as observed in the middle of the 7-base-pair duplex. The intra- and sequential distance constraints of the individual blocks and the sequential distance constraints

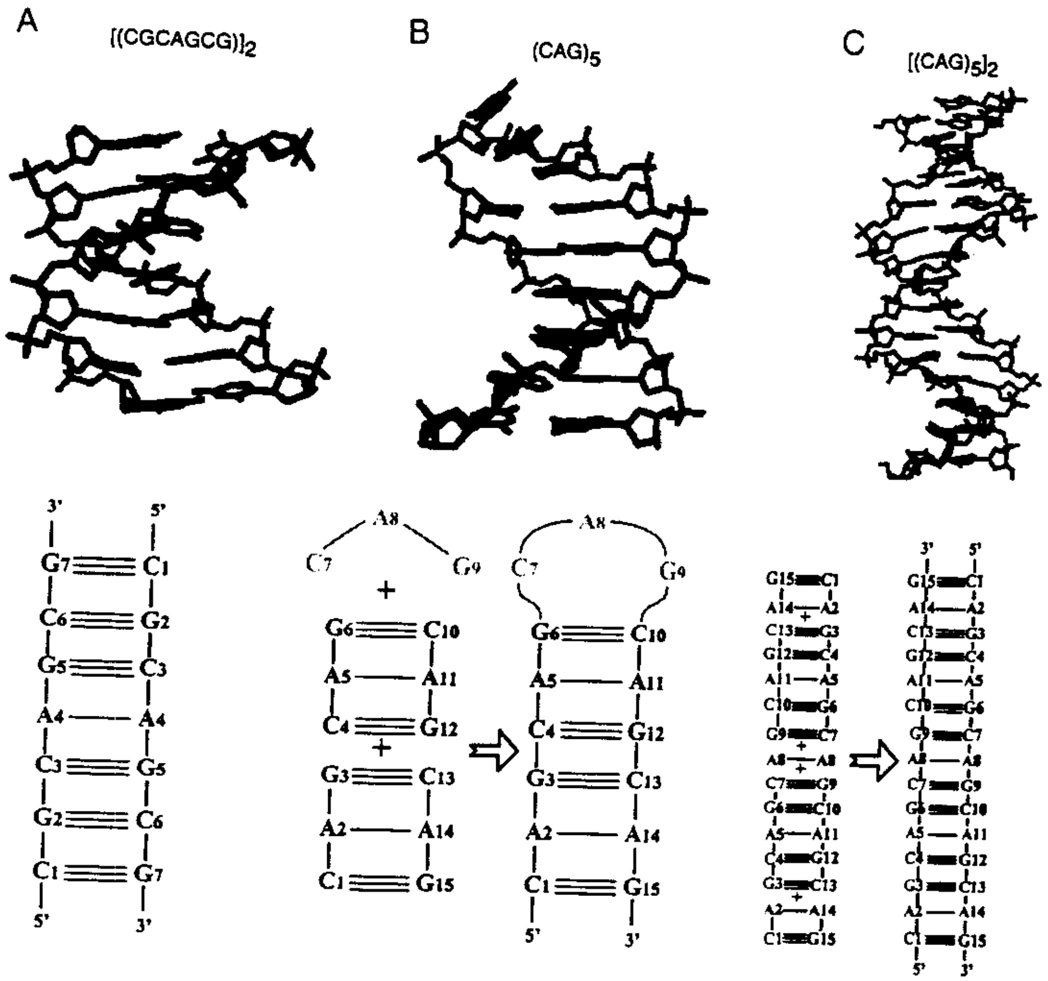

FIGURE 41-14 Energy minimized structures of (left) (CGCAGCG)$_2$, (middle) [(CAG)$_5$]$_2$, and (right) (CAG)$_5$ hairpin. The initial structures are constructed for duplex and hairpin in the same way as for (CGCAGCG)$_2$. Because of severe resonance overlap, we are unable to derive distance constraints either for the hairpin or for the duplex of (CAG)$_5$. The stem of the hairpin and the duplex are divided into overlapping motifs of (GCAGC)$_2$ duplex. Distance constraints for the 5-base-pair-long motif are from the NMR data of the duplex, (CGCAGCG)$_2$. The overlapping structural blocks defined by the 5-base-pair motif are shown below the structure of duplex and hairpin. No constraints are added for the interproton distances involving the loop nucleotides.

of the connecting steps of the individual blocks are the same as those of [(CGCAGCG)]$_2$. The distance constraints, assigned in this way together with the hydrogen bonding constraints, have been employed for the constrained minimization using AMBER (version 4.0); 20,000 conjugate gradient steps were used for minimization. Figures 41-14B and 41-14C show the duplex and hairpin structures of (CAG)$_5$ modeled by restrained minimization. In modeling the hairpin structures, no distance constraints have been imposed on the loop nucleotides. Similar structural features of the 5-base-pair structural motif present in the [(CGCAGCG)]$_2$ prevail in the duplex and the stem of the hairpin.

Based upon their replication and mismatch repair assays, Wells and co-workers [28] have suggested that the (CAG) repeats form less stable hairpins than the (CTG) repeats. It has been argued that the As in the (CAG)$_n$ hairpins are extrahelical, which according to Wells and co-workers is in agreement with the NMR data of Gao and co-workers [22]. However, the NMR data of Gao and co-workers show continuous sequential NOE connectivity in the (CAG)$_3$ duplex. This rules out the possibility of extrahelical As in the duplex. However, Gao and co-workers [22] did not identify the H-bonding nature of the two As facing each other in the duplex. Our NMR studies on the (CAG)$_5$ duplex and the model (CGCAGCG) duplex unequivocally show that the As in these duplexes are intrahelical and singly H-bonded.

B. Effect of Repeat Length on (CAG)$_n$ Structures

Nondenaturing gel electrophoresis, digestion by the single-strand-specific P1 nuclease, and 1D NMR studies have been carried out on (CAG)$_{10}$ and (CAG)$_{11}$ in order to determine the effects of repeat length, *n*, on the structures formed by the (CAG)$_n$ repeats. Under all solution conditions, (CAG)$_{10}$ and (CAG)$_{11}$ remain exclusively in the hairpin conformation. Figures 41-15A–41-15C show three possible structures of (CAG)$_{10}$ that would be consistent with the electrophoretic mobility data. Note that a (CAG)$_{10}$ hairpin with one single-stranded loop (Fig. 41-15A) has 9 G · C pairs (the terminal one is susceptible to end-fraying), whereas a (CAG)$_{10}$ hairpin with two single-stranded loops (Fig. 41-15C) has 8 G · C pairs (none is susceptible to end-fraying). We have carried out P1 digestions to probe the single-stranded regions in the (CAG)$_{10}$ and (CAG)$_{11}$ hairpins which reveal the presence of hairpin folding of (CAG)$_{10}$ and (CAG)$_{11}$ because substantial portions of these sequences remain undigested at a low enzyme to DNA ratio. However, what is more interesting is the observation of three protected fragments, i.e., digests of 22–24, 14–16, and 7–8 nucleotides for (CAG)$_{10}$ and digests of 23–25, 15–17, and 8–9 nucleotides for (CAG)$_{11}$. As shown in Fig. 41-15C, of the three fragments the longest one (over 22 nucleotides) is not expected if (CAG)$_{10}$ and (CAG)$_{11}$ form hairpins with one single-stranded loop. However, if (CAG)$_{10}$ and (CAG)$_{11}$ form hairpins with two single-stranded loops, fragments (>22 nucleotides) are expected after P1 digestion. Important evidence for the formation of double hairpin structures comes from the ^{1}H NMR spectra of (CAG)$_{10}$ and (CAG)$_{11}$ at different temperatures and pHs. In (CAG)$_{10}$, the presence of four G-imino signals within 12.0–13.0 ppm with intensity distribution (1:1:2:4) is consistent with 8 G · C base pairs. The G-imino resonances at 10.75 and 10.95 ppm disappear above 10°C and they are highly sensitive to pH. These two Gs may, therefore, belong to two different loops. Thus, the imino proton profile of (CAG)$_{10}$ is consistent with a hairpin containing two single stranded loops (Fig. 41-15C). The presence of two imino proton resonances for the guanines in the loop is also consistent with a (CAG)$_{10}$ hairpin with a single loop of 6 nucleotides (Fig. 41-15B). However, for such a hairpin only 7 (and not 8) G · C base pairs are expected.

Although, the 5′ and 3′ ends come close to each other in doubly looped (CAG)$_{10,11}$ hairpins, it does not mean that hairpin formation would have to be punctuated beyond *n* = 10/11 since several doubly looped hairpins of length 10 or 11 may be linked by double-stranded spacers as shown in Fig. 41-15D. Organization of several such 10- or 11-repeat-long hairpins may occur for longer repeat lengths. This would suggest that the stability of longer (CAG)$_n$ repeats largely depends upon the stability of individual doubly looped (CAG)$_{10,11}$ hairpins. Interestingly, Petruska *et al.* [34] made a similar observation in that the stability of (CAG)$_{10}$ is quite similar to that of (CAG)$_{30}$. Organization of several doubly looped (CAG)$_{10,11}$ hairpins may also describe the slippage structure during replication of long (CAG)$_n$ repeats (see Fig. 41-15D).

C. (CAG)$_n$ Hairpins Cause Slippage During Replication: An *in Vitro* Assay

As previously described in Fig. 41-8, we have used *in vitro* replication of M13 single-stranded DNA templates by *Taq* polymerase to assay for the intrinsic preference of hairpin formation by the (CAG)$_n$ strands in presence of its complementary Watson–Crick partner. Triplet repeats, (CAG)$_n$ or (CTG)$_n$ (*n* = 8 or 21), are inserted into the single-stranded M13 phage vectors (M13mp18 or M13mp19).

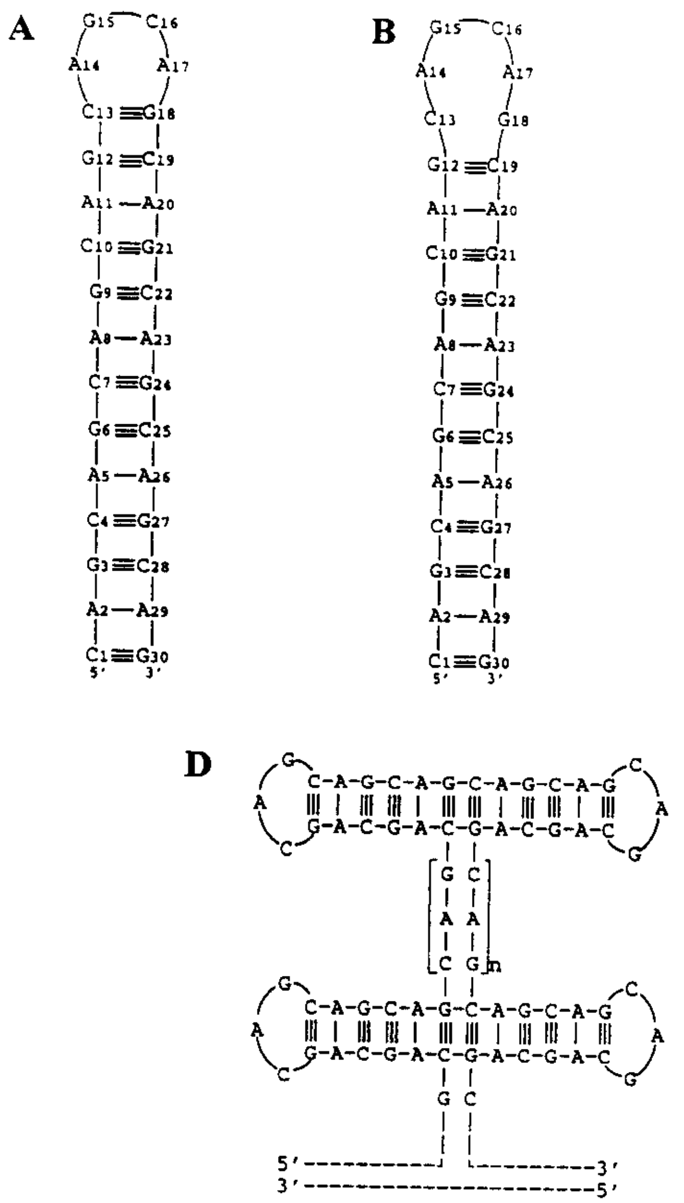

FIGURE 41-15 Three possible hairpin structures of $(CAG)_{10}$: (A) a singly folded hairpin with four nucleotides in the loop, (B) a singly folded hairpin with six nucleotides in the loop, and (C) a doubly folded hairpin with three nucleotides in each loop. The last loop folding is supported by gel electrophoresis, P1 digestion, and NMR data. Exchange properties of the loop imino signals are characterized by monitoring them at different pH and temperatures [46, 84-85]. $(CAG)_{11}$ also forms a doubly folded hairpin. (D) Possible arrangement of the doubly folded $(CAG)_n$ hairpins in a slippage structure.

½ / ?

In this replication assay, the majority of the replication product sequences belong to one size category. The full length of the insert and the flanking sequences are replicated for $(CAG)_8$ at a reaction temperature of 60°C, which should completely melt the $(CAG)_8$ hairpin. For $(CAG)_{21}$, 17 out of 21 repeats in the insert and the flanking sequences before and after two *Bam*H1 restriction sites are replicated, i.e., the remaining 4 repeats in the insert are bypassed. This observation is explained by the formation of a hairpin structure by the $(CAG)_{21}$

insert in the template. The extent of end-fraying increases with increasing temperature and as a result, the length of the central $(GCC)_{21}$ insert that is bypassed by *Taq* polymerase decreases with increasing temperature. In our replication assay, a repeat length (*n*) of 21 is required to demonstrate the formation of a hairpin although the individual $(CAG)_n$ strands showed exclusive presence of hairpins for $n = 10–11$. Therefore, the presence of the complementary strand and the polymerase pushes the critical threshold of *n* required for hairpin formation by the $(CAG)_n$ strand to a higher value. This threshold value of *n* may be higher when in addition to the polymerase all the replication accessory proteins (i.e., helicase, single-strand binding proteins) are also present during chain elongation.

V. FRIEDREICH'S ATAXIA (FRDA), TRIPLET REPEATS, $(GAA)_n/(TTC)_n$: CHARACTERIZATIONS OF THE TRIPLEX STRUCTURES BY NMR SPECTROSCOPY AND *IN VITRO* REPLICATION

Friedreich's ataxia (FRDA) is an autosomal recessive degenerative disease. The disease susceptible gene, c25, have recently been identified in the FRDA locus on 9q13-q21.1 [6, 88, 89]. c25 encodes for a 210-amino-acid-long protein called frataxin. FRDA is associated with the expansion of GAA/TTC triplet in the first intron of the c25 gene (Fig. 41-2). The number of GAA/TTC repeats in normal chromosomes varies between 7 and 22, whereas FRDA alleles ranges between 201 and 1186. Usually there is no detectable mRNA expression of the frataxin gene in FRDA alleles. Point mutations that either hinder intron–exon splicing or abruptly terminate the mRNA synthesis of X25 also lead to FRDA. This suggests that FRDA is a single-gene disorder.

Unlike the GC-rich $(CXG)_n$ triplet repeats which are capable of forming hairpin structures with X · X mispairs in the stem, the individual strands of $(GAA)_n/(TTC)_n$ do not form hairpin structures. Instead, they tend to form triplex structures as shown Fig. 41-2. Note that the folding of the longer TTC strand in this triplex leads to the formation of $C^+ \cdot G \cdot C$ and $T \cdot A \cdot T$ triads (Fig. 41-2). Here we discuss the determination of the high-resolution structure of such a triplex by homonuclear and heteronuclear NMR experiments. We also demonstrate by an *in vitro* replication assay the presence of similar triplexes with pyr · pur · pyr triads when long $(TTC)_n$ repeats (i.e., $n = 28–36$) are present in the template. In addition, by similar assays we show the presence of triplexes with pur · pur · pyr triads (see Fig. 41-2) when long $(GAA)_n$ repeats (i.e., $n = 28–36$) are present in the template although these triplexes are less stable than those with pur · pur · pyr triads.

Previously, Wells and co-workers performed systematic studies (reviewed in [90]) on various triplex-forming sequences including $(GAA)_n/(TTC)_n$. The triplex-forming ability of the $(GAA)_n/(TTC)_n$ repeat was studied by monitoring the induction of superhelicity in circular plasmids containing this repeat. It was shown that the $(GAA)_n/(TTC)_n$ repeat (for $n > 40$) formed a triplex with pur · pur · pyr triads even at neutral pH. Here we show that a triplex of $(GAA)_n/(TTC)_n$ containing as few as only six triads (two $C^+ \cdot G \cdot C$ and four $T \cdot A \cdot T$) is also stable at neutral pH.

A. Structural Characterization of a Triplex with pyr · pur · pyr Triads by NMR Spectroscopy

The possibility of a triplex structure for the $(GAA)_n/(TTC)_n$ repeat becomes obvious because a finite population of a triplex is observed even in the 1D NMR spectra of the exchangeable imino (NH) and amino (NH_2) protons of [1 : 1] $(GAA)_3 \cdot (TTC)_3$ at pH 7.0 and 15°C. The presence of the NH_2 signals from the protonated $C^+ \cdot G$ Hoogsteen pairs indicates the presence of $C^+ \cdot G \cdot C$ triads as expected in a triplex. In addition, the NH region is split into two groups, one belonging to the Watson–Crick duplex (the major population) and the other belonging to the triplex (the minor population). With (1 : 2) $(GAA)_3 \cdot (TTC)_3$ stoichiometry, the spectrum of the NH and NH_2 protons shows the exclusive presence of a triplex conformation at neutral pH. A triplex conformation is also exclusively present at neutral pH when $(GAA)_3$ is mixed with $(TTC)_7$ in equimolar amounts. A triplex structure for [1 : 1] $(GAA)_3 \cdot (TTC)_7$ implies that the $(TTC)_7$ strand probably folds with a T_2 loop. However, the spectral qualities in (1 : 2) $(GAA)_3 \cdot (TTC)_3$ and [1 : 1] $(GAA)_3 \cdot (TTC)_7$ do not allow the determination of high-resolution triplex structures by 2D NMR. We have, therefore, chosen an intramolecularly folded triplex, $(GAAGAA)T_4(TTCTTC)$-$T_4(CTTCTT)$, that models the folding of the $(TTC)_n$ strand with a T_4 loop (see Fig. 41-16 and [91]). In an actual system (see Fig. 41-2), either a T_2 or a T_2CT_2 loop is present. The spectral quality is dramatically improved in this unimolecular triplex.

A 295 interproton distance constraint has been derived by analyzing WATERGATE NOESY at 100 and 200 ms of mixing and NOESY in D_2O at 25, 50, 75, 100, 125, 200, and 400 ms of mixing. NOE data reveal that

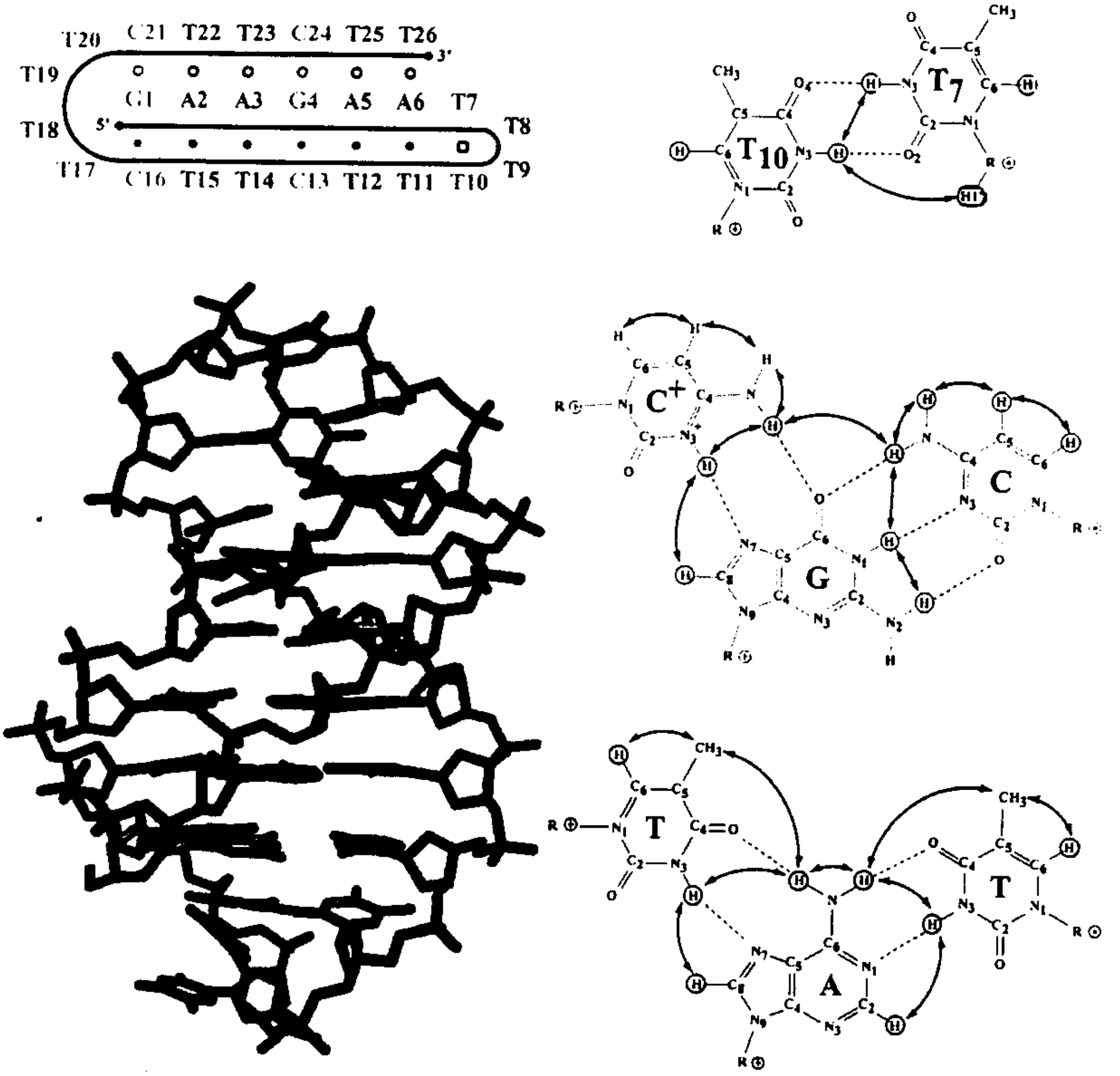
T20 C21 T22 T23 C24 T25 T26
3'
T19
G1 A2 A3 G4 A5 A6 T7
T18 5'
T8
T17 C16 T15 T14 C13 T12 T11 T10 T9
T7
T10
C+
G
C
T
A
T

FIGURE 41-16 Schematic chainfolding, descriptions of H-bonding in the T · T pair and in the C^+ · G · C and T · A · T triads, and three-dimensional average structures of the triple helix forming sequence $(GAA)_2T_4(TTC)_2T_4(CTT)_2$ derived using 1D/2D NMR spectroscopy. The stem of the triplex contains four T · A · T triads, and two C^+ · G · C. The stem is connected by two T_4 loops, each of them containing one T-T base pair. All nucleotides are in (*C2'-endo, anti*) conformations. Analyses of the 2D NMR data of $(GAA)_2T_4(TTC)_2T_4(CTT)_2$ lead to about 300 NOEs which are converted into distances using full-relaxation matrix analysis; 20 structures compatible with the distance constraints are generated. All the structures were found to belong to the same cluster with and average MSD of 0.52 Å^2.

all the nucleotides are in (*C2'-endo, anti*) conformations. We have also determined [91] the specific interactions of the amino protons belonging to the 4 Cs in this triplex. For this, we have ^{15}N4-labeled the 4 Cs in (GAAGAA)T_4(TTCTTC)T_4(CTTCTT) and performed (^{15}N-^{1}H-^{1}H) HMQC-NOESY experiments. Figure 41-16 shows the stereo view of the superimposition of 20 minimized structures. Figure 41-16 also shows the nature of $T_7 \cdot T_{10}$ pair that prevents the end-fraying of the T_{11} · A6 · T_{26} triad of the triplex. The (TTCTTC) arm involved in Watson–Crick pairing and the (CTTCTT) arm involved in Hoogsteen pairing show several interarm NOEs with the (GAAGAA) arm that uniquely lock the structure of the stem. Therefore, the structure in Fig. 41-16 represents a quantitative model for the triplex with pyr · pur · pyr triads that can be formed by the GAA/TTC repeats.

B. Identification of Triplexes with pyr · pur · pyr and pur · pur · pyr Triads by an *in Vitro* Replication Assay

Because our NMR studies unequivocally demonstrate that GAA/TTC repeats can form triplex structures, it is obvious that the formation of such structures during replication would cause slippage and replication. We have demonstrated the presence of such structures during replication by an *in vitro* replication using the single-stranded M13 DNA template. This assay is different from the one described in Fig. 41-8 which is performed in presence of *Taq* polymerase at various temperatures (45–80°C) and in absence of any replication accessory proteins. As shown in Fig. 41-17, in this room temperature assay, three different proteins (or protein assemblies) are used: (i) DNA polymerase that extends the primer on the single-stranded template, (ii) the single-strand binding protein (SSBP, such as the *E. coli* SSB or the human RP-A) that binds the polymerase on the 3′ side of the primer and tends to destabilize any secondary structure in the template, and (iii) the accessory and ATP-dependent protein complex that binds the polymerase on the 5′ side of the primer. This multicomponent protein–DNA complex ensures efficient chain elongation and processivity in DNA replication. Abnormal replication in this assay indicates that the replication machinery is unable to perform template-directed synthesis due to the presence of secondary (unusual) structure in the template. In fact, the nature of the replication product reflects the nature of the unusual structure present in the template. For example, a replication bypass is expected if a DNA repeat in the template forms a hairpin (as shown previously in Fig. 41-8). On the other hand if a DNA repeat in the template forms a triplex, a replication arrest is expected in the middle of the repeat (see Fig. 41-17). Finally, if a DNA repeat in the template forms a G-quartet, a replication arrest is expected in the beginning of the repeat. As previously stated, this assay has been successfully used to demonstrate the presence of (i) simple hairpins in the fragile X triplet repeats, (ii) triplex in FRDA triplet repeats, and (iii) hairpin G-quartet and hairpin i-motif in the insulin minisatellite [92–94].

As shown in Fig. 41-17, the slippage of $(TCC)_n$ in the template may cause the formation of a triplex with pyr · pur · pyr triads whereas the slippage of $(GAA)_n$ in the template may lead to the formation of a triplex with pur · pur · pyr triads. Since the triplex with pyr · pur · pyr triads contains $C^+ \cdot G \cdot C$, an acidic pH is likely to enhance the stability of such a triplex. Also in this triplex, the second pyr strand is located in the major groove of the pur strand of the Watson–Crick duplex and one of the H bonds in the major groove involves N7 of purines (see Fig. 41-2). Therefore, when $(TTC)_n$ is present in the template the use of 7-deaza GTP (instead of dGTP) in the precursor pool should eliminate the possibility of a H bond involving N7 of G and, therefore, should drastically lower the stability of the triplex with pyr · pur · pyr triads.

In the triplex with pur · pur · pyr triads (see Fig. 41-2), the second pur strand is located in the major groove of the first pur strand of the Watson–Crick duplex. Note that the formation of this triplex requires no protonation. However, here also N7 in the first pur strand is involved in H-bonding. Therefore, the presence of $(7\text{-deaza GAA})_n$ in the template instead of $(GAA)_n$ should eliminate one of the two H-bonding potentials of Gs in the major groove and should, therefore, drastically lower the stability of such a triplex.

The replication of $(TTC)_8$ in the M13 single-stranded DNA template gives a full replication product containing $(GAA)_8$ when the primer extension is carried out in the presence of the T_7 DNA polymerase within the pH range 5.5–8.0. Similar results are obtained with the T_4 DNA polymerase and the *E. coli* Klenow fragment. However, the replication pattern of $(TTC)_{36}$ in the template in the presence of the T_4 DNA polymerase, the T_7 DNA polymerase, or the *E. coli* Klenow fragment at pH 5.5–7.0 shows a strong replication arrest in the middle of the $(TCC)_{36}$ tract in the template (i.e., at $n = 18$). Similarly, when $(TCC)_{28}$ is present in the template, a strong replication arrest is observed at $n = 14$. Therefore, it appears that in our assay at least 28 repeats of (TTC) are necessary to cause the formation of a triplex during replication. The addition of either the *E. coli* SSB/human RP-A or the T_4 SSB (i.e., the product of gene 32) and the T_4 accessory proteins (i.e., the products of genes 44/62 and gene 45) to the replication assay releases the arrest for $(TTC)_{28,36}$ in the template. The replication arrest in the $(TTC)_{36}$ template is also released when 7-deaza G is used in the precursor pool. This is consistent with a triplex with a H bond through N7 of G in the major groove (see Fig. 41-2).

A replication arrest occurs in the middle of the $(GAA)_{28,36}$ tract (i.e., at $n = 14$ or 17) only for the T_4 polymerase. This is consistent with the formation of a triplex with pur · pur · pyr triads as shown in Fig. 41-2; such a triplex is detected in our replication assay only for $(GAA)_n$ templates with repeat number 28 or larger. Like the triplex with pyr · pur · pyr triads, this triplex is also unwound when either the *E. coli* SSB/human RP-A or the T_4 SSB plus other T_4 accessory proteins is

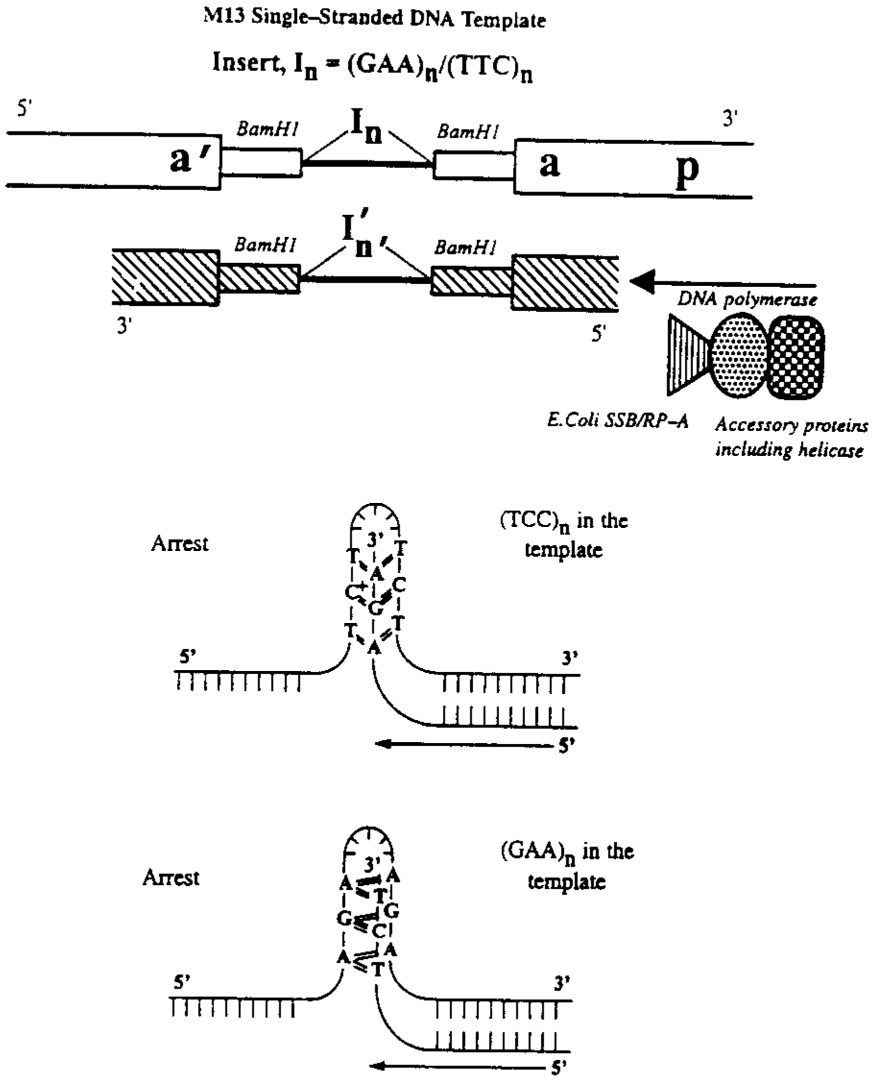
M13 Single-Stranded DNA Template
Insert, I_n = $(GAA)_n/(TTC)_n$
5'
3'
a'
BamHI
I_n
BamHI
a
p
I'_n
BamHI
n'
BamHI
3'
5'
DNA polymerase
E.Coli SSB/RP-A
Accessory proteins including helicase
Arrest
$(TCC)_n$ in the template
5'
3'
5'
Arrest
$(GAA)_n$ in the template
5'
3'
5'

FIGURE 41-17 An *in vitro* replication assay with $(GAA)_n$ or $(TTC)_n$ inserts in the single stranded M13 DNA templates. The folding of $(GAA)_n$ or $(TTC)_n$ creates a triplex with pur · pur · pyr or pyr · pur · pyr triads.

added during replication. The replication of (7-deaza GAA)$_{36}$ in the template shows no replication arrest. This indicates that in this triplex the Gs of the second pur strand is interacting with the Gs of the first pur strand through H bonds involving N7 (see Fig. 41-2). However, this triple helix is less stable than the triplex with pyr · pur · pyr triads since in the presence of the T$_7$ DNA polymerase no replication arrest is observed when (GAA)$_{28,36}$ are present in the template whereas as previously described strong replication arrests are observed for the same enzyme when (TTC)$_{28,36}$ are present in the template.

VI. CONCLUDING REMARKS

A. On the Mutagenic Unusual DNA Structures

Even before the discovery of the triplet related diseases, Sinden and Wells [95–97] proposed that unusual DNA structures may cause slippage during replication, resulting in deletions or expansions in the genomic DNA. After the discovery of the triplet-related diseases and the accumulation of a vast body of sequence, biochemical, and genetic data, the possibility of unusual

DNA structures as "dynamic mutagens" could be put to rigorous tests. In fact, Wells and co-workers [97] have attempted to explain these data associated with various triplets on the basis of their ability to form hairpin, triplex, or noodle DNA structures. In her review McMurray [11] provides key experimental support for the hypothesis that the molecular mechanism for (CXG) triplet-related neurological disorders is different from that for colon cancers [98–101]. In the triplet-related diseases, large expansions in the triplet repeats are caused by the formation of unusual DNA structures, whereas in colon cancers small deletions or expansions in the dinucleotide repeats are caused by the defects in the mismatch repair system. The experimental support comes from a combination of thermodynamic and genetic analyses [11]. It appears that the stabilities of the (CXG)$_n$ hairpins are length dependent. The stability vs length correlation agrees well with the disease susceptibility vs length correlation. This means that stable unusual DNA structures may form during replication or recombination and that the longer the triplet repeat, the higher the probability of the formation of unusual DNA structures. This structural feature of the triplet repeats is quite distinct from that of the dinucleotide repeats in colon cancer. In other words, slippage in the dinucleotide repeats may involve a small number of repeats. Perhaps such a slippage is efficiently corrected by the mismatch repair system in normal cells. However, in colon cancer cells a genetic defect in the mismatch repair system causes small deletions or expansions. Slippage of the triplet repeats, on the other hand, involves a large number of repeats which may not be corrected by the mismatch repair system. Indeed, it has been shown that the genetic instability in the triplet repeat can be blocked by the mismatch repair system only when he repeat length is small [102–104]. The repair system ecomes ineffectual when the repeat number of the riplets exceeds a certain critical threshold.

1. On the (CXG)$_n$ Hairpins

In these hairpins, the X · X mismatches are periodilly present in the stem of the hairpins. The pattern of se pairing remains the same for odd and even numbers repeats. This is brought about by the three-nucleotide op in the (CXG)$_n$ hairpin with an odd number of peats and the four-nucleotide loop in the (CXG)$_n$ irpins with an even number of repeats. Generally, the ger the repeat, the higher the probability of hairpin mation. However, different (CXG) repeats have different length requirements for hairpin formation. For mple, the individual strands of (CCG)$_n$ or (CTG)$_n$ form hairpins starting at $n = 5$ or 6, whereas the vidual strands of (CGG)$_n$ or (CAG)$_n$ form hairpins ting only at $n = 10$ or 11. It appears that when X is A or G, a homoduplex conformation of (CXG)$_n$ is preferred to a hairpin conformation for shorter ($n < 10$) repeat lengths. A larger repeat lengths, the equilibrium shifts toward the hairpin. *In vitro* replication assay also demonstrates that (CCG)$_n$ forms a hairpin at $n = 21$ whereas (CGG)$_n$ does not form a hairpin at $n = 21$. Note that in the presence of the complementary strand (i.e., in our replication assay), a larger repeat number is required for hairpin formation than for the individual strand, i.e., $n = 5$ is required for the individual (CCG)$_n$ strand whereas $n = 21$ is required for (CCG)$_n$ in the replication assay. Although in the replication assay we detect the presence of an unusual DNA structure in the template, the same structure can also be present in the growing chain during replication. For example, if in the presence of a hairpin the (CCG)$_n$ template causes a deletion then the formation of the same hairpin in the growing (CCG)$_n$ chain should cause expansion. This is exactly what Laird and co-workers [39] have observed in their replication assay with various DNA polymerases. In addition they have observed that the (CGG)$_n$ strand is less prone to expansion than the (CCG)$_n$ strand which is again consistent with our finding that the (CCG)$_n$ strand more readily forms a hairpin structure than the (CGG)$_n$ strand.

2. On the Triplexes Formed by the Friedreich's Ataxia Triplet Repeats

Our studies on the GAA/TTC repeats show the possibility of two types of triplexes: one with pyr · pur · pyr triads and the other with pur · pur · pyr triads. The triplex with pyr · pur · pyr triads is more stable than the one with pur · pur · pyr triads. Also in the triplex with pyr · pur · pyr triads the second pyr strand runs parallel in the major groove of the Watson–Crick duplex whereas in the triplex with pur · pur · pyr triads the second pur strand runs anti-parallel to the Watson–Crick duplex. In accordance with the data of McMurray and co-workers [26], we also find the possibility of a stable triplex with pur · pur · pyr triads when the second pur strand runs parallel to the Watson–Crick duplex. We are currently in the process of acquiring high-resolution NMR data of an intramolecular triplex with pur · pur · pyr triads. The formation of such a DNA–RNA triplex inside the intron with the GAA/TTC repeats may abruptly halt transcription. In addition, the intramolecular DNA triplex (GAAGAA)T$_4$(TTCTTC)T$_4$(CTTCTT) is a good model to study the effec sequence change in the triplex with pyr · pur · pyr tr It is shown that the analog (GAAGGA)T$_4$(TTCC $_4$(CTTCCT) which models the GAA to GGA tion, lowers the stability of the triplex at ne Note that the same substitutions in the m GAA repeat confer stability to the repea

Acknowledgments

This work was supported by the Human Genome Project of the Office of Health and Environmental Research (OHER) of the Department of Energy and OHER support to E. M. B. We thank Ms. Sue Thompson for synthesis and purification of various DNA oligomers used in this work. We thank Dr. Clifford Unkefer for maintenance of the 500 MHz Bruker DRX instrument.

References

1. Caskey, C. T., Pizzuti, A., Fu, Y-H., Fenwick, R. G., and Nelson, D. L. (1992). Triplet repeat mutations in human disease. *Science* **256,** 784–789.
2. Ross, C. A., McInnis, M. G., Margolis, R. L., and Li, S-H. (1993). Genes with triplet repeats: candidate mediators of neuropsychiatric disorders. *TINS* **16,** 254.
3. Bell, M. V., Hirst, M. C., Nakahori, Y., MacKinnon, R. N., Roche, A..,Flint, T. J, Jacobs, P. A., Tommerup, N., Tranebjaerg L., Froster-Iskenius, U., Kerr, B., Turner, G., Lindenbaum, R. H., Winter, R., Pembrey, M., Thibodeau, S., and Davies, K. E. (1991). Physical mapping across the Fragile X: hypermethylation and clinical expression of the Fragile X syndrome. *Cell* **64,** 861–866.
4. Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Marang, M., Barcelo, J., O'Hoy, K., Leblond, S., Earle-Macdonald, J., De Jong, P. J., Wieringa, B., and Korneluk, R. G. (1992). Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255,** 1253–1255.
5. Duyao, M., Ambrose, A., Myers, R., Novelleto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., Cray, J., Conneally, P., Young, A., Penney, J., Hollingsworth, Z., Shoulson, I., Lazzarini, A., Falek, A., Koroshetz, W., Sax, D., Bird, E., Vonsattel, J., Bonilla, Alvir, J., Conde, J. B., Cha, J-H., Dure, L., Gomez, F., Ramos, M., Sanchez-Ramos, J., Snodgrass, S., De Young, M., Wexler, N., Moscowitz, C., Penchaszadeh, G., MacFarlane, H., Anderson, M., Jenkins, B., Srinidhi, J., Barnes, G., Gusella, J., and MacDonald, M. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature Genet.* **4,** 387–397.
6. Campuzano, V., Montermini, L, Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Canizres, J., Koutnikova, H., Bidichandani, S. I., Gellera, C., Brice, A., Trouillas, P., De Michele, G., Filla, A., De Fruots, R., Palau, F., Patel, P. I., Di Donato, S., Mandel, J-L., Cocozza, S., Koenig, M., and Pandolfo, M. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271,** 1423–1427.
7. Richard, R. I., and Sutherland, G. R. (1994). Simple repeat DNA is not replicated simply. *Nature Genet.* **6,** 114–116.
8. Dover, G. (1995). Slippery DNA runs on and on and on . . . *Nature Genet.* **10,** 254–256.
9. Richards, R. I., and Sutherland, G. R. (1992). Dynamic mutations causing human disease. *Cell* **70,** 709–712.
10. Wells, R. D. (1996). Molecular basis of genetic instability of triplet repeats. *J. Biol. Chem.* **271,** 2875–2878.
11. McMurray, C. T. (1995). Mechanisms of DNA expansion. *Chromosoma* **4,** 2–13.
12. Fry, M., and Loeb, L. A. (1994). The Fragile X syndrome d(CGG)$_n$ nucleotide repeats form a stable tetrahelical structure. *Proc. Natl. Acad. Sci. USA* **91,** 4950–4954.
13. Chen, X., Mariappan, S. V. S., Catasti, P., Ratliff, R., Moyzis, R. K., Laayoun, A., Smith, S. S., Bradbury, E. M., and Gupta, G. (1995). Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. *Proc. Natl. Acad. Sci. USA* **92,** 5199–5203.
14. Mariappan, S. V. S., Chen, X., Castasti, P., Ratliff, R., Moyzis, R. K., Laayoun, A., Smith, S. S., Bradbury, E. M., and Gupta, G. (1996). Hairpin and junction structures of fragile X triplets. *In* "Proceedings of the 9th Conversation in Biomolecular Stereodynamics, June 1995, Albany, NY" (R. H. Sarama and M. H. Sarma, Eds.), pp. 105–119. Adenine Press.
15. Mitas, M., Yu, A., Dill, J., Kamp, T. J., Chambers, E. J., and Haworth, I. S. (1995). Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: (CTG)$_{15}$. *Nucleic Acids Res.* **23,** 1050–1059.
16. Yu, A., Dill, J., Wirth, S. S., Huang, G., Lee, V. H., Howorth, I. S., and Mitas, M. (1995). The trinucleotide repeat sequence d(GTC)$_{15}$ adopts a hairpin conformation. *Nucleic Acids Res.* **23,** 2706–2714.
17. Joworski, A., Rosche, W. A., Gellibolian, R., Kangk, S., Shimizu, M., Bowater, R. P., Sinden, R. R., and Wells, R. D. (1995). Mismatch repair in Escherichia coli enhances instability of (CTG)$_n$ triplet repeats from human hereditary diseases. *Proc. Natl. Acad. Sci. USA* **92,** 111019–11023.
18. Chastain, P. D., II, Eichler, E. E., Kang, S., Nelson, D. L., Levene, S. D., and Sinden, R. R. (1995). Anomalous rapid electrophoeretic mobility of DNA containing triplet repeats associated with human disease genes. *Biochemistry* **34,** 16125–16131.
19. Smith, G. K., Jie, J., Fox, G. E., and Gao, X. (1995). DNA CTG triplet repeats involved in dynamic mutations of neurologically related gene sequences form stable duplexes. *Nucleic Acids Res.,* **23,** 4303–4311.
20. Gao, X., Huang, X., Smith, G. K., Zheng, M., and Liu, H. (1995). New antiparallel duplex motif of DNA CCG repeats that is stabilized by extrahelical bases symmetrically located in the minor groove. *J. Am. Chem. Soc.* **95,** 1517–8883.
21. Mitas, M., Yu, A., Dill, J., and Haworth, I. S. (1995). The trinucleotide repeat sequence d(CGG)$_{15}$ forms a heat-stable hairpin containing G^{syn}G^{anti} base pairs. *Biochemistry* **34,** 12803–12811.
22. Aheng, M., Huang, X., Smith, G. K., Yang, X., and Gao, X. (1996). Genetically unstable CXG repeats are structurally dynamic and have a high propensity for folding. An NMR and UV spectroscopic study. *J. Mol. Biol.* **264,** 323–336.
23. Kettani, A., Kumar, R. A., and Patel, D. J. (1995). Solution structure of a DNA quadruplex containing the Fragile X syndrome triplet repeat. *J. Mol. Biol.* **95,** 2822–2836.
24. Gacy, A. M., Goellner, G., Juranic, N., Macura, S., and McMurray, C. T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* **8,** 533–540.
25. Goldberg, Y. P., McMurray, C. T., Zeisler, J., Almqvist, E., Sillence, D., Richards, F., Gacy, A. M., Buchanan, J., Telenius, H., and Hayden, M. R. (1995). Increased instability of intermediate alleles in families with sporadic Huntington disease compared to similar sized intermediate alleles in the general population. *Hum. Mol. Genet.* **4,** 1911–1918.
26. McMurray, C. T., Gacy, A. M., Goellner, G., Spiro, C., Dyer, R., Mikesell, M., Yao, J., Johnson, A. J., Juranic, N., Macura, S., Richter, A., and Melancon, S. B. (1997). DNA structures associated with class I expansion of GAA in Friedreich's ataxia. *In* "Proceedings of the 10th Conversation in Biomolecular Stereodynamics, June 1997, Albany, NY" (R. H. Sarama and M. H. Sarma, Eds.). Adenine Press. [In press]

27. Goellner, G. M., Tester, D., Thibodeau, S., Almqvist, E., Goldberg, Y. P., Hayden, M. R., and McMurray, C. T. (1997). Different mechanisms underlie DNA instability in Huntington's disease and colorectal cancer. *Am. J. Hum. Genet.* **60,** 879–890

28. Kang, S., Jaworski, A., Ohshima, K., and Wells, R. D. (1995). Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in E. coli. *Nature Genet.* **10,** 213–218.

29. Otten, A. D., and Tapscott, S. J. (1995). Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure. *Proc. Natl. Acad. Sci. USA* **92,** 5465–5469.

30. Wang, Y.-H., and Griffith, J. (1994). Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. *Genomics* **25,** 570–573.

31. Kang, S., Ohshima, K., Shimize, M., Amirhaeri, S., and Wells, R. D. (1995). Pausing of DNA synthesis in vitro at specific loci in CTG and CCG triplet repeats form human hereditary disease genes. *J. Biol. Chem.* **270,** 27104–27021.

32. Mariappan, S. V. S., Catasti, P., Chen, X., Ratliff, R., Moyzis, R. K., Bradbury, E. M., and Gupta, G. (1996). Solution structures of the individual single strands of the fragile X DNA triplets $(GCC)_n$- $(GGC)_n$. *Nucleic Acids Res.,* **24,** 784–792.

33. Mariappan, S. V. S., Garcia, A. E., and Gupta, G. (1996). Structure and dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy. *Nucleic Acids Res.* **24,** 775–783.

34. Petruska, J., Arnheim, N., and Goodman, M. F. (1996). Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nucleic Acids Res.* **24,** 1992–1998.

35. Nadel, Y., Weisman-Shomer, P., and Fry, M. (1995). The fragile X syndrome single strand $d(CCG)_n$ nucleotide repeats readily fold back to form unimolecular hairpin structures. *J. Biol. Chem.* **270,** 28970–28977.

36. Yu, A., Dill, J., and Mitas, M. (1995). The purine-rich trinucleotide repeat sequences $d(CAG)_{15}$ and $d(GAC)_{15}$ form hairpins. *Nucleic Acids Res.* **23,** 4055–4057.

37. Pearson, C. E., and Sinden, R. R. (1996). Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and Fragile X loci. *Biochemistry* **35,** 5041–5053.

38. Lian, C., Robinson, H., and Wang, A. H. J. (1996). Structure of ACTINOMYCIN D binding with (GAAGCTTC)2 and (GATGCTTC)2 and its binding to $(CAG)_n \cdot (CTG)n$ triplet sequence determined by NMR analysis. *J. Am. Chem. Soc.*

39. Ji, J., Clegg, N. J., Peterson, K. R., Jackson, A. L., Laird, C. D., and Loeb, L. A. (1996). In vitro expansion of GGC : GCC repeats: identification of the preferred strand of expansion. *Nucleic Acids Res,* **24,** 2835–2840.

40. Wang, Y.-H., Gellibolian, R., Shimizu, M., Wells, R. D., and Griffith, J. (1996). Long CCG triplet repeat blocks exclude nucleosomes: a possible mechanism for the nature of fragile sites in chromosomes. *J. Mol. Biol.* **263,** 511–516.

41A. Wang, Y.-H., and Griffith, J. (1996). Methylation of expanded CCG triplet repeat DNA from Fragile X syndrome patients enhances nucleosome exclusion. *J. Biol. Chem.,* **271,** 22937–22940.97.

41B. Wang, Y.-H., and Griffith J. (1995). Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. *Genomics* **25,** 570–573.

42. Schweitzer, J. K., and Livingston, D. M. (1997). Destabilization of CAG trinucleotide repeat tracts by mismatch repair mutations in yeast. *Hum. Mol. Genet.* **6,** 349–355.

43. Kang, S., Ohshima, K., Jaworski, A., and Wells, R. D. (1996). CTG triplet repeats from the myotonic dystrophy gene are expanded in Escherichia coli distal to the replication origin as a single large event. *J. Mol. Biol.* **258,** 543–547.

44. Rosche, W. A., Jaworski, A., Kang, S., Kramer, S. F., Larson, J. E., Geidroc, D. P., Wells, R. D., and Sinden, R. R. (1996). Single-stranded DNA-binding protein enhances the stability of CTG triplet repeats in Escherichia coli. *J. Bacteriol.* **178,** 5042–5044.

45. Ohshima, K., Kang, S., and Wells, R. D. (1996). CTG triplet repeats from human hereditary diseases are dominant genetic expansion products in Escherichia coli. *J. Biol. Chem.* **271,** 1853–1856.

46. Mariappan, S. V. S., Silks, L. A., III, Chen, X., Springer, P. A., Wu, R., Moyzis, R. K., Bradbury, E. M., Garcia, A. E., and Gupta, G. (1997). Solution structure of the Huntington's disease DNA triplets, $(CAG)_n$. *J. Biol. Struct. Dynamics.* [in press]

47. Brahmachari, S. K., Meera, G., Sarkar, P. S., Balugurumoorthy, B., Tripathi, J., Raghavan, S., Shaligram, U., and Pataskar, S. S. (1995). Simple repetitive sequences in the genome: structural and functional significance. *Electrophoresis* **16,** 1705–1714.

48. Brahmachari, S. K., Sarkar, P. S., Shaligram, U., Matsuddi, M., Raghavan, S., Bhandari, Pataskar, S. S., Narayan, M., and Quasar, S. P. (1997). Genome instability: structural basis of triplet repeat expansion and genetic disorder. *In* "Proceedings of the 10th Conversation in Biomolecular Stereodynamics, June 1997, Albany, NY" (R. H. Sarama and M. H. Sarma, Eds.). Adenine Press. [In press]

49. Pieretti, M., Zhang, F., Fu, Y.-H., Warren, S. T., Oostra, B. A., Caskey, C. T., and Nelson, D. L. (1991). Absence of expression of the FMR-1 gene in fragile X syndrome. *Cell* **66,** 817–822.

50. Laird, C., Jaffe, E., Karpen, G., Lamb, M., and Nelson, R. Fragile sites in human chromosomes as regions of late-replicating DNA. *TIG* **3,** 274–280.

51. Laird, C. D. (1987). Proposed mechanism of inheritance and expression of the human fragile-X syndrome of mental retardation. *Genetics* **117,** 587–599.

52. Oberle, I. Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer., A., Boue, J., Bertheas, M. F. and Mandel, J. L. (1991). Instability of a 550-base pair DNA segement and abnormal methylation in fragile X syndrome. *Science* **252,** 1097–1102.

53. Gecz, J. Gedeon, A. K., Sutherland, G. R., and Mulley, J. C. (1996). Identification of the gene FMR2, associated with FRAXE mental retardation. *Nature Genet.* **13,** 105–110.

54. Eichler, E. E., Holden, J. J. A., Popovich, B. W., Reiss, A. L., Snow, K. Thibodeau, S. N., Richards, C. S., Ward, P. A., and Nelson, D. L. (1994). Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nature Genet.* **8,** 88–92.

55. Cognet, J. A. H., Gabarro-Arpa, J., Bret, M. L., van der Marel, G. A., van Boom, J. H., and Fazakerley, G. V. (1991). Solution conformation of an oligonucleotide containing G · G mismatches determined nuclear magnetic resonance and molecular mechanics, *Nucleic Acids Res.* **19,** 6771–6779.

56. Yu, A., Barron, M. D., Romero, R. M., Christy, M., Gold, B., Dai, J., Gray, D. M., Haworth, I. S., and Mitas, M. (1997). At physiological pH, d(CCG)15 forms a hairpin containing protonated cytosines and a distorted helix. *Biochemistry* **36,** 3687–3699.

57. Smith, S. S., Kaplan, B. E., Sowers, L. C., and Newman, E. M. (1992). Mechanism of human methyl-directed DNA methyltransferase and the fidelity of cytosine methylation. *Proc. Natl. Acad. Sci. USA* **89,** 4744–4748.

58. Baker, D. J., Kan, J. L. C., and Smith, S. S. (1988). Recognition of structural perturbations in DNA by human DNA (cytosine-5) methyltransferase. *Gene* **74,** 207–210.

59. Baker, D. J., Laayoun, A., and Smith, S. S. (1993). Transition state analogs as affiity labels for human DNA methyltransferases. *Biochem. Biophys. Res. Commun.* **196,** 864–871.

60. Posfai, J., Bhagwat, A. S., Posfai, G., and Roberts, R. J. (1989). Predictive motifs derived from cytosine methyltransferases. *Nucleic Acids Res.* **17,** 2421–2435.

61. Klimassauskas, S., Kumar, S., Roberts, R. J., and Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix. *Cell* **76,** 357–369.

62A. Roberts, R. J. (1995). On base flipping. *Cell* **82,** 639–645.

62B. Chen, X., Mariappan, S. V. S., R., Moyzis, R. K., Bradbury, E. M., and Gupta, G. (1997). Hairpin induced slippage and hypermethylation of the fragile X DNA triplets. *J. Biomol. Struct. Dynamics.* [in press]

63A. Depamphilis, M. L., and Wasserman, P. M. (1980). Replication of eukaryotic chromosomes: a close-up view of the replication fork. *Annu. Rev. Biochem.* **49,** 627–666.

63B. Stillman, B. (1994). Smart machines at the DNA replication fork. *Cell* **78,** 725–728.

64. Zhao, J., Cheng, W., Gibb, C. L. D., Gupta, G., and Kallenbach, N. R. (1996). HMG box proteins interact with multiple tandemly repeated $(GCC)_n \cdot (GGC)n$ DNA sequences. *J. Biomol. Struct. Dynamics* **14,** 235–238.

65. Anteguera, F., and Bird, A. (1993). Number of Cp'G island and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* **90,** 11995–11999.

66. Bird, A. P. (1986). CpG-rich island and the function of DNA methylation. *Nature* **321,** 209–213.

67. Boyes, J., and Bird, A. (1992). Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J.* **11,** 327–333.

68. Meehen, R. R., Lewis, J. D., McKay, S. Kleiner, E. L., and Bird, A. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **58,** 499–507.

69. Lewis, J. D., Meehan, R. R., Henzel, W. J., Maurer-Fogy, I., Jappesen, P., Klein, F., and Bird, A. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* **69,** 905–914.

70. Lie, L. F., and Wang, J. C. (1987). Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci. USA* **84,** 7024–7027.

71. Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'Hoy, K., Leblond, S., Earle-Macdonald, J., De Jong, P. J., Wieringa, B., and Korneluk, R. G. (1992). Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255,** 1253–1255.

72. Fu, Y.-H., Pizzuti, A., Fenwick, R. G., Jr., King, J., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., De Jong, P., Wieringa, B., Korneluk, R., Perryman, M. B. , Epstein, H. F., and Caskey, C. T. (1992). An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* **255,** 1256–1258.

73. Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J., Hudson, T., Sohn, R., Zemelman, B., Snell, R. G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P. S., Shaw, D. J., and Housman, D. E. (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* **68,** 799–808.

74. Briat, J-F., Bollag, G., Kearney, C. A., Molineuz, I., and Chamberlin, M. J. (1987). Tau factor from Escherichia coli mediates accurate and efficient termination of transcription at the bacteriophage T3 early termination site in vitro. *J. Mol. Biol.* **198,** 43–49.

75. Eckner, R., and Birnstiel, M. (1992). Evolutionary conserved multiprotein complexes interact with the 3' untranslated region of histone transcripts. *Nucleic Acids Res.* **20,** 1023–1030.

76. Taneja, K. L., McCurrach, M., Schalling, M., Housman, D., and Singer, R. H. (1995). Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and nuclei. *J. Cell Biol.* **128,** 995–1002.

77. Kouchakdjian, M., Li, B. F. L., Swan, P. F., and Patel, D. J. (1988). Pyrimidine · pyrimidine base-pair mismatches in DNA: a nuclear magnetic resonance study of T · T pairing at neutral pH and C · C pairing at acidic pH in dodecnucleotide DNA duplexes. *J. Mol. Biol.* **202,** 139–155.

78. Lane, A. N., and Forster, M. J. (1989). Determination of internal dynamics of deoxyriboses in the DNA hexamer d(CGTACG)2 by 1H NMR. *Eur. Biophys. J.* **17,** 221–232.

79. Lipari, G., and Szabo, A. (1982). Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.* **104,** 4546–4558.

80. Lipari, G., and Szabo, A. (1982). Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *J. Am. Chem. Soc.* **104,** 4559–4570.

81. Krahe, R., Ashizawa, T., Abbruzzese, C., Roeder, E., Garango, P., Giacaelli, M., Funanage, V., and Siciliano, M. J. (1995). Effect of myotonic dystrophy trinucleotide repeat expansion on DMPK transcription and processing. *Genomics* **28,** 1–14.

82. The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72,** 971–983.

83. Leeflang, E. P., Zhang, L., Tavare, S., Hubert, R., Srinidhi, J., MacDonald, M. E., Myers, R., de Young, M., Wexler, N. S., Gusella, J. F., and Arnheim, N. (1995). Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency spectrum. *Hum. Mol. Genet.* **4,** 1519–1526.

84. van de Ven, F. J. M., and Hilbers, C. W. (1988). Nucleic acids and nuclear magnetic resonance. *Eur. J. Biochem.* **270,** 1–38.

85. Orbons, L. P. M., van der Marel, G. A., van Boom, J. H., and Altona, C. (1987). An NMR study of the polymorphous behavior of the mismatched octamer d(m^5C-G-m^5C-G-T-G-m^5C-G) in solution: the B, Z, and hairpin forms. *J. Biomol. Structure Dynamics* **4,** 939–963.

86. Ikuta, S., Takagi, K., Wallace, R. B., and Itakura, K. (1987). Dissociation kinetics of 19 base-paired oligonucleotide-DNA duplexes containing different single mismatched base pairs. *Nucleic Acids Res.* **15,** 797–811.

87. Gervais, V., Cognet, J. A. H., Le Bret, M., Sowers, L. C., and Fazakerley, G. V. (1995). Solution structure of two mismatches A · A and T · T in the K-ras gene context by nuclear magnetic resonance and molecular dynamics. *Eur. J. Biochem.* **228,** 279–290.

88. Cossee, M., Schimtt, M., Campuzano, V., Reutenauer, L., Moutou, C., Mandel, J. L., and Koenig, M. (1997). Evolution of the Friedreich's ataxia trinucleotide repeat expansion. *Proc. Natl. Acad. Sci. USA* **94,** 7452–7457.

89. Bidichandani, S. I., Ashizawa, T., and Patel, P. (1997). Atypical Friedreich's ataxia caused by compound heterozygosity for a

novel missense mutation and the GAA-triplet repeat expansion. *Am. J. Hum. Genet.* **60,** 1251–1256.

90A. Filla, A., de Michele, G., Cavalcanti, F., Pianese, L., Monticelli, A., Campanella G., and Cocozza, S. (1996). The relationship between trinucleotide (GAA)repeat length and clinical features in Friedreich ataxia. *Am. J. Hum. Genet.* **59,** 554–560.

90B. Wells, R. D., Collier, D. A., Hanvey, J. C., Shimizu, M., and Wohlrab, F. (1988). The chemistry and biology of unusual DNA structures adopted by oligopurine · oligopyrimidine sequences. *FASEB J.* **2,** 2939–2949.

91. Mariappan, S. V. S., Catasti, P., Silks, L. A., Bradbury, E. M., and Gupta, G. (1997). High resolution structure of triplex formed by the GAA/TTC triplets associated with the Friedreich's Ataxia. [Submitted for publication]

92. Baran, N., Lapidot, A., and Manor, H. (1991). Formation of DNA triplexes accounts for arrests of DNA synthesis at d(TC) and d(GA) tracts. *Proc. Natl. Acad. Sci.USA* **88,** 507–511.

93. Catasti, P., Chen, X., Moyzis, R. K., Bradbury, E. M., and Gupta, G. (1997). Structure-function correlations of the insulin-linked polymorphic region. *J. Mol. Biol.* **263,** 534–545.

94. Catasti, P., Chen, X., Deaven, L. L., Moyzis, R. K., Bradbury, E. M., and Gupta, G. (1997). Cytosine-rich strands of the insulin minisatellite adopt hairpins with intercalated cyotosine+ · cytosine pairs. *J. Mol. Biol.* **272,** 369–382.

95. Wells, R. D. (1988). Unusual DNA structures. *J. Biol. Chem.* **263,** 1095–1098.

96. Sinden, R., R., and Wells, R. D. (1992). DNA structure, mutations, and human genetic disease. *Biotech.* **3,** 612–622.

97. Trinh, T. Q., and Sinden, R. R. (1991). Preferential DNA secondary structure mutagenesis in the lagging strand of replication in E. coli. *Nature* **352,** 544–547.

98. Thibodeau, S. N., Bren, G., and Schaid, D. (1993). Microsatellite instability in cancer of the proximal colon. *Science* **260,** 816–819.

99. Fishel, R., Lescoe, M. K., Rao, M. R. S., Copeland, N. G., Jenkins, N. A., Garber, J., Kane, M., and Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75,** 1027–1038.

100. Bonner, C. E., Baker, S. M., Morrison, P. T., Waren, G., Smith, L. G., Lescoe, M. K., Kane, M., Erabino, C., Lipford, J., Lindblom, A., Tannergard, Pl., Bollag, R. J., Godwin, A. R., Ward, D. C., Nordenskjeid, M., Fishel, R., Kolodner, R., and Liskay, R. M. (1994). Mutation in the DNA mismatch repair gene homologue hMLH1 is assosicated with hereditary non-polyposis colon cancer. *Nature* **368,** 258–261.

101. Papadopoulos, N, Nicolaides, N. C., Wei, Y-F., Ruben, S. M., Carter, K. C., Rosen, C. A., Haseltine, W. A., Fleischmann, R. D., Fraser, C. M., Adams, M. D., Venter, J. C., Hamilton, S. R., Petersen, G. M., Watson, P., Lynch, G. T., Peltomaki, P., Mecklin, J-P., de la Chapelle, A., Kinzler, K. W, and Vogelstein, B. (1994). Mutation of a *mutL* homolog in hereditary colon cancer. *Science* **263,** 1625–1629.

102. Kramer, P. R., Pearson, C. E., and Sinden, R. R. (1996). Stability of triplet repeats of myotonic dystrophy and fragile X loci in human mutator repair cell lines. *Hum. Genet.* **98,** 151–157.

103. Fishel, R., Ewel, A., Lee, S., Lescoe, M. K., and Griffith, J. (1994). Binding of mismatched microsatellite DNA sequences by the human MSH2 protein. *Science* **266,** 1403–1405.

104. Schweitzer, J. K., and Livingston, D. M. (1997). Destabilization of CAG trinucleotide repeat tracts by mismatch repair mutations in yeast. *Hum. Mol. Genet.* **6,** 349–355.

105. Mariappan, S. V. S., Silks, L. A., Bradbury, E. M., and Gupta, G. (1997). Hairpins of fragile X GCC DNA triplet form single hydrogen-bonded C · C mispairs: selective ^{15}N4-labeled cytosines and isotope-edited nuclear magnetic resonance spectroscopy. [Submitted for publication]

106. Godde, J. S., and Wolffe, A. P. (1996). Nucleosome assembly on CTG triplet repeats. *J. Biol. Chem.* **271,** 15222–15229.

Reversible Histone Modifications and the Chromosome Cell Cycle

E. Morton Bradbury

Summary

During the eukaryotic cell cycle, chromosomes undergo large structural transitions and spatial rearrangements that are associated with the major cell functions of genome replication, transcription and chromosome condensation to metaphase chromosomes. Eukaryotic cells have evolved cell cycle dependent processes that modulate histone:DNA interactions in chromosomes. These are; i) acetylations of lysines; ii) phosphorylations of serines and threonines and iii) ubiquitinations of lysines. All of these reversible modifications are contained in the well-defined very basic N- and C-terminal domains of histones. Acetylations and phosphorylations markedly affect the charge densities of these domains whereas ubiquitination adds a bulky globular protein, ubiquitin, to lysines in the C-terminal tails of H2A and H2B. Histone acetylations are strictly associated with genome replication and transcription; histone H1 and H3 phosphorylations correlate with the process of chromosome condensation. The subunits of histone H1 kinase have now been shown to be cyclins and the $p34^{CDC2}$ kinase product of the cell cycle control gene CDC2. It is probable that all of the processes that control chromosome structure:function relationships are also involved in the control of the cell cycle.

Introduction

The cell division cycle constitutes a series of interrelated processes that have evolved to create two genetically identical daughter cells from a mother cell. A major function of this cycle is the faithful replication of the genome and its packaging into chromosomes. During the cell cycle, chromosomes undergo major structural transitions ranging from the more dispersed functional states in S phase to the fully condensed inactive state of metaphase chromosomes. The factors that control these transitions most probably have major involvements in the molecular control of the cell cycle.

The magnitude of the problem of understanding the structures and functions of mammalian chromosomes can be appreciated from the fact that the diploid human genome contains more than 200 cm of DNA molecules packaged into 46 chromosomes, each several μm's in length, contained in a cell nucleus about 5 μm radius. Although chromosomes have been studied intensively for more than a century, we have only a sketchy understanding at the molecular level of their organization and different structural states. This situation reflects the inherent difficulties of analyzing the complex hierarchy of protein:DNA interactions involved in packaging DNA molecules into chromosomes. The first level of complexity involves histone: DNA interactions that must be able to accommodate a multitude of different DNA sequences, with each sequence differing in its physical property of flexibility. Subsequent interactions are required to generate the higher orders of chromatin structure found in metaphase chromosomes. Furthermore, specific DNA sequences most probably determine or influence the long range organization of DNA in chromosomes. It is to be hoped that one outcome of extensive human genome sequencing will be the identification of the sequences involved in DNA packaging and chromosome organization. Solutions of these complex structural problems are central also to our understanding of the major cell cycle events of replication, transcription and the process of condensation to the metaphase state. In this article, I will discuss our current understanding of the chromatin structure/function relationships of the cell cycle-dependent reversible chemical modifications of histones; acetylations, phosphorylations and ubiquitinations. Lysine residues in histones are subjected also to irreversible methylations of unknown function and to levels of polyadenosine diphosphate ribosylation, which are vanishingly low in undamaged cells.

Long Range Order in Chromosomes

The special case of polytene chromosomes with their bands, interbands and puffs of active regions provides visual evidence of long range organization in chromosomes. Long range order exists also in mammalian chromosomes. The gentle removal of the histones and all but the most tightly bound of the chromosomal proteins from human metaphase chromosomes reveals in electron micrographs a residual proteinaceous scaffold of the original chromosome which constrains a 'halo' of DNA loops[1]. These observations and biochemical evidence for long range order in the chromatin of interphase cells[2] have led to the proposal of a DNA loop or chromatin domain model for the organization of DNA in chromosomes with each DNA loop containing a gene or small set of linked genes. DNA loop sizes from a variety of studies have been estimated to range from 5 to 200 kbp with an average size of 50 kbp (reviewed in ref. 3). Thus, the human haploid genome of 3×10^9 bp DNA would contain 60 000 loops, a number comparable to the 50 000 to 100 000 genes thought to be contained in the human genome. The proteinaceous scaffold comprises 2% of the chromosomal proteins and contains two major scaffold proteins Sc1 and Sc2 that constrain the DNA loops at their bases[4]. It is significant that Sc1 has been identified[5] as topoisomerase II, an enzyme that relaxes both negatively and positively supercoiled DNA. Further yeast genetic studies have shown that topoisomerase II is essential in mitosis for the separation of daughter chromosomes[6–8]. Recent studies from my laboratory[9] and other laboratories[10–12] also implicate topoisomerase II in the physical process of chromosome condensation in mammalian cells.

Histones

If DNA is constrained by loop attachment or any other mechanism, then the state of DNA topology becomes an important consideration in the understanding of chromosome functions[13]. Histones are of central importance because they package eukaryotic DNA into nucleosomes and several higher orders of chromatin structures. An essential role for histones in the structure/function relationships of chromosomes is implied by the rigid sequence conservation of the entire histones H3 and H4 and of the well-defined globular domains of the other histones H1, H2A and H2B. However, histone diversity is also of functional importance and families of histone subtypes that have been found for histones H1, H2A and H2B provide the potential to introduce considerable variability into the physical packaging of genes in DNA loops in chromosomes of different cell types; for example, mammalian cells contain six subtypes of histone H1 (reviewed in ref. 14). Additional variability derives from the reversible chemical modification of histones that change markedly the nature of the modified amino acid. The major reversible modifications of histones that are cell cycle dependent are phosphorylation, acetylation and ubiquitination.

NMR spectroscopy[15] and controlled proteolysis[16] of specific histone complexes, nucleosomes and chromatin have demonstrated that histones are multidomain proteins. In isolated complexes in solution, H3 and H4 have well-defined, basic, flexible N-terminal domains extending from defined apolar globular structures; H2A and H2B have variable flexible N and C-terminal domains and conserved central globular domains, and H1 has a similar conformation, but with longer extended flexible N and C terminal domains. It is significant that *all* of the sites of reversible chemical modifications of histones, acetylation, phosphorylation and ubiquitinations, are located in these flexible basic domains. The specific histone complexes, the dimer [H2A, H2B], tetramer [$H3_2$, $H4_2$] and octamer [(H2A, $H2B)_2(H3_2$. $H4_2$)] are held together by interactions between the apolar globular domains. Cartoon models for the histone dimer and tetramer and H1 showing the different domains and sites of reversible modifications are given in Fig. 1. Clearly these reversible histone modifications that affect the charge density of the flexible N and C-terminal domains or introduce bulky ubiquitin moieties have considerable potential to modulate histone:DNA interactions in chromatin.

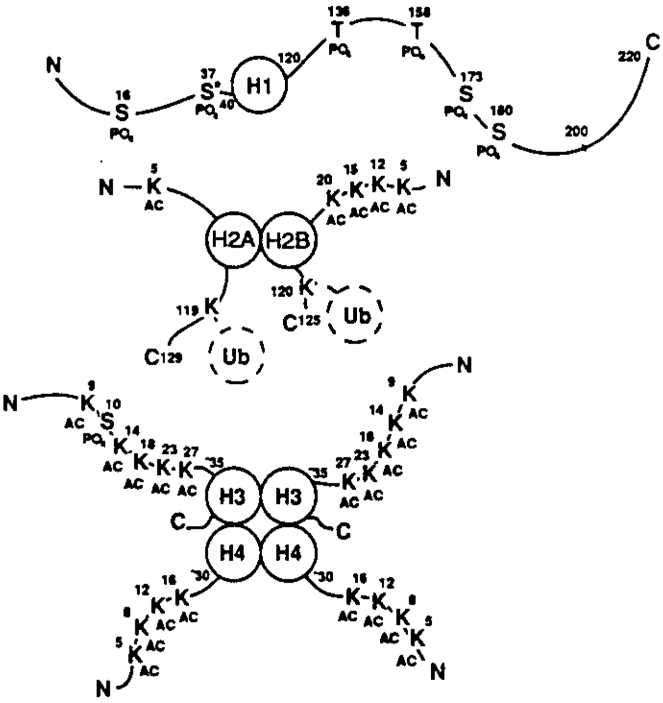

Fig. 1. Outline structures of histone H1, and the (H2A, H2B) dimer and ($H3_2$, $H4_2$) tetramer showing the well-defined globular domains, the basic flexible N and C-terminal domains and sites of reversible acetylation, phosphorylation and ubiquitinations.

Nucleosome and Chromatin Structures

The first level of histone:DNA interactions generates the basic structural unit of eukaryotic chromosomes, the nucleosome. Virtually all of the genome is packaged into nucleosomes. Nucleosomes from most cells of higher eukaryotes contain 195±5 bp DNA, the histone octamer [(H2A, $H2B)_2(H3_2,H4_2)$] and one H1 molecule. Nuclease digestion of nucleosomes reveals two well-defined subnucleosome particles; the chromatosome with 168 bp DNA and the full complement of histones and the nucleosome core particle with 146 bp DNA and the histone octamer (reviewed in ref. 3). Neutron scatter studies of the low resolution structure of the core particle in solution (reviewed in ref. 17) reveal it to be a flat disc 11.0 nm diameter, 5.5–6.0 nm thick with 1.7±0.1 turns of DNA of pitch 3.0 nm coiled on the outside of the histone octamer; this structure is the same as the crystal structure now solved to 0.7 nm[18] and 0.8 nm[19] resolutions. At these higher resolutions, the DNA is seen to be not uniformly bent around the octamer, but to follow a path of gentle curves and tighter bends. It is to be noted that not all of the histone octamer electron density is accounted for by the core particle crystal structure[10]. This has been attributed to disorder in the flexible N-terminal domains, probably because their *in vivo* sites of interactions lie outside of the 146 bp DNA core particle. These N-terminal domains can be proteolysed away

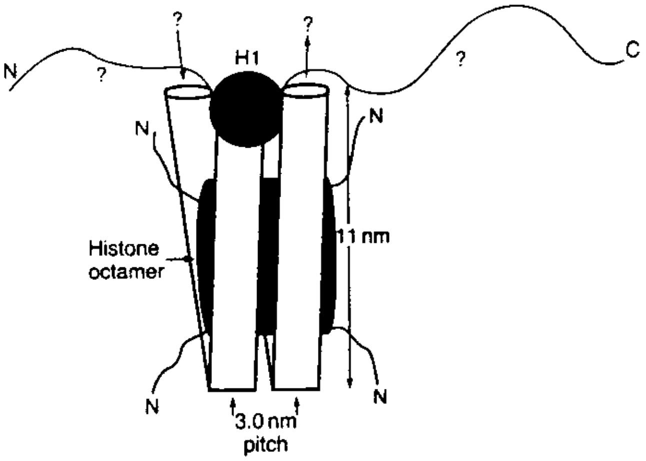

Fig. 2. Model for the chromatosome based on the known structure of the nucleosome core particle and the probable binding probe of histone H1.

from core particles with little effect on core particle structures[16]. Thus, the conserved core particle appears to be stabilized by interactions between the 146 bp DNA and the rigidly conserved H3 and H4 and conserved globular domains of H2A, H2B. At this time the modes of interactions of the flexible basic N and C-terminal domains in nucleosomes and chromatin are not understood. Although these regions are depicted as random coils in isolated histone complexes (Fig. 1), they could be partially helical when complexed with the phosphate groups of DNA because of charge neutralization of lysines and arginines[20]. Such interactions, however, must be modulated by the reversible chemical modifications.

Extrapolating from the 1.7–1.8 turns of 146 bp of DNA of pitch 3.0 nm coiled around the core particle, then the 168 bp of DNA in the chromatosome corresponds to 2.0 turns of DNA that are sealed off by the binding of the central globular domains of the fifth histone H1, as shown in the model given in Fig. 2. The long flexible 'arms' of the H1 molecule clearly have the potential to be involved in long range interactions in chromosomes.

At low ionic strength, chromatin unfolds and forms an 11 nm diameter fibril consisting of a string of nucleosome discs arranged roughly edge to edge (reviewed in 17). The mass/unit length gives a DNA packing ratio of 6 to 7 to 1, which corresponds to a nucleosome every 10±2 nm. With increase of ionic strength, the 11 nm fibril makes a transition to a '30 nm' fibril. Most of the DNA in interphase and metaphase chromosomes is packaged into 30 nm fibrils. The diameter of the hydrated form of this fibril from chicken erythrocytes is 34 nm and its mass/unit length corresponds to 6 to 7 nucleosomes per turn of a supercoil of pitch 11.0 nm, i.e. a DNA packing ration of 40 to 50:1 (reviewed in refs 3 and 17). Because of the paucity of hard structural data, a range of models have been proposed for this 34 nm fibril (reviewed in ref. 3). Basic questions that still need to be answered are the location of histone H1 and internucleosomal linker DNA.

DNA Supercoiling

In bacteria the state of DNA topology in the nucleoid is maintained by the balance of activities of gyrases which increase DNA supercoiling, and topoisomerases, which relax supercoiled DNA (reviewed in ref. 21). Despite many efforts, however, the eukaryotic homologue of the bacterial gyrase has not been found. A possible reason for this is that the evolution of the nucleosome provided other mechanisms for the control of DNA supercoiling in eukaryotic chromosomes. This is evident from the structures of the core particle and nucleosome. The core particle contains 1.7±0.1 turns of DNA coiled around the histone octamer. Addition of H1 to form the nucleosome increases the number of turns of DNA coiled around the octamer from 1.7 to 2.0 (Fig. 2). Thus the state of DNA supercoiling within a contrained chromatin loop will depend, in part, on the number of DNA supercoils taken up by the nucleosomes. Other factors are given by the equation for the linking number, Lk, which is the number of times that each DNA strand crosses the other strand in a DNA circle[22]. It follows that for a DNA circle, Lk must be an integer that can be changed only by cleaving one or both DNA strands, rotating the DNA ends and religating the circle. Lk is given by

$$Lk = Tw + Wr$$

where the twisting number, Tw, is the number of DNA helical repeats in the circle and is usually not an integer and Wr, the writhing number, is the number of superhelical turns of DNA in the circle. In a core particle, if the helical twist parameters of the DNA double helix remained unchanged on dissociation from the histone octamer, then the change in Wr or Lk would be −1.7±0.1, the number of DNA superhelical turns on the core particle. The linking number change associated with a core particle can be determined by reassembling core particles onto a closed circular plasmid and relaxing superhelical strain in the DNA between core particles with topoisomerase I. (an enzyme that nicks one of the two polynucleotide chains on duplex DNA, allows the DNA ends to rotate and then religates the cut DNA ends). The average number of core particles on these relaxed circles is obtained by counting them in E.M. pictures of a large number of assembled circular minichromosomes. The core particles are dissociated by exposure of these circular minichromosomes to high salt and the average number of previously constrained DNA supercoils is obtained from the distribution of topoisomers on agarose gels. By dividing the average number of DNA supercoils by the average number of core particles, the linking number change associated with a core particle has been determined to be −1.01±0.08[23]. This less-than-expected change in Lk is attributed to a change in helical twist between free DNA in solution and the DNA coiled around the histone octamer (see ref. 24). The helical repeat of free DNA in solution has been determined to be 10.6 bp/turn, whereas the DNA on the core particle has an average value of 10.1 bp turn. With these values for helical twists the expected linking number change on release of DNA from a core particle would be

$$\Delta Lk = \Delta Tw + \Delta Wr = \left(\frac{146}{10.1} - \frac{146}{10.6}\right) - 1.7 = -1.02 \text{ the observed value.}$$

The overall state of supercoiling of a chromatin domain will depend in part on the number of DNA supercoils constrained by the nucleosomes. If nucleosomes release previously contrained DNA then this will increase the negative DNA supercoiling in the free linker DNA in the chromatin loop and favor the unwinding of a chromatin domain of nucleosomes, whereas if nucleosomes take up additional DNA, that is add on DNA supercoils, then the opposite effect would be expected.

Histone Acetylation

Some irreversible acetylations occur during histone synthesis that block the N-termini of H1, H2A and H4. The sites of reversible histone acetylations are given in Fig. 1. A small subset, 5–10 %, of the core histones are acetylated and because we know that transcriptionally competent nucleosomes are in hyperacetylated states, it follows that acetylation affects only a small but significant chromatin component. The turnover of acetylation on the core histones is usually, but not always (see ref. 25) rapid and determined by the net activities of histone acetyltransferases and deacetylases.

Histone acetylation has been correlated with all aspects of DNA processing in eukaryotes; replication, transcription and spermiogenesis (reviewed in refs 26 and 27). A role for histone acetylation in transcription was first proposed by Allfrey and co-workers (see ref. 27) who have now demonstrated a strict association of hyperacetylated H3 and H4 with the chromatin state of active genes. Such an association has also been shown for H4 acetylation in *Physarum polycephalum*[28] and antibodies, specific for acetylated lysines, bind nucleosomes that are 15–30 fold enriched in active gene sequences[29]. The most detailed cell cycle studies of histone acetylations have used the precise nuclear division cycle found in the macroplasmodium of *Physarum polycephalum* (reviewed in ref. 30). Three patterns of histone acetylations have been observed: i) S-phase acetylations of all the sites in all four core histones, H2A, H2B, H3 and H4, that have been associated with chromatin replication and S-phase gene expression; ii) G_2 phase acetylations to the higher states of acetylation of only H3 and H4 associated with G_2 phase gene expression, and iii) the M-phase deacetylations of all four core histones, presumably to allow the correct packaging of nucleosomes into metaphase chromosomes.

Yeast Genetics and Histone Acetylation

Grunstein and co-workers[31] have used the powerful approach site-directed mutagenesis in yeast (*S. cerevisiae*) genetics, generating residue deletions and replacements, to study the functions of the different domains and sites of acetylations of core histones. It should be noted, however, that a very much larger component of the yeast genome is transcriptionally active compared to the genomes of higher organisms, and this is paralleled by a five fold increase in hyperacetylated forms of histones in this organism[32] and an apparent absence of histone H1. Although deletion of either N-terminal domain of H2A or of H2B results in yeast cells that are viable, deletion of both N-terminal domains is lethal, suggesting some redundancy in the functions of these domains in yeast[33]. Deletions in the apolar globular domains of the core histones are all lethal, showing that interactions of these domains and the integrity of the core particle are essential to chromosome functions. Particularly revealing are the effects of deletions and replacement of residues in the N-terminal domain of histone H4. In a recent report, Durrin *et al.*[34] have shown that the N-terminal domain of H4 is required for GAL 1 gene promoter activation and deletion of the region 4–23, which contains all four sites of acetylation (Fig. 1), causes a marked reduction in GAL 1 gene activation. Deletions in the N-terminal domains of the other core histones do not cause similar effects. Smith and co-workers[35] have shown that the directed replacement of all four sites of acetylation in the N-terminal domain of histone H4 with either arginine or asparagine is lethal. The replacement of any individual lysine, however, resulted in viable cells, but with longer periods of DNA replication. When all four lysines were replaced by glutamines, several effects on the cells were observed; the cells were sterile, they had prolonged S- and G_2/M phases of the cdc, and were temperature sensitive. These studies provided genetic support for specific roles of H4 acetylation events in gene expression, chromatin replication and the cell cycle.

Effects of Histone Acetylation on Chromatin Structure

An interesting early observation that has relevance for structure/function relationships of acetylation, is of the much enhanced nuclease sensitivity of DNA in hyperacetylated chromatin in nuclei[36]. Thus, histone hyperacetylation results in increased accessibility of DNA to nuclease attack presumably through the unfolding of chromatin. However, at the level of the nucleosome core particle, only small structural effects of histone acetylation have been found; hyperacetylated core particles migrate slightly slower than control particles on particle gels(37). Such an effect could be attributed to a change in core particle shape or to the viscous drag of the acetylation released N-terminal domains of the core histones. Low resolution neutron scatter studies, however, showed no effects of hyperacetylation on the overall shape of the core particle in solution(38). Although there is evidence that hyperacetylation of core histones weakens their interactions with the central region of DNA(39), it does not result in an unfolding of the core particle.

A detailed understanding of the structure/function relationships of histone acetylation requires the development of fully defined chromatin systems with both closed circular and linear DNAs. Such studies are now possible through the construction of a DNA circle with 18 repeats of 207 bp DNA, each repeat containing a nucleosome locating sequence(23). Using this construct, we have found that the linking number change ΔLk of an hyperacetylated nucleosome particle is -0.82 ± 0.05, compared to -1.04 ± 0.08 for the control particle(40). These studies have been extended by the fractionation of all states of acetylations of the core histones and the assembly of 207 bp nucleosome particles and 18×207 bp closed circular DNA with octamer containing tetra-acetylated H3 and H4 and bulk H2A and H2B(41). The tetra-acetylated states of H3 and H4 induced the same linking number change as found for the hyperacetylated states of all four core histones. As suggested by Grunstein and coworkers(34), from the different functional effects of the deletions in the N-terminal domains of H3 and H4, the effects of histone acetylation on linking number change may be attributed to H4. There are several possible explanations for histone acetylation induced change in ΔLk. The most likely is that the histone octamer, through its N-terminal domains, binds and coils DNA regions outside of the 146 bp core particle DNA and acetylation releases these segments, resulting in changes of both DNA writhe and twist in these regions. Core histone acetylations may also affect the proposed DNA binding site of the central globular domain of histone H1 (Fig. 2), destabilizing higher order chromatin structures and making previously shielded DNA control regions accessible to transacting factors and polymerases. Some understanding of the biological functions of histone acetylation should come from the identification of a potent and specific inhibitor of mammalian histone deacetylase, trichostation A, which causes cells to arrest in both G$_1$ and G$_2$ phase of the cell cycle(42).

Histone Ubiquitination

Another major cell cycle dependent modification of histone H2A and H2B is the reversible ubiquitination of lysines located in the C-terminal tails of these histones (Fig. 1). Ubiquitin, the most conserved of all eukaryotic proteins, is a very stable, globular 76-amino acid protein. It is covalently attached to H2A and H2B by an isopeptide bond between the C-terminus of ubiquitin and the ϵ amino group of the target lysine side chain. Thus, ubiquitin forms unusual bifurcated nuclear proteins with H2A and H2B. The functions of these nuclear histone ubiquitinations are not known. There is no evidence that they result in the degradation of the labelled H2A and H2B, as is found for the ubiquitin mediated degradation pathways of cytoplasmic proteins. It has been reported that fractionated nucleosomes containing active DNA sequences are also enriched in uH2A and particularly in uH2B(43). The ubiquitinated histones may 'label' active or potentially active genes of a particular gene family.

The additions of bulky globular ubiquitins to the C-terminal tails of both H2A molecules have been shown to have little effect on the structure of the core particle, suggesting that the ubiquitins attached to H2A lie on the faces of the disc-shaped core particles(44). Such ubiquitin locations would be expected to interfere with the close packing of nucleosomes in the 34 nm diameter supercoil. From such structural considerations, it is of some interest that metaphase chromosomes of higher eukaryotes totally lack uH2A(45). The precise nuclear division cycle in the macroplasmodium of *Physarum polycephalum* has allowed detailed studies of the ubiquitination cycle of H2A and H2B(46). uH2A and uH2B are present through S-phase and G$_2$ phase up to prophase. From prophase to metaphase, they are deubiquitinated, but are then reubiquitinated in anaphase. It appears that ubiquitin tags have to be removed to allow the close packaging of nucleosomes in metaphase chromosomes. Additional evidence for an important cell cycle role for ubiquitination comes from the mouse G$_2$ phase mutant cell line ts85(47), which has a temperature sensitive lesion in the ubiquitin-activating enzyme(48). At the non-permissive temperature, ts85 cells arrest close to the S/G$_2$ boundary and this arrest is accompanied by the loss of uH2A. It appears that following S-phase replication of chromosomes, the ubiquitination of a subset of H2A and H2B and/or other proteins is an essential step for progression through the S/G$_2$ boundary.

Results from the cell cycle studies of another ts cell mutant tsBN2(49) suggest an interrelationship between the cell cycle controls of histone phosphorylation and ubiquitination in the process of chromosome condensation. Incubations of tsBN2 cells at the non-permissive temperature induces premature chromosome condensation (PCC) in S and G$_2$-phases of the cell cycle(50). The defective gene, RCC1, has been identified and codes for a 45 kD DNA binding protein of, as yet, unknown function(51). This temperature-induced PCC in tsBN2 in S and G$_2$ phases is accompanied by both the mitosis-related phosphorylations of histone H1 and H3(49) and the deubiquitination of uH2A (Th'ng, J., Bradbury, E. M., unpublished) suggesting that these events are linked in the process of chromosome condensation.

Histone Phosphorylations

Histone H1 has the unusual conformation shown in Fig. 1 and part of its mode of binding to the chromatosome is through the central globular domain, Fig. 2. H1 is known to stabilize the 34 nm supercoil of nucleosomes and is further involved in the salt-induced compaction of chromatin, a process that must involve, in part, charge neutralization between the histones and DNA. Because the phosphorylations of serines and threonines in H1 could be expected to influence this process, precise cell cycle studies of H1 phosphorylations were undertaken on the macroplasmodium of *Physarum polycephalum*(52). It was found that H1 was phosphorylated through S-phase, phosphorylation events that are probably associated with H1 deposition and chromatin replication. Through G_2 phase, all H1 molecules were subjected to increasing phosphorylations to reach a hyperphosphorylated state at metaphase, that was shown later to reach 22–24 phosphates per H1 molecule(53). Immediately following nuclear division, these hyperphosphorylated H1's were rapidly dephosphorylated to the S-phase levels. H1 kinase activity was also shown to vary cyclically, increasing 15 fold from S-phase to just before metaphase and then rapidly falling off(52). These results led to the proposal that an increase in H1 kinase activity initiated and controlled mitosis and the physical process of chromosome condensation was driven by H1 phosphorylation. Support for these proposals came from the demonstration that mitosis could be advanced by up to 40 min when heterologous H1 kinase activity was added to macroplasmodia 3 h prior to the expected metaphase(54). A similar pattern of H1 hyperphosphorylation was observed in mammalian cells and, in addition, all histone H3 molecules were phosphorylated at a single site, serine 10, just prior to metaphase(55). This phosphorylation is probably required for a late stage of chromosome condensation. Over the past ten years, three distinct lines of research have merged to advance significantly our understanding of eukaryotic cell cycle controls: i) cell cycle studies of the reversible chemical modifications of histones and their effects on chromatin structure and function; ii) the powerful approach of yeast genetics in identifying the genes and their products that act at the major decision points in the cell cycle(56); and iii) studies of the rapid cell cycles of fertilized embryos of sea urchins, clams and other organisms, which led to the identification of cyclins that increased in amount through the cell cycle to be specifically degraded at metaphase(57–59).

Yeast genetic studies have identified a protein kinase, p34, the product of the cdc2 gene in *S. pombe*, that is required at decision points in both G_1 and G_2 phases of the cell cycle. Through S phase, phosphorylated forms of p34 complex with mitotic cyclins to form an inactive kinase. This kinase is activated through G_2 phase by the dephosphorylation of a phosphotyrosine in p34. The isolated H1 kinase activity increases through G_2 phase to peak at metaphase and then rapidly falls off, a behavior identical to H1 kinase activity through the nuclear division cycle of *Physarum polycephalum*(52). We have recently characterized a temperature-sensitive G_2 phase mutant FT210 cell(60) that derives from a mouse mammary tumor cell line FM3A(61). Two lesions have been found in the CDC2 gene and the H1 kinase isolated from FT210 is temperature labile. At the non-permissive temperature, FT210 cells block only in early G_2 phase. Histone H1 does not undergo mitosis-related phosphorylations, suggesting that H1 is an *in vivo* substrate for the p34/cyclin kinase. Our finding that there is no temperature-induced G_1 block in the FT210 mutant cell suggests that in transformed mammalian cells, there is either no CDC2 gene involvement in G_1 regulation or that other CDC2-like genes may act with G_1 cyclins in the G_1 phase of the cell cycle. Another p34-like kinase, p33, has been reported that is activated earlier in the cell cycle(62). However, a very important recent study(63) shows that there is a cascade of kinase-mediated controls of progression through G_1 phase of normal cells that is completely absent from transformed cells.

H1 Phosphorylation and Chromatin Structure

Based on the known chromatin structural effects of histone H1, we have proposed that H1 phosphorylation is a major factor involved in physical process of chromosome condensation. The cause-and-effect relationships, however, have yet to be demonstrated. The major reasons for this situation are the complexity of chromosomes, the well-known difficulties in handling isolated chromatin, the unsolved problem of how flexible histone domains interact with DNA, and the unknown role of divalent cations, such as calcium. A similar situation applied to an understanding of histone acetylation in nucleosome and chromatin structure, until it was shown by using fully defined chromatin model systems that hyperacetylation of H3 and H4 caused a reduction in the nucleosome core particle linking number change(40,41), thus releasing negative DNA supercoiling into a constrained DNA loop facilitating the unfolding of chromatin domains. Because the reverse process would have the opposite effect, an attractive model under test is that H1 phosphorylation increases nucleosome linking number change. thereby introducing positive DNA supercoiling into a chromatin loop. It has been shown that SPKK sequences in H1, the H1 growth-associated kinase phosphorylation sites, have a specific interaction with AT-rich DNA segments(64). The nature of this interaction is not understood. but it is to be expected that the phosphorylation of these sites would probably abolish their interaction with DNA.

If it is accepted that the major function of the cell cycle is to package genomes or to repackage them during the developmental process, then it is logical that cell cycle dependent changes in chromosomal proteins are important in understanding cell cycle controls. Changes in charge are introduced into the histone sequences through the reversible modifications of acetylation and phosphorylation. The third major modification, ubiquitination, introduces a major structural perturbation in chromosomes. The metabolic costs to the cell in the syntheses of the proteins required to control these modifications are not trivial and underscores their importance in the cell cycle.

Acknowledgements

The work referred to from this laboratory is supported by grants from NIH (PHS GM-26901), ACS (CD-484F), and DOE (DE-FG03–88ER60673). I am grateful to Timothy O'Neill for his comments on the manuscript.

References

1 **Paulson, J. R. and Laemmli, U. K.** (1977). The structure of histone-depleted metaphase chromosomes. *Cell* **12**, 817-828.

2 **Igo-Kemenes, T. and Zachau, H. G.** (1977). Domains in chromatin structure. *Cold Spring Harb. Symp. Quant. Biol.* **42**, 109-118.

3 **Van Holde, K. E.** (1988). *Chromatin*. Springer-Verlag, New York. Heidelberg.

4 **Lewis, C. D. and Laemmli, U. K.** (1982). Higher order metaphase chromosome structure: Evidence for metalloprotein interactions. *Cell* **29**. 171-181.

5 **Earnshaw, W. C. and Heck, M. M. S.** (1985). Localization of topoisomerase II in mitotic chromosomes. *J. Cell Biol.* **100**, 1716-1725.

6 **DiNardo, S., Voelkel, K. and Sternglanz, R.** (1984). DNA topoisomerase II mutant of *Saccharomyces cerevisiae*: Topoisomerase II is required for segregation of daughter molecules at the termination of DNA replication. *Proc. Natl Acad. Sci. USA* **81**, 2616-2620.

7 **Uemura, T. and Yanagida, M.** (1986). Mitotic spindle pulls but fails to separate chromosomes in type II DNA topoisomerase mutants: Uncooordinated mitosis. *EMBO J.* **5**. 1003-1010.

8 **Holm, C., Goto, T., Wang, J. C. and Botstein, D.** (1985). DNA topoisomerase II is required at the time of mitosis in yeast. *Cell* **41**, 553-563.

9 **Roberge, M., Th'ng, J., Hamaguchi, J. and Bradbury, E. M.** (1990). The topoisomerase II inhibitor VM-26 induces marked changes in histone H1 kinase activity, histones H1 and H3 phosphorylations and chromosome condensation in G2 phase and mitotic BHK cells. *J. Cell Biol.* **111**, 1753-1762.

10 **Uemura, T., Ohkura, H., Adachi, Y., Morino, K. and Shiozaki, Y.** (1987). DNA topoisomerase II is required for condensation and separation of mitotic chromosomes in *S. pombe*. *Cell* **50**, 917-925.

11 **Charron, M. and Hancock, R.** (1990). DNA topoisomerase II is required for formation of mitotic chromosomes in chinese hamster ovary cells: Studies using the inhibitor 4'-demethylepipodophyllotoxin 9-(4,6-0-thenylidene-β-D-glucopyranoside). *Biochemistry* **29**, 9531-9537.

12 **Adachi, Y., Luke, M. and Laemmli, U. K.** (1991). Chromosome assembly *in vitro*: Topoisomerase II is required for condensation. *Cell* **64**, 137-148.

13 **Wang, J. C.** (1985). *DNA Topoisomerases. Annu. Rev. Biochem.* **54**. 665-697.

14 **Cole, R. D.** (1984). A minireview of microheterogeneity in H1 histone and its possible significance. *Anal. Biochem.* **136**, 24-30.

15 **Schroth, G. P., Yau, P., Imai, B. S., Gatewood, J. M. and Bradbury, E. M.** (1990). NMR study of mobility in the histone octamer. *FEBS Lett.* **268**. 117-220.

16 **Böhm, L. and Crane-Robinson, C.** (1984). Proteases as structural probes for chromatin: The domain structure of histones (review). *Bioscience Reports* **4**. 365-386.

17 **Bradbury, E. M. and Baldwin, J. P.** (1986). Neutron scatter studies of chromatin structure. In *Supramolecular Structure and Function* (ed. G. Pifat-Mrzljak). Springer-Verlag, Berlin, Heidelberg.

18 **Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. and Klug, A.** (1984). Structure of the nucleosome core particle at 7Å resolution. *Nature* **311**. 532-537.

19 **Uberbacher, E. C. and Bunick, G. J.** (1989). Structure of the nucleosome core particle at 8A resolution. *J. Biomol. Structure & Dyn.* **7**, 1-20.

20 **Lewis, P. N. and Bradbury, E. M.** (1974). Effect of electrostatic interactions on the predictions of helices in proteins. *Biochem. Biophy. Acta* **335**. 19-29.

21 **Gellert, M.** (1981). DNA topoisomerases. *Annu. Rev. Biochem.* **50**, 879-910.

22 **Bauer, W. and Vinograd, J.** (1968). The interaction of closed circular DNA with intercalative dyes. *J. Mol. Biol.* **33**, 141-171.

23 **Simpson, R. T., Thoma, F. and Brubaker, J. M.** (1985). Chromatin reconstituted from tandemly repeated cloned DNA fragments and core histones: A model system for study of higher order structure. *Cell* **42**. 799-808.

24 **See letters from White, J. M., Bauer, W. R., also Klug, A. and Travers, A. A.** (1989). The helical repeat of nucleosome-wrapped DNA. *Cell* **56**, 9-11.

25 **Covault, J. and Chalkley, R.** (1980). The identification of distinct populations of acetylated histones. *J. Biol. Chem.* **255**, 9110-9116.

26 **Csordas, A.** (1990). On the biological role of histone acetylation. *Biochem. J.* **265**, 23-38.

27 **Allfrey, V. G.** (1980). Molecular aspects of the regulation of eukaryotic transcription-nucleosomal proteins and their postsynthetic modification in the control of DNA conformation and template function. In *Cell Biology: A Comprehensive Treatise*, Vol. 3 (eds L. Goldstein, and D. M. Prescott). Academic Press Inc.

28 **Chahal, S., Matthews, H. R. and Bradbury, E. M.** (1970). Acetylation of histone H4 and its role in chromatin structure and function. *Nature* **287**. 76-79.

29 **Hebbes, T. R., Thorne, A. W. and Crane-Robinson, C.** (1988). A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J.* **7**, 1395-1402.

30 **Matthews, H. R. and Waterborg, J. M.** (1985). Reversible modifications of nuclear proteins and their significance. In *The Enzymology of Post-translational Modifications of Proteins*, Vol. 2, Academic Press Inc., London.

31 **Grunstein, M.** (1990). Nucleosomes: regulators of transcription. In *TIG* **6**, 395-400.

32 **Davie, J. R., Saunders, C. A., Walsh, J. M. and Weber, S. C.** (1981). Histone modifications in the yeast *S. cerevisiae*. *Nucl. Acids Res.* **9**, 3205-3216.

33 **Schuster, T., Han, M. and Grunstein, M.** (1986). Yeast histone H2A and H2B amino termini have interchangeable functions. *Cell* **45**, 445-451.

34 **Durrin, L. K., Mana, R. K., Kayne, P. S. and Grunstein, M.** (1991). Yeast histone H4 N-terminal sequence is required for promoter activation *In vivo*. *Cell* **65**, 1023-1031.

35 **Megee, P. C., Morgan, B. A., Mittman, B. A. and Smith, M. M.** (1990). Genetic analysis of histone H4: Essential role of lysines subject to reversible acetylation. *Science* **247**. 841-845.

36 **Vidali, G., Boffa, L. C., Bradbury, E. M. and Allfrey, V. G.** (1978). Supression of histone deacetylation leads to accumulation of multiacetylated forms of histones H3 and H4 and increased DNase 1 sensitivity of associated DNA sequences. *Proc. Natl Acad. Sci. USA* **75**, 2239-2244.

37 **Bode, J., Gomez-Lira, M. M. and Schröter, H.** (1983). Nucleosome particles open as the histone core becomes hyperacetylated. *Eur. J. Biochem.* **130**, 437-445.

38 **Imai, B. S., Yau, P., Baldwin, J. P., Ibel, K., May, R. P. and Bradbury, E. M.** (1986). Hyperacetylation of core histones does not cause unfolding of nucleosome: Neutron scatter data accords with disc structure of the nucleosome. *J. Biol. Chem.* **261**, 8784-8792.

39 **Ausio, J. A. and van Holde, K. E.** (1986). Histone hyperacetylation: Its effects on nucleosome conformation and stability. *Biochem.* **25**, 1421-1428.

40 **Norton, V. G., Imai, B. S., Yau, P. and Bradbury, E. M.** (1989). Histone acetylation reduces nucleosome core particle linking number change. *Cell* **57**. 449-457.

41 **Norton, V. G., Marvin, K. W., Yau, P. and Bradbury, E. M.** (1990). Nucleosome linking number change controlled by acetylation of histones H3 and H4. *J. Biol. Chem.* **265**, 19 848-19 852.

42 **Yoshida, M., Kijima, M., Akita, M. and Beppu, T.** (1990). Potent and specific inhibition of mammalian histone deacetylase both *in vivo* and *in vitro* by trichostatin A. *J. Biol. Chem.* **265**, 17 174-17 179.

43 **Nickel, B. E., Allis, D. C. and Davie, J. R.** (1989). Ubiquitinated histone H2B is preferentially located in transcriptionally active chromatin. *Biochem.* **28**, 958-963.

44 **Kleinschmidt, A. M. and Martinson, H. G.** (1981). Structure of nucleosome core particles containing uH2A. *Nucl. Acids Res.* **9**. 2423-2431.

45 **Matsui, S. I., Seon, B. K. and Sandberg, A. A.** (1979). Disappearance of a structural chromatin protein A24 in mitosis: Implication for molecular basis of chromatin condensation. *Proc. Natl Acad. Sci. USA* **76**, 6386-6390.

46 **Mueller, R. D., Yasuda, H., Hatch, C. L., Bonner, W. M. and Bradbury, E. M.** (1985). Identification of ubiquitinated histones H2A and H2B in *Physarum polycephalum*. Disappearance of these proteins at metaphase and reappearance at anaphase. *J. Biol. Chem.* **260**, 5147-5153.

47 **Yasuda, H., Matsumoto, Y., Mita, S., Marunouchi, T. and Yamada, M.** (1981). A mouse temperature-sensitive mutant defective in H1 histone phosphorylation is defective in DNA synthesis and chromosome condensation. *Biochem.* **20**, 4414–4419.

48 **Finley, D., Ciechanover, A. and Varshavsky, A.** (1984). Thermolability of ubiquitin-activating enzyme from the mammalian cell cycle mutant ts85. *Cell* **37**, 43-55.

49 Nishimoto, T., Ajiro, K., Davis, F. M., Yamashita, K., Kai, R., Rao, P. N. and Sekiguchi, M. (1987). Mitosis-specific protein phosphorylation associated with premature chromosome condensation in a ts cell cycle mutant. In *Molecular Regulation of Nuclear Events in Mitosis and Meiosis*. Academic Press, Inc., 295-318.

50 Nishimoto, T., Ishida, R., Ajiro, K., Yamamoto, S. and Takahashi, T. (1981). The synthesis of protein(s) for chromosome condensation may be regulated by a post-transcriptional mechanism. *J. Cell. Physiol.* **109**, 299-308.

51 Ohtsubo, M., Okazaki, H. and Nishimoto, T. (1989). The RCC1 protein, a regulation for the onset of chromosome condensation locates in the nucleus and binds DNA. *J. Cell Biol.* **109**, 1389-1397.

52 Bradbury, E. M., Inglis, R. J. and Matthews, H. R. (1974). Control of cell division by very lysine rich histone phosphorylation. *Nature* **247**, 257-261.

53 Mueller, R. D., Yasuda, H. and Bradbury, E. M. (1985). Phosphorylation of histone H1 through the cell cycle of *Physarum polycephalum*: 24 sites of phosphorylation at metaphase. *J. Biol. Chem.* **260**, 5081-5086.

54 Bradbury, E. M., Inglis, R. J., Matthews, H. R. and Langan, T. A. (1974). Molecular basis of control of mitotic cell division in eukaryotes. *Nature* **249**, 553-556.

55 Gurley, L. R., Walters, R. A. and Tobey, R. A. (1975). Sequential phosphorylation of histone subfractions in the chinese hamster cell cycle. *J. Biol. Chem.* **250**, 3936-3944.

56 Nurse, P. (1990). Universal control mechanism regulating onset of M-phase. *Nature* **344**, 503-507.

57 Hunt, T. (1989). Maturation promoting factor, cyclin and the control of M-phase. *Curr. Opin. Cell Biol.* **1**, 268-274.

58 Pines, J. and Hunter, T. (1990). $p34^{cdc2}$: The S and M kinase? *The New Biologist* **2**, 389-401.

59 Lohka, M. J. (1989). Mitotic control by metaphase-promoting factor and cdc proteins. *J. Cell Sci.* **92**, 131-135.

60 Th'ng, J. P. M., Wright, P. S., Hamaguchi, J., Lee, M. G., Norbury, C. J., Nurse, P. and Bradbury, E. M. (1990). The FT210 cell line is a mouse G2 phase mutant with a temperature-sensitive cdc2 gene product. *Cell* **63**, 313-324.

61 Mineo, C., Murakami, Y., Ishimi, Y., Hanoaka, F. and Yamada, M. (1986). Isolation and analysis of a mammalian temperature-sensitive mutant defective in G2 functions. *Exp. Cell Res.* **167**, 53-62.

62 Pines, J. and Hunter, T. (1990). Human cyclin A is Adenovirus E1A-associated protein p60 and behaves differently from cyclin B. *Nature* **346**, 760-763.

63 Crissman, H. A., Gadbois, D. M., Tobey, R. A. and Bradbury, E. M. (1991). Transformed mammalian cells are deficient in kinase-mediated control of G1 phase progression. *Proc. Natl Acad. Sci. USA*, in press.

64 Churchill, M.E.A. and Suzuki, M. (1989). 'SPKK' motifs prefer to bind to DNA at A/T-rich sites. *EMBO J.* **8**, 4189-4195.

E. Morton Bradbury is at the Dept. Biological Chemistry, School of Medicine, University of California, Davis, CA 95616, and Div. of Life Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.