UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL ATOMIC ENERGY AGENCY
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/916 -15

SEVENTH COLLEGE ON BIOPHYSICS:

Structure and Function of Biopolymers: Experimental and Theoretical Techniques.

4 - 29 March 1996

Structure and Function of repeated DNA sequences and of nucleosomes

Morton BRADBURY
Life Sciences Division
Los Alamos National Laboratory
P.O.Box 1663
New Mexico
87545 Los Alamos
U.S.A.

UNRAVELING THE CHROMOSOME

E. Morton Bradbury

Central to biology is an understanding of the organization, structure, and functions of the chromosomes of higher organisms. Chromosomes contain the DNA molecules of the genome and are themselves contained within the cell nuclei of all eukaryotes, from single-celled yeast all the way up the evolutionary ladder to human beings. As pointed out by David Galas (pages 164–165 of "Mapping the Genome"), to understand the functions of the multitude of protein-coding and noncoding DNA sequences that will be determined by the Human Genome Project, we will need detailed knowledge of the three-dimensional structure of chromosomes and the structural changes that chromosomes undergo during the various phases of the cell cycle. Major advances in biology will be at the interfaces between the Human Genome Project, structural biology, and molecular biology of the cell.

The size of the human genome suggests the magnitude of the problem. The diploid human genome contains 6×10^9 base pairs or 204 centimeters of DNA molecules packaged into 46 chromosomes. It is generally believed that each chromosome contains a single DNA molecule several centimeters in length.

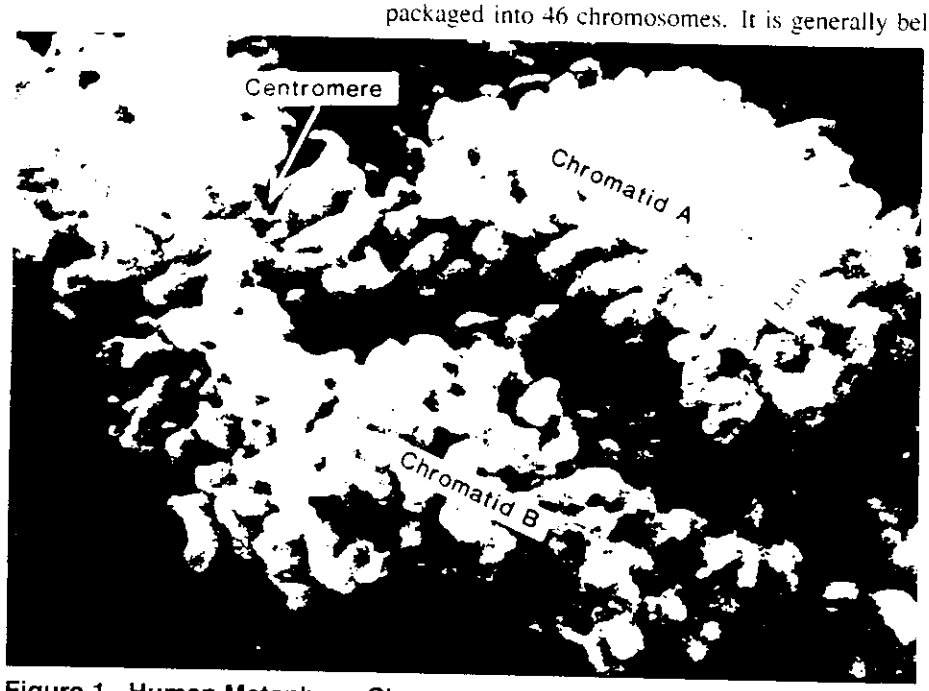


Figure 1. Human Metaphase Chromosome

A scanning transmission electron micrograph of a metaphase chromosome showing two sister chromatids attached at the centromeres. Each compact projection is thought to be a long loop of DNA (see Figure 2) packaged along with various proteins into a thick chromatin fiber. (Reprinted courtesy of U.K. Laemmli, Université de Genève.)

Studies of the yeast *S. cerevisiae*, a lower eukaryote that can be easily manipulated, have revealed three chromosomal elements that are essential to the faithful replication of each chromosome and to the subsequent separation of the two duplicate chromosomes into daughter cells during cell division. These are: (1) the very ends of chromosomes, called the telomeres; (2) a central region of constriction called the centromere that, after replication of a chromosome, is the last point of attachment between the resulting pair of sister chromatids; and (3) a DNA sequence required to initiate DNA replication, called an origin of replication.

Figure 1 is a scanning transmission electron micrograph of a human metaphase chromosome, the highly condensed structure adopted by the chromosome during metaphase, one of the last phases of cell division. The chromosome has already replicated into two sister chromatids. The centromere connecting the sister chromatids (seen in the micrograph as a region of constriction) provides the point of attachment

for the spindle apparatus that contracts and separates the replicated chromosomes into the daughter cells. The telomeres at the ends of each chromatid contain tandem repeated DNA sequences that cap, protect, and maintain the linear DNA ends of the chromosomes during replication.

Each of the 46 human chromosomes can be identified during metaphase by its length, the location of its centromere, and the particular banding pattern produced by staining the DNA of that chromosome. (Banding patterns can be seen in "Chromosomes: The Sites of Hereditary Information" in "Understanding Inheritance.") The origins of the distinctive banding patterns are not well understood but probably reflect a reproducible pattern of DNA folding induced by DNA-protein interactions specific to each chromosome. The DNA molecule is very tightly wound during metaphase. For example, human chromosome 16 is 2.5 micrometers long, whereas the DNA molecule in each sister chromatid is 3.7 centimeters long. In other words, the packing ratio of the linear DNA molecule in the metaphase chromosome is 15,000 to 1.

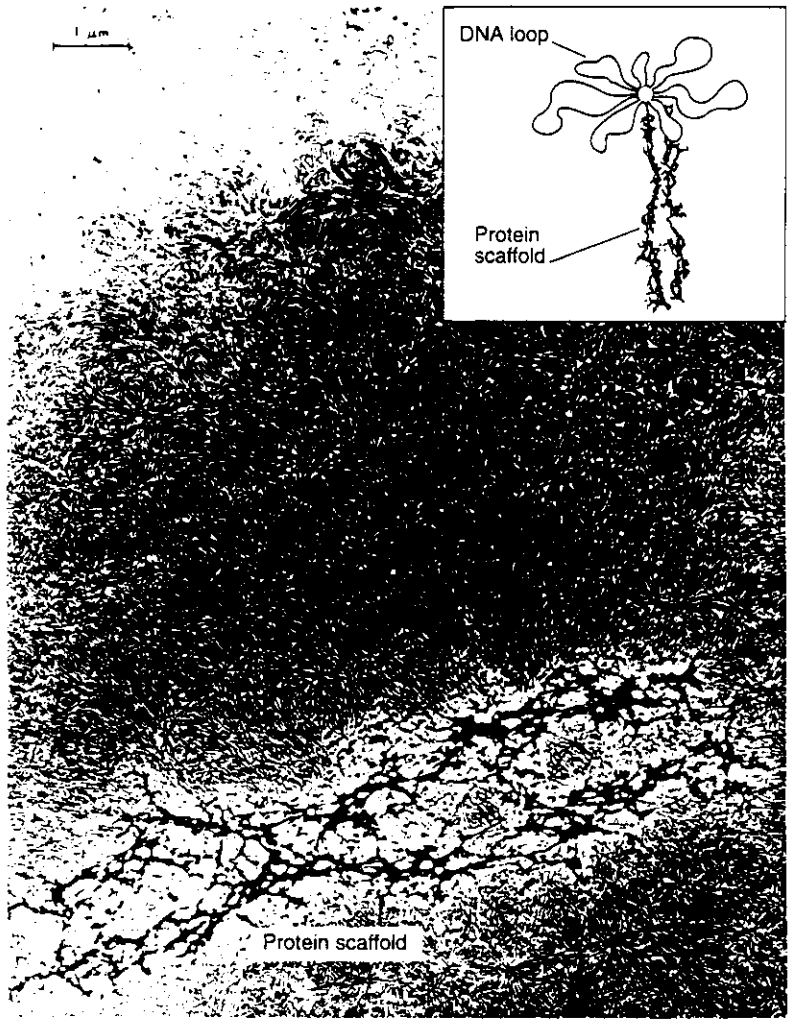


Figure 2. Chromosome Loops and Protein Scaffold
Above is a metaphase chromosome depleted of almost all chromosomal proteins. The remaining 2 to 3 percent of the proteins form a scaffold that retains the shape of the intact chromosome. Around the scaffold is a halo of loops of naked DNA. Each loop appears to begin and end at the same point along the protein scaffold (see insert). The number and sizes of these loops suggest that each may contain a single gene or a group of linked genes. (Reprinted courtesy of U.K. Laemmli, Université de Genève.)

Chromosomal DNA Loops

When chromosomal material is isolated from the nucleus, the long DNA molecules are found to be associated with chromosomal proteins, whose weight is up to twice that of the DNA. The five histones, the many copies of which are equal in weight to that of DNA, are found in all eukaryotes and as explained below are involved in packaging the DNA in the chromosomes. The non-histone proteins are a heterogeneous group and many are associated with various chromosome functions, such as replication, gene expression, and chromosome organization. Among the latter are a small group that bind most tightly to the DNA and form a scaffold for the chromosome. This protein scaffold has been made visible by gently treating metaphase chromosomes with detergents to remove the histones and most other nonhistone proteins. The remarkable structure that remains is shown in Figure 2. The residual protein scaffold, or "ghost," of the metaphase chromosome is surrounded by a halo of DNA. At higher

resolution DNA loops can be observed to emerge from and return to the same point on the protein scaffold (see inset in Figure 2).

Two major scaffold proteins have been isolated, Sc1 and Sc2. Sc1 has been identified as topoisomerase II, an enzyme that relaxes supercoiled DNA by cutting through both strands of the DNA, thereby enabling the cut DNA ends to rotate, and then resealing the cut. The cuts made by topoisomerase II are essential for the separation of sister chromatids to the daughter cells.

The DNA loops in Figure 2 range in size from 5,000 to 120,000 base pairs and have an average size of about 50,000 base pairs. Thus the haploid human genome of 3×10^9 base pairs of DNA corresponds to 60,000 loops, which is close to the estimated numbers of genes, 50,000 to 100,000, in the human genome. Perhaps each DNA loop contains one or a small number of linked genes and therefore serves as both a genetic and a structural unit of eukaryotic chromosomes. This tantalizing conjecture was first made in 1978, and although it remains unproven, evidence in its favor has been accumulating.

Chromatin Contains a Repeating Subunit Structure

Having looked at some of the largest structural features of the chromosome, we now turn to what we know about the small, repeating substructures within a chromosome. DNA with its associated chromosomal proteins, histones, and nonhistone proteins, is called chromatin. In 1973 chromatin in isolated nuclei was first digested with micrococcal nuclease, an enzyme that cuts double-stranded DNA. The digestion yielded a ladder of DNA lengths in multiples of about 190 to 200 base pairs. Evidently DNA sequences spaced by 190 to 200 base pairs were more accessible to attack by micrococcal nuclease than the intervening DNA. This seminal observation showed that chromatin contained a simple, repeating subunit, known as the nucleosome.

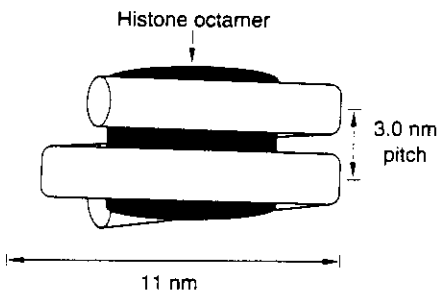


Figure 3. Nucleosome Core Particle
Structure of the nucleosome core particle determined from neutron scattering. The core particle is a flat disc, 100 angstroms in diameter and 55 to 60 angstroms thick.

For most somatic tissues, the nucleosome contains three elements, a stretch of DNA containing 195±5 base pairs, one copy of the histone octamer $[(H3_2H4_2)(H2A.H2B)_2]$ and one copy of the histone H1. More prolonged micrococcal nuclease digestion reduces the length of the DNA in the nucleosome, thereby creating a slightly smaller unit, called the chromatosome, which contains 168±2 base pairs of DNA, the histone octamer, and H1. Such digestion often reduces the nucleosome to an even smaller unit contained within the chromatosome and called the nucleosome core particle. It contains 146±1 base pairs of DNA and the histone octamer (see Figure 3).

The nucleosome core particle has been obtained in large quantities and subjected to extensive structural studies. In 1974 neutron-scattering studies of the core particle in aqueous solution showed that it was a flat disc of diameter 100 angstroms and thickness 55 to 60 angstroms, with 1.7 turns of DNA coiled on the outside of a core of the histone octamer at a pitch of about 30 angstroms

(Figure 3). Subsequent x-ray-diffraction studies of crystallized core particles achieved a resolution of 6 to 7 angstroms. The crystal structure (Figure 4) not only confirmed the lower resolution solution structure achieved by neutron scattering but also showed that histones are in contact with the minor groove of DNA and leave the major groove available for interactions with the proteins that regulate gene expression and other DNA functions. The 7-angstrom-resolution crystal structure also revealed that DNA does not bend uniformly but rather bends gently and then more sharply around the histone octamer. Such a path implies that flexibility, or bendability, of DNA may be sequence-dependent and that the underlying DNA sequence along the molecule may determine the positions of some nucleosomes. The most recent work on nucleosome positioning shows that the bulk of nucleosome core particles are able to move along the DNA molecule between a cluster of positions separated by about 10 base pairs. This mobility is probably required during DNA replication and transcription to allow DNA polymerases and other enzymes access to specific DNA sequences.

Figure 4. Crystal Structure of Core Particle
The structure of the nucleosome core particle as determined by x-ray diffraction is shown above. At a resolution of 6 to 7 angstroms, this top view of the core particle shows that the DNA (brown) does not follow a smooth path around the histone octamer (blue and turquoise) but rather bends sharply and then more gently. (Reprinted courtesy of Uberbacher and Bunick, Oak Ridge National Laboratory.)

Despite considerable effort to achieve higher resolution, the best data for the core particle structure is at a resolution of about 6 angstroms. However, the crystal structure of the isolated histone octamer has been solved to the higher resolution of 3.3 angstroms. This structure shows shapes of the individual histones and the nature of interhistone interaction of most but not all of the histone polypeptide chains. In particular, the basic N-terminal domains, comprising some 20 to 25 percent of the histone octamer, are not "seen" in the crystal structure, probably because they bind to DNA, and in the absence of DNA, they are disordered. These N-terminal domains contain all of the sites of the cell-cycle-dependent acetylation of lysines and phosphorylation of serines or threonines. Acetylation of lysine converts it from a positively charged residue, which can therefore bind to DNA, to a neutral acetylysine. It has been shown first that lysine acetylation is strictly correlated with transcription and DNA replication, and more recently, that histone acetylation drives the uncoiling of part of the DNA from the nucleosome to allow the initiation and progression of DNA replication and transcription.

Chromatosomes and Nucleosomes

A model of the structure of the chromatosome (Figure 5) has been inferred from the structures of the nucleosome core particle and the histone H1. The core particle has 1.7 turns of DNA at a pitch of 3.0 nanometers (30 angstroms) coiled around the histone octamer. Consequently, the chromatosome's 168 base pairs of DNA are long

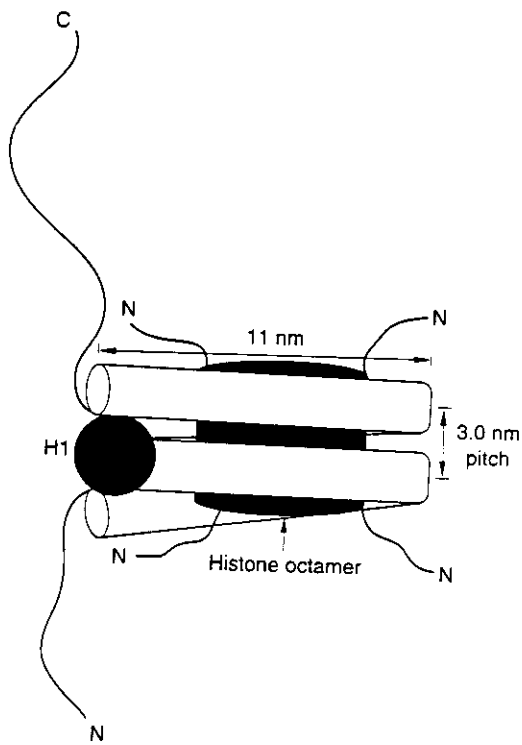


Figure 5. Model of the Chromatosome
The model includes the nucleosome core particle, an extra stretch of DNA, and the histone H1. The DNA makes two complete turns around the histone octamer, and H1 is bound to the outside of the coil at the place where the coil begins and ends. In this position H1 might server to modulate long-range interactions that modify chromosome structure during the cell cycle.

enough to complete two turns of DNA around the histone octamer. The chromatosome also includes the fifth histone H1. In the model structure shown in Figure 5, the histone H1 is bound to the outside of the coiled DNA where it might serve to modulate long-range interactions associated with reversible changes in chromosome structure during the cell cycle. During cell division chromosomes become more and more condensed until they reach metaphase. Then, when cell division is completed and the daughter cells enter interphase, the chromosomes assume a less-condensed configuration (see "Mitosis" in "Understanding Inheritance"). The long, flexible "arms" of H1 undergo a pattern of phosphorylations through this cycle, which may well modulate the long-range interactions required to coordinate these structural changes in the chromosomes. In support of this hypothesis is the fact that an increase in H1 phosphorylation has been correlated with the process of chromosome condensation to metaphase chromosomes. To describe the nucleosome beyond the model for the chromatosome requires a knowledge of the paths of the DNA that link one nucleosome to another. Our present lack of knowledge about those paths impedes our ability to pin down the higher-order chromatin structures that make up the chromosome.

Higher-Order Chromatin Structures

Although higher-order structures of chromatin cannot be resolved in the chromosome itself, they can be studied in solution. Chromatin, when placed in low ionic strength, 10-millimolar NaCl, forms a 10-nanometer-diameter fibril of nucleosomes, which is sometimes referred to as "beads on a string." This form is also observed when chromatin spills out of lysed nuclei. Neutron-scattering studies of the 10-nanometer chromatin fibril give a mass per unit length equivalent to one nucleosome per 10±2 nanometers of fibril, or a DNA packing ratio of between 6 and 7 to 1. When ionic strength is increased to 150-millimolar NaCl, corresponding to normal physiological conditions, the 10-nanometer fibril undergoes a transition to the "30-nanometer" fibril. Neutron-scattering studies indicate that the diameter for this fibril in solution is 34 nanometers and the mass-per-unit length is equivalent to 6 to 7 nucleosomes per 11 nanometers of fibril, or a DNA packing ratio of between 40 and 50 to 1. Figure 6 shows the simplest model of the 34-nanometer fibril that is consistent with available structural data: it is a supercoil or solenoid of 6 to 7 radially arranged disc-shaped nucleosomes with a pitch of 11.0 nanometers and a diameter of 34 nanometers. Basic questions concerning the location of histone H1 and the linker DNA connecting the nucleosomes remain unanswered.

Packaging of Chromosome Loops

With these higher order chromatin structures in mind, we can imagine how the large transverse DNA loops present in the histone-depleted metaphase chromosome (see Figure 2) might be packaged in the normal chromosome. Since the average size of the

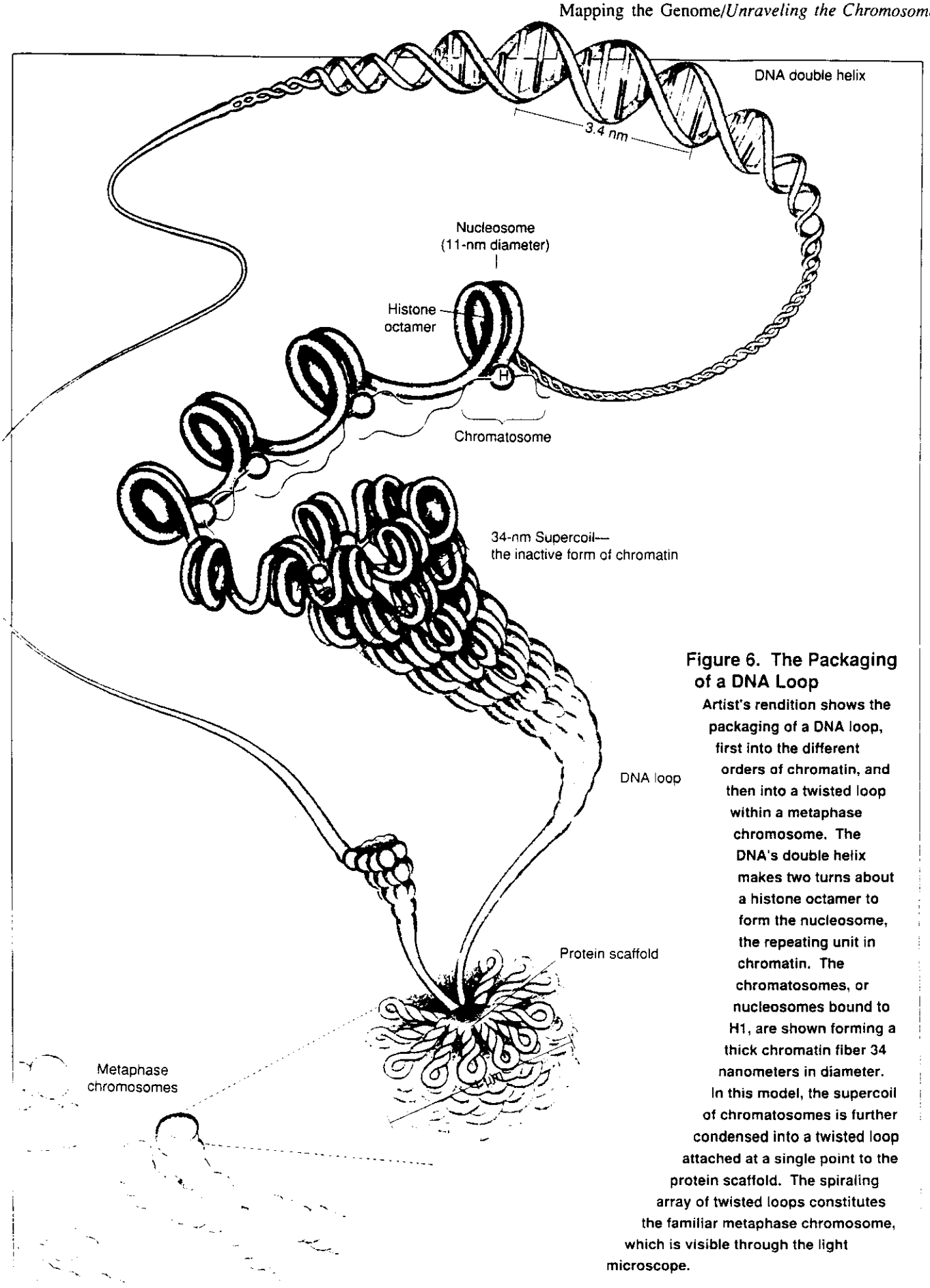


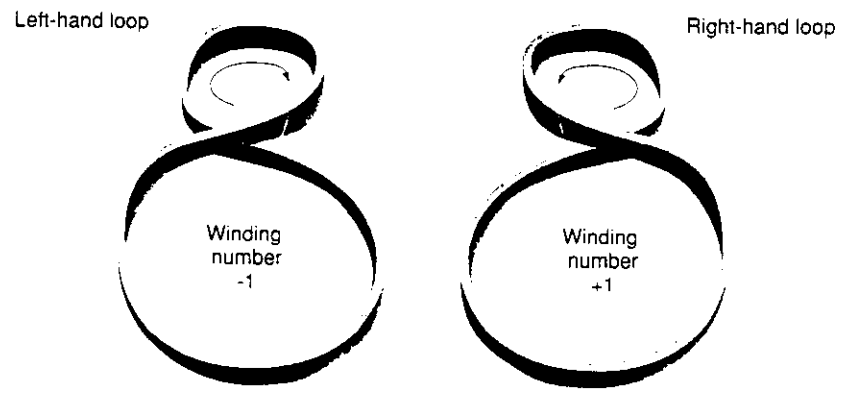
Figure 6. The Packaging of a DNA Loop

Artist's rendition shows the packaging of a DNA loop, first into the different orders of chromatin, and then into a twisted loop within a metaphase chromosome. The DNA's double helix makes two turns about a histone octamer to form the nucleosome, the repeating unit in chromatin. The chromatosomes, or nucleosomes bound to H1, are shown forming a thick chromatin fiber 34 nanometers in diameter. In this model, the supercoil of chromatosomes is further condensed into a twisted loop attached at a single point to the protein scaffold. The spiraling array of twisted loops constitutes the familiar metaphase chromosome, which is visible through the light microscope.

DNA loops is 50,000 base pairs, or 17 micrometers in length, each loop of DNA can form a string of nucleosomes that are either coiled to form 2.6 micrometers of a 10 nanometer fiber, or supercoiled into 0.4 micrometers of a 34 nanometer fiber. Thus, to create the thickness of a sister chromatid (Figure 1), which is 1 micron in diameter, would require just one more order of chromatin folding above the 34 nanometer supercoil. Figure 6 shows a possible model of this final level of chromatin folding.

How is the packaging of DNA loops controlled in response to chromosome functions? Evidence suggests that the inactive form of chromatin is the 34-nanometer supercoil or solenoid of nucleosomes. For both DNA transcription and genome replication this supercoil of nucleosomes must first be uncoiled to the linear array of nucleosomes and then the DNA must uncoil even further to allow access of the transcriptional machinery or the replication machinery to the DNA sequences. Whenever DNA is constrained by proteins to form a loop, DNA supercoiling becomes an important consideration in understanding DNA structure-function relationships. DNA supercoiling has been subjected to extensive experimental and mathematical analysis.

Figure 7. Configurations of a Closed Loop of Ribbon with Winding Numbers +1 and -1



Consider a model in which each DNA loop is firmly attached to the protein scaffold of a chromosome and is therefore somewhat analogous to a closed loop of ribbon. A closed loop of ribbon has a topologically invariant property known as the winding number, which is the number of twists in the ribbon plus the number of times the ribbon crosses itself, that is, coils about itself. The winding number is an integer or half-integer and remains constant unless the ribbon is cut. Each complete twist and each complete crossing adds $+1$ or -1 to the winding number depending on the direction of the twist or crossing. A right-handed twist (the same direction as the thread of a standard screw and the standard helical structure of a double-stranded DNA molecule) is positive, and a left-handed twist is negative. Similarly, a crossing that produces an extra right-handed loop in a loop of ribbon is positive, and a crossing that produces an extra left-handed loop in a loop of ribbon is negative (see Figure 7).

Now consider a loop of double-stranded DNA. Unconstrained DNA has 10.4 to 10.6 base pairs in each complete turn of the double helix. Taking the value 10.6 base pairs per helical turn, the twist (Tw) of a loop of unconstrained DNA consisting of N base pairs would be $N/10.6$. Because a double-stranded DNA molecule already has a helical structure, a loop of DNA further coiled about itself is said to be supercoiled. The linking number (Lk) of a closed loop of DNA is defined in terms of the twist and the number of supercoils, or writhe (Wr), through the equation $Lk = Tw + Wr$. Twists can be converted into supercoils, but Lk must remain constant in a DNA loop whose ends are fixed, in analogy with the constancy of the winding number of the loop of ribbon. If the loop is closed, the linking number must be an integer.

As an example, suppose three helical turns of a linear stretch of DNA are unwound and the ends are then joined. The linking-number change resulting from the unwinding is -3, and the loop can take on any of the three configurations shown in Figure 8. Moreover, the three configurations can be converted into one another without cutting the DNA. DNA configured as in (b) and (c) is said to be negatively supercoiled.

As shown in Figure 3, the DNA in the nucleosome core particle has 1.7 left-handed supercoils and in early studies it was expected that the linking-number change associated with the dissociation of a core particle would be -1.7. However, the experimentally determined linking-number change was -1.02. Although this difference was unexpected and initially controversial, it is easily explained by the change in twist between the DNA constrained in the core particle and free DNA in solution. The average DNA helical repeat on the core particle as measured from its crystal structure is 10.1 base pairs per turn. If we take the average helical repeat of free DNA as 10.6 base pairs per turn, the difference in twist between the DNA in the core particle and free DNA would be $146/10.1 - 146/10.6$ that is, 0.68. Thus the linking-number change associated with the core particle $\Delta Lk = -1.7+0.68 = -1.02$ as observed.

Now we can suggest how a DNA loop packaged as a 34-nanometer supercoil of nucleosomes (see Figure 6) could be unwound during interphase. If negative supercoils previously constrained by the nucleosomes are released, then negative supercoiling must be taken up by the linker DNA joining one nucleosome to another. This negative supercoiling would favor the unwinding of a 34-nanometer supercoil of nucleosomes. As suggested above, the acetylation of histones releases DNA that was negatively supercoiled about the histone octamer, presumably by unwinding DNA from the ends of the nucleosome.

The reverse process of chromosome condensation to the metaphase configuration (see Figure 1) requires that the 34-nanometer supercoil be further coiled into higher orders of coiling(s). Perhaps histone-H1 phosphorylation introduces additional supercoiling into a packaged DNA loop causing the higher order of coilings of metaphase chromosomes.

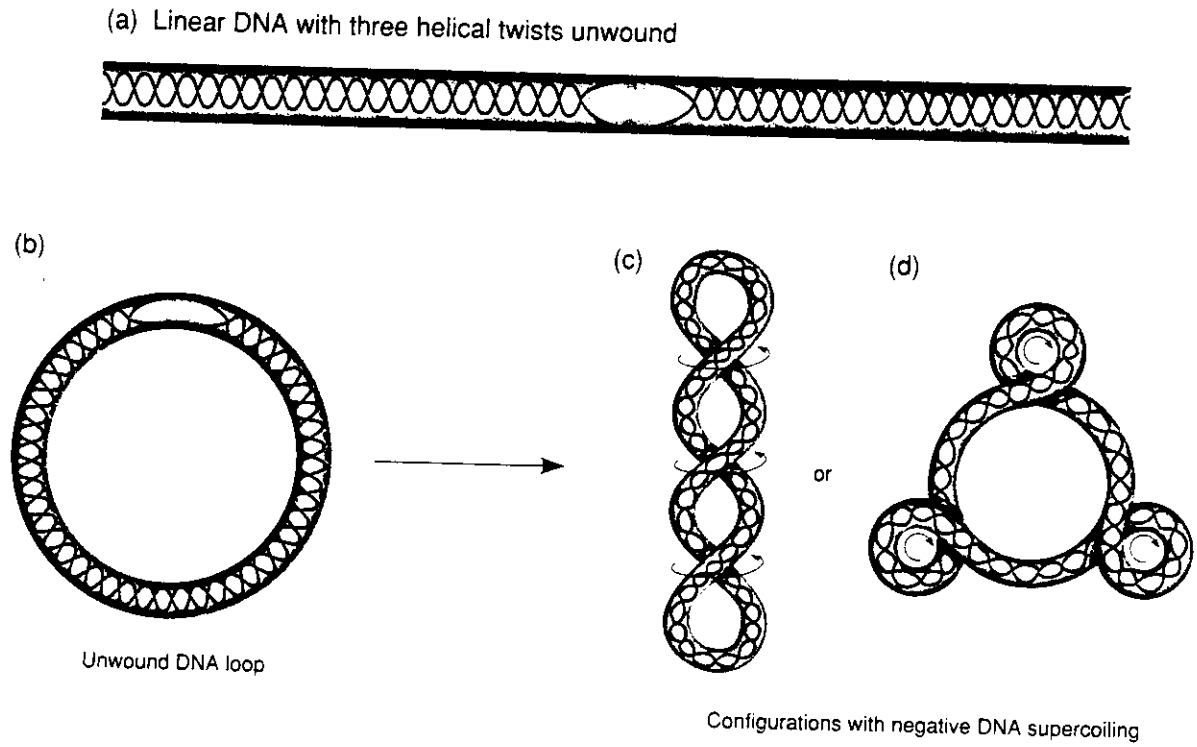


Figure 8. Negative Supercoiling of a Closed DNA Loop
If three helical twists of a linear, double-stranded DNA molecule are unwound as shown in (a) and the ends are then joined, the resulting DNA loop can take on the configurations shown in (b), (c), and (d). All three have the same linking number. In (b) the circular molecule is missing three helical twists that would be present in the normal structure. In (c) the three twists are restored and the loop forms a right-handed superhelix with three crossings in (d) the three twists are restored, but the loop forms three extra left-handed loops. Configurations (c) and (d) are referred to as negative DNA supercoiling.

Figure 8 shows in outline the different orders of packaging of DNA loops into the different orders of chromatin structure and into metaphase chromosomes. It appears that the reversible chemical modifications of acetylation and phosphorylation of histones are involved in the structural transitions undergone by a chromosome during the cell cycle. These structural transitions are dictated by the functional requirements of chromosomes.

Conclusion

Despite recent advances in understanding centromeres and telomeres, we are still a long way from understanding the relationships between structure and function of eukaryotic chromosomes. Relevant to this understanding will be the sequence information from the Human Genome Project. Although much interest is now focused on the mapping and sequencing of genes, the noncoding DNA regions clearly

contain information involved in the organization and functions of chromosomes. The constancy of the banding patterns of individual metaphase chromosomes reflects a highly reproducible pattern of long-range DNA folding, most probably directed by specific DNA-protein interactions and possibly by unusual DNA structures such as bent DNA segments. Superimposed on the very long-range order suggested by banding patterns is the packaging of the DNA loops by the histones together with other structural and regulatory proteins.

The existence of several subtypes of each histone raises the possibility that DNA loops containing different gene families could be packaged with different types of histones according to the requirements of the different cells. DNA control regions of active genes must be packaged in a fashion that makes them accessible to gene-regulating proteins, whereas regions containing permanently repressed genes of a particular cell type may be packaged so that they are inaccessible to such proteins. Such packaging may also determine the availability of DNA regions to chemical damage. Thus a knowledge of the organization of chromosomes is essential to an understanding of the central processes of cell differentiation and the orderly development of complex organisms as well as the processes of DNA damage in chromosomes. ■

Further Reading

E. Morton Bradbury, "Reversible Histone Modifications and the Chromosome Cell Cycle." *BioEssays*, Volume 14, No. 1, January 1992.

Morton Bradbury received a bachelor of science degree in physics and a Ph.D. in biophysics from King's College, University of London, in 1955 and 1958, respectively. After completing his postdoctoral research at Courtauld Research Laboratory, he was appointed head of the Department of Molecular Biology at Portsmouth (England) Polytechnic in 1962, where he remained until his appointment at UC Davis in 1979. He became leader of the Life Sciences Division at Los Alamos in 1988. Bradbury's research has been devoted to understanding whether chromosome organization and chromosome structure are involved in determining how a cell looks and behaves: the structure and function of active chromatin; and the process by which chromosomes condense prior to cell division. In pursuing his investigations, Bradbury has combined the results of measurements derived from the use of a wide range of techniques, including optical spectroscopy, nuclear magnetic resonance, x-ray diffraction, electron diffraction, and neutron diffraction. The recipient of numerous award and honors, Bradbury has also chaired a number of scientific organizations, including the British Biophysical Society, the International Council for Magnetic Resonance in Biology, and the Neutron/Biology Committee of the Institut Laue-Langevin. Bradbury is a member of HERAC and a member of the HERAC subcommittee on structural biology.

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 multi-domain 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.

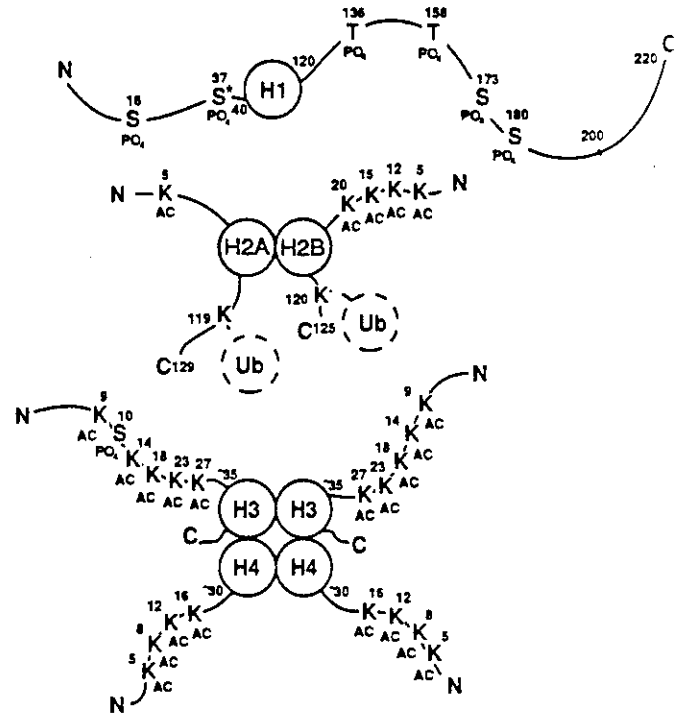


Fig. 1. Outline structures of histone H1, and the (H2A, H2B) dimer and ($H3_2$, $H4_2$) tetramer showing the well-defined globular domains, the basic flexible N and C-terminal domains and sites of reversible acetylation, phosphorylation and ubiquitinations.

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

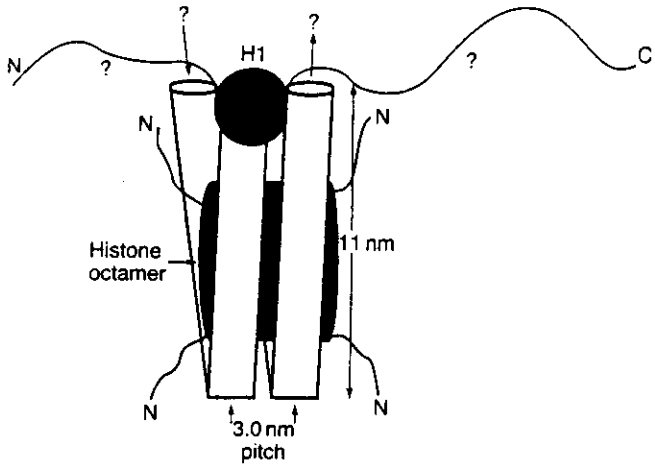


Fig. 2. Model for the chromatosome based on the known structure of the nucleosome core particle and the probable binding probe of histone H1.

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; hyperacety-

lated 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, trichostatin 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.

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: Uncoordinated 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 5Å 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**, 19-34.

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., Mann, 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 I 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**, 19848-19852.

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**, 17174-17179.

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.

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

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.

Chromosoma (1993) 102:509–518

CHROMOSOMA

Organization of centromeres in the decondensed nuclei of mature human sperm

Andrei O. Zalensky[1], **John W. Breneman**[1], **Irina A. Zalenskaya**[1], **B.R. Brinkley**[2], **E. Morton Bradbury**[1,3]

[1] Department of Biological Chemistry, School of Medicine, University of California, Davis, CA 95616, USA
[2] Baylor College of Medicine, Texas Medical Center, Houston, TX 77030, USA
[3] Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received: 25 February 1993; in revised form: 27 May 1993 / Accepted: 18 June 1993

Abstract. The localization of centromeres in mature human sperm was shown by immunofluorescent labeling and nonisotopic in situ hybridization. In the decondensed nucleus structural elements (dimers, tetramers, linear arrays and V shape structures) formed by individual centromeres of nonhomologous chromosomes were observed. They organize the compact chromocenter, which was shown for nuclei decondensed to a low extent. The chromocenter is buried inside the nucleus; in contrast, telomeric regions of chromosomes were tentatively localized on the periphery. Thus, a gross architecture, which can influence selective unpackaging of the paternal genome upon fertilization, exists in human sperm.

Introduction

Mammalian centromeres are distinctive chromosomal regions consisting of specific centromeric proteins (CENP), histones and tandemly repeated alphoid DNA (Schulman and Bloom 1991; Brinkley 1990; Earnshaw 1991; Zinkowski et al. 1991). This nucleoprotein complex has fundamental roles in kinetochore assembly, in the movement of chromosomes during mitosis and meiosis and in sister chromatid exchange (Schulman and Bloom 1991). There is conflicting evidence that during spermatogenesis mammalian centromeres, like other chromosomal nucleoprotein structures, are reorganized and eventually lost. Brinkley et al. (1986), using indirect immunofluorescence and serum from CREST scleroderma patients, have shown that during the maturation of mouse spermatids the number of stained centromeres progressively decreases, individual centromeres cluster into a conspicuous chromocenter and then finally disappear in mature spermatozoa. Virtually the same result was obtained for spermatogenic cells from primates, hamsters and mice (Haaf et al. 1990). In marked contrast centromeres are preserved in the mature spermatozoa of amphibians (Haaf et al. 1990). In other studies discrete, immunofluorescent spots were observed in human (Sumner 1987) and bovine (Palmer et al. 1990) spermatozoa. Recently, a detailed study (Courtens et al. 1992) utilizing immunoelectron microscopy, immunofluorescence and confocal microscopy showed that centromeric proteins (CENP-A, B, C) detach from their DNA loci and are released from nuclei at the same time as histones and transitional proteins during rabbit spermatogenesis. In the mammalian sperm nucleus it is difficult to investigate centromeric existence or topology because of the extreme compaction of most of the DNA into a nucleoprotamine structure. One of the explanations of the previous contradictory observations is not species-specific behaviors of centromeric complexes, but the result of different accessibility of epitopes to antibodies depending on experimental conditions.

Several fundamental questions remain to be answered: (i) Are centromeres preserved through the late stages of spermatogenesis and passed on as nucleoprotein complexes to the next generation or are they formed de novo after fertilization? (ii) Do centromeric proteins remain associated with their corresponding DNA sequences? (iii) If centromeres are preserved do they associate into higher order structures similar to the chromocenter as described for midsperamtid stages in mouse (Brinkley et al. 1986)? (iv) Do centromeres occupy specific position(s) within nuclei, i.e., is there a nuclear architecture in mammalian sperm (Powell et al. 1990; Ward and Coffey 1991)?

We approached these questions using purified nuclei of ejaculated human sperm that had been subjected to heparin decondensation. In contrast to the uniform distribution of protamine we showed that the centromeric protein A (CENP-A) is preserved in mature human sperm and localized within decondensed nuclei into discrete regions. The number of putative centromeres approaches 23, the value expected for a haploid genome. Using nonisotopic fluorescent in situ hybridization (FISH) we demonstrated that centromeric DNA se-

Communicated by: W.C. Earnshaw

Correspondence to: A. Zalensky

quences are also located in morphologically similar discrete regions in sperm nuclei. We describe the dispersion of the compact chromocenter formed by interacting centromeres of nonhomologous chromosomes to individual elements during progressive decondensation of nuclei. Of considerable interest are the patterns of both centromeric DNA and CENP-A localization in highly swollen nuclei, which show the existence of dimers, tetramers and linear arrays of nonhomologous centromeres.

Further we showed that the compact, localized distribution of centromeric DNAs in sperm nuclei differs significantly from the dispersed one of telomeric DNAs. Telomeric DNA hybridization rapidly disappears upon micrococcal nuclease and DNAase I digestion of sperm nuclei. In contrast centromeric DNA was resistant to these nucleases. This observation suggests a more external location for telomeres compared with an internal location for centromeres and the existence of a well defined nuclear architecture in human sperm nuclei.

Materials and methods

Sperm nuclei isolation and decondensation. Nuclei were isolated from semen obtained from healthy donors according to a modified (Zalensky et al. 1993) procedure of Gusse et al. (1986). Briefly, cells were washed with 0.25 M sucrose, 0.1 M NaCl, 10 mM EDTA, 20 mM Tris-HCl, pH 7.5, 0.1 mM PMSF (phenylmethylsulfonyl fluoride), 2 µg/ml TLCK (tosyl-lysine chloromethylketone), and sonicated to remove tails. Nuclei were further purified by differential centrifugation and partially demembraned by treatment with 10 mM CHAPS {3-[(3-cholamidopropyl)-dimethylammonio]-1-propanesulfonate}. Nuclei were decondensed by incubation in PBS containing 1 mM DTT (dithiothreitol) and 0.02–1 mg ml (typically 0.05 mg/ml) of heparin at room temperature during the indicated periods of time. To evaluate the degree of decondensation nuclei were put onto microscopic slides precovered with 0.015% polyethylenimine and allowed to sit for 5 min. Afterwards the excess suspension was aspirated, slides were washed with PBS and stained for DNA with DAPI (4',6-diamidino-2-phenylindole). Similarly prepared preparations of nuclei were used both for immunofluorescence and hybridization experiments. Fluorescent nuclei were photographed on Kodak TMAX 3200 film.

Protein isolation, gel electrophoresis and immunoblotting. Samples of total nuclear proteins for gel electrophoresis were obtained by lysis of nuclei in SDS-containing sample buffer (2% SDS, 0.125 M Tris-HCl, pH 6.8, 30% glycerol). Acid soluble sperm nuclear proteins were extracted with 0.4 N HCL after reduction with β-mercaptoethanol and precipitated with 6 volumes of acetone. Proteins were separated in 15% SDS-polyacrylamide gels (Laemmli 1970) and were transferred to Immobilon P membrane (Millipore) using a Jensen semidry blotter for 1–2 h at 0.8 mA/cm^2. Transfer buffer was 50 mM Tris, 40 mM glycine, 0.04% SDS, 20% methanol. Blocking, treatment with antibodies and washing of membranes were performed as described (Towbin et al. 1979). Primary antibodies (centromere specific CREST serum at 1:2000 dilution) were detected using secondary antibodies conjugated to alkaline phosphatase and NPT/BCIP (nitroblue tetrazolium/5-bromo-4-chloro-3-indolyl phosphate) as substrate.

Nuclease digestion. Sperm nuclei were suspended in the decondensation buffer supplemented with 1 mM $CaCl_2$ and excess (5 units/µl per 1 mg/ml of DNA in the suspension) of DNAase I or micrococcal nuclease. The suspension was incubated for 30 min at room temperature, then put onto microscopic slides and processed further for hybridization, as described. In control experiments with HeLa nuclei it was shown that both nucleases are active in the presence of heparin in the concentrations used.

Immunocytochemistry. Preparation of HeLa cells for indirect immunofluorescence was as described by us earlier (Breneman et al. 1993). Sperm nuclei were decondensed for 30 min and put onto microscopic slides as described above. Sperm nuclei were immunolabeled simultaneously with mouse monoclonal antibodies against human protamine 1 and CREST serum. Both primary antibodies were at 1:100 dilution. Secondary antibodies were used at 1:60 dilution. A detailed protocol was described earlier (Breneman et al. 1993). Fluorescent nuclei were photographed on Kodak Ektachrome 800/1600 or TMAX 3200 film.

Fluorescent in situ hybridization (FISH). Sperm nuclei were prepared on microscopic slides as described above and dehydrated in a series of ethanol (70%, 80%, 95%) washes. Slides were allowed to dry and afterward were incubated in 70% formamide, 2 × SSC at 70° C for 2 min. (1 × SSC is 0.15 M NaCl, 0.015 M sodium citrate.) After denaturation slides were successively rinsed in cold (70%, 80%, 90%, 100%) ethanol solutions. Hybridization with biotinylated DNA probes was in 50% formamide, 2 × SSC at 37° C for 12–16 h. Post hybridization washes (low stringency), detection with fluorescein labeled avidin and amplification with anti avidin antibodies, were as recommended by the Oncor manual and using Oncor reagents. DNA was counterstained either with propidium iodide or DAPI.

Antibodies and hybridization probes. Mouse monoclonal antibodies against human protamine 1 were provided by Dr. R. Balhorn (Lawrence Livermore National Laboratory). Anticentromere antibodies were from patients with CREST scleroderma, and in particular serum S.H., which was described earlier (Brinkley et al. 1986; Balczon and Brinkley 1990). Secondary antibodies (Calbiochem) were goat anti mouse IG rhodamine conjugated and goat anti human IG FITC (fluorescein isothiocyanate) conjugated for immunofluorescence. For the immunoblotting we used goat anti human IG alkaline phosphatase conjugated antibodies (Calbiochem). Biotinylated DNA probes "all human centromeres" and "all human telomeres" were from Oncor.

Results

Persistence of CENP-A in human sperm nuclei

To evaluate whether some specific proteins of somatic cell centromeres are preserved in mature human sperm nuclei we used sera from patients with CREST scleroderma (Cox et al. 1983; Earnshaw and Rothfield 1984). We screened about 20 different sera and chose three samples with almost identical antigenic properties. One of these three, S.H., had been to some extent characterized by us previously (Balczon and Brinkley 1990). This serum detected three polypeptides with apparent Mrs of 19000, 23000 and 28000 in HeLa nuclear extracts (Fig. 1a, lane 4) and produced the characteristic pattern of HeLa metaphase chromosome centromere labeling (Fig. 1b). The polypeptide of Mr 17000–19500 is centromeric protein A (CENP-A) (Earnshaw and Rothfield 1984; Guldner et al. 1984; Palmer et al. 1987). The range of molecular weights reflects differences in estimations given by the authors; according to our data the Mr value of 19000 also reported by Guldner et al. (1984) is a more realistic one. Polypeptides of Mr 23000 and 28000 are not components of the centromere kinetochore complex

(Guldner et al. 1984). CENP-A was recognized also by a number of the other CREST sera tested, for example, serum E.K., lane 2, Fig. 1a.

We have previously described (Zalensky et al. 1993) the isolation of highly purified, sonication resistant nuclei from human sperm. When total nuclear proteins were separated by SDS-PAGE (SDS-polyacrylamide gel electrophoresis) and immunoblotted with the above described sera, a strong reaction with an Mr 19000 protein was observed (Fig. 1a, lanes 3 and 5). The immunoblot results indicate also that serum S.H. was largely monospecific for this protein in human sperm. Several high molecular weight polypeptides recognized in HeLa nuclear extracts are absent from sperm nuclei (compare lanes 4 and 5 in Fig. 1a). By SDS-PAGE the sperm protein is identical in mobility to CENP-A from HeLa nuclei, it is acid soluble (Fig. 1d,d'), which is a characteristic feature of CENP-A (Palmer et al. 1987), and the serum specific to this protein produces characteristic centromeric staining in sperm nuclei (see below). Therefore we assume that the Mr 19000 protein of human sperm nuclei is CENP-A. Immunoblotting of acid soluble proteins after separation by PAGE containing acetic acid, urea, Triton X100 shows that the mobilities of the CENP-A from sperm and HeLa cells are slightly different (data not shown) indicating that the sperm protein might be represented by either a specific variant or a modified form. Work on isolation and sequencing of human sperm CENP-A is in progress.

Establishment of the decondensation assay

Immunoblotting data clearly show that CENP-A is present in nuclei of mature human sperm. At the same time, in a number of previous reports authors have failed to recognize CENP-A in mature mammalian sperm cells (Brinkley et al. 1986; Haaf et al. 1990; del Mazo et al. 1987; Courtens et al. 1992). Our suggestion was that these attempts were unsuccessful because of the extremely condensed state of sperm chromatin, which makes epitopes unavailable for reaction with antibodies. To overcome this problem we tried several commonly used methods of sperm nuclei decondensation (treatment with 1–10 mM DTT, mild digestion with trypsin, hypotonic shock). None of them resulted in reproducible centromeric staining. Finally, we arrived at the protocol of treating demembraned human sperm nuclei with solutions containing heparin and DTT, which is suitable for both immunolocalization of nuclear proteins and in situ hybridization (see below). Such decondensation increases the nuclear volume (length of the longer elliptical axis increases up to tenfold) while preserving overall nuclear morphology (Fig. 1c). Decondensation by low (0.02–0.05 mg/ml) concentrations of heparin is largely isomorphous since it does not change: (i) the characteristic ellipsoidal shape of the human sperm nucleus (ratio of axes remains 1.7 ± 0.2); (ii) the characteristic z-dimension profile, with a flattened anterior part of the nuclei, as revealed by atomic force microscopy (data not shown). This procedure was based on earlier observations (Reyes et al. 1989; Jager et al. 1990) of the effect of heparin on human sperm cell decondensation. It is important to note that the heparin concentrations used in this work do not lead to a noticeable loss of nuclear proteins (Fig. 1d, lanes 2, 4 and 5). Most important, CENP-A remains in the nuclei as shown by immunoblotting (Fig. 1d') and immunofluorescence (Fig. 2c). Protamines do not dissociate from sperm nuclei as demonstrated by the same method (Fig. 2a). In summary, the sperm nucleus differs significantly from somatic nuclei with respect to heparin treatment. Histones (Paulson and Laemmli 1987) and centromeric proteins (Valdivia and Brinkley 1985) were removed from the latter by 0.02 mg/ml and 2 mg/ml heparin respectively.

Localization of CENP-A in decondensed sperm nuclei

We used the S.H. CREST serum described above to localize CENP-A in human sperm nuclei that were decondensed by heparin treatment. Figure 2 shows double immunolabeling with the S.H. serum and monoclonal antibodies against human protamine 1. Protamine, which is the major structural protein of human sperm chromatin, is evenly distributed throughout nuclei (Fig. 2a). The pattern of protamine labeling is exactly superimposed on DNA as shown by DAPI staining (Fig. 2a, b). In contrast, CENP-A is localized in defined loci (Fig. 2c). In some nuclei spots apparently corresponding to individual centromeres tend to group into aggregates, resembling the chromocenters described for midspermatid stages in some mammals (Brinkley et al. 1986; Courtens et al. 1992). Heparin concentration optimal for CENP-A localization was 0.05 mg/ml. Nuclei decondensed to a lesser extent by lower heparin concentrations do not produce reproducible labeling, though larger centromere agregates were occasionally observed. We estimated the total number of individual stained spots by inspection of about 100 nuclei that were decondensed to a higher degree (see Fig. 4a for an example). The number of CENP-A loci per nucleus approached 23, the number of chromosomes in the haploid human genome.

Localization of centromeric DNA

To determine whether the pattern of CENP-A localization differs from that of centromeric DNA we used FISH. A biotinylated probe complementary to all human centromeric sequences was hybridized with heparin untreated and decondensed nuclei (Fig. 3). Only a small fraction of nondecondensed nuclei gave a positive hybridization response (Fig. 3a). Such nuclei represent about 5% of the population and, we assume, are characterized by a less compact state of chromatin. Heterogeneity of the mature human sperm cells with respect to differences in DNA condensation is a well-known phenomenon (Bedford et al. 1973), though its biological function is unclear.

When the nuclei had been subjected to mild decondensation (0.02 mg/ml heparin) prior to FISH the effi-

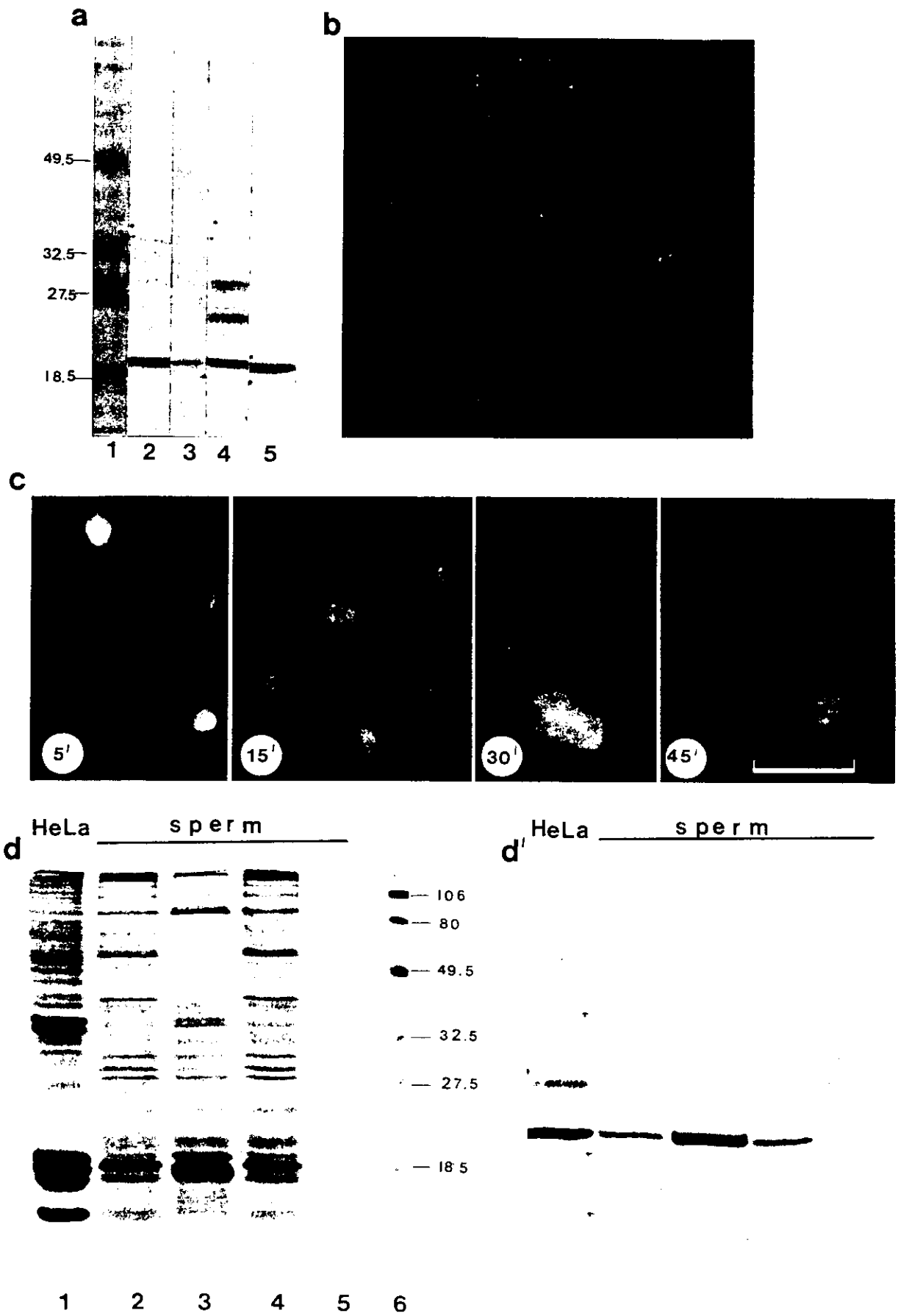

a
49.5
32.5
27.5
18.5
1 2 3 4 5
b
c
5'
15'
30'
45'
d
HeLa
sperm
106
80
49.5
32.5
27.5
18.5
1 2 3 4 5 6
d'
HeLa
sperm

Fig. 1 a–d

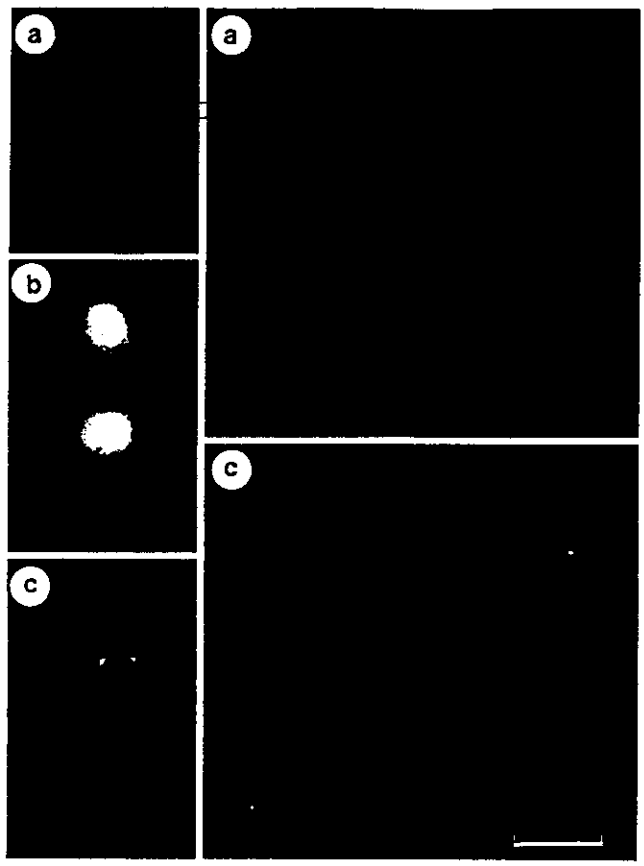


Fig. 2a–c. The even distribution of protamines (**a**) contrasts with the spot-like localization of CENP-A (**c**) within decondensed human sperm nuclei as demonstrated by double indirect immunofluorescence. DNA is counterstained with DAPI (**b**). Bar represents 5 μm

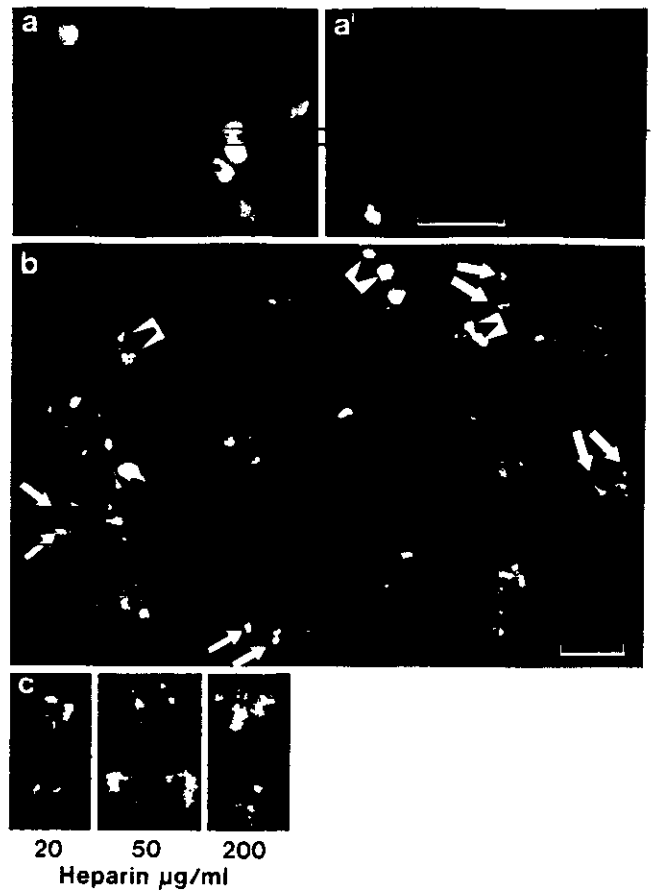


Fig. 3a–c. Non-radioactive fluorescence in situ hybridization (FISH) demonstrates localization of centromeric DNA within specific regions of human sperm nuclei. **a, a'** Sperm nuclei without heparin decondensation (**a** DAPI staining, **a'** FISH). **b** FISH of nuclei decondensed by 0.02 mg/ml heparin. Hybridization signals are yellow, total DNA is counterstained with propidium iodide, red. *Arrows* and *double arrows* indicate examples of the characteristic patterns of centromeric DNA localization. **c** Typical FISH patterns after decondensation of nuclei by increasing heparin concentrations. Bar represents 5 μm

ciency of hybridization increased so that almost all nuclei produced a hybridization response (Fig. 3b). Because of the random orientation of nuclei on the support and intrinsic heterogeneity in DNA condensation, localization patterns seem to be nonuniform. At the same time, several characteristic patterns of hybridization are seen repeatedly and reproducibly – dense linear arrays forming an open V shape and two-spot formations, indicated by arrows and double arrows respectively in Fig. 3. The same types of hybridization patterns were noticed in a fraction of intact sperm cells subjected to hybridization without heparin pretreatment (A.O. Zalensky et al., manuscript in preparation). We think that this is good indication that the nuclear isolation and heparin decondensation procedures do not produce a specific spatial rearrangement of DNA sequences within sperm nuclei.

Packaging of centromeric DNA into compact formations is most prominent in nuclei decondensed to a low extent. When the concentration of heparin is increased nuclei expand and the progressive dissociation of the centromeric network is observed (Fig. 3c). The data clearly show that centromeric DNAs are concentrated within defined regions of the nucleus. This observation suggests that: (i) a nonrandom spatial positioning exists within the sperm nuclear volume for these sequences; and (ii) nonhomologous centromeric DNAs are packed

Fig. 1a–d. Identification of CENP-A among nuclear proteins of mature human sperm using CREST serum, and establishment of decondensation assay. **a** Total nuclear proteins of HeLa cells (lanes *2*, *4*) and human sperm (lanes *3*, *5*) separated by SDS-PAGE (SDS-polyacrylamide gel electrophoresis) and immunoblotted with the CREST serum E.K. (lanes *2*, *3*) and S.H. (lanes *4*, *5*); lane *1*, molecular weight markers (kilodaltons). **b** Indirect immunofluorescent labeling of HeLa metaphase cell with CREST serum S.H. provides the characteristic double-dot pattern of homologous chromosome staining. **c** Time course (5–45 min) of decondensation of human sperm nuclei by treatment with 0.05 mg/ml heparin. Nuclei stained for DNA with DAP (4',6-diamidino-2-phenylindole); all panels of the figure are at the same magnification. Bar represents 5 μm. **d, d'** Coomassie stained gel after SDS-PAGE (**d**) and corresponding immunoblot treated with the CREST serum S.H. (**d'**). Total nuclear proteins of HeLa (lane *1*) and sperm (lane *2*); acid soluble nuclear proteins of sperm (lane *3*); total sperm nuclear proteins remaining after treatment with 0.05 mg/ml Heparin (lane *4*) and solubilized by such treatment (lane *5*)

together in several chromocenters (Brinkley et al. 1986), most probably by specific protein-protein interactions.

Elements of internal organization of centromeric complexes

The complex internal organization of the chromocenter(s) was further revealed from detailed inspection of both the immunolabeling and FISH patterns in nuclei subjected to greater decondensation (0.05–0.08 mg/ml heparin). Upon nuclear expansion compact chromocenter(s) stretched so that apparent individual centromeres became distinguishable (Fig. 4). In many nuclei characteristic dimer and tetramer arrangements of nonhomologous centromeres could be observed and quite often the spots, corresponding to individual centromeres, are arranged in several linear arrays (shown by arrows in Fig. 4). In most cases there are two such arrays per nucleus. We suggest that they correspond to the dense V shape localization of centromeres in less decondensed nuclei and that the same arrays give rise to the bright two-spot fluorescence when viewed end on (Fig. 3; see also examples and model Fig. 6). We believe the observed patterns of the organization of centromeres in human sperm nuclei are not artifacts. Most convinc-

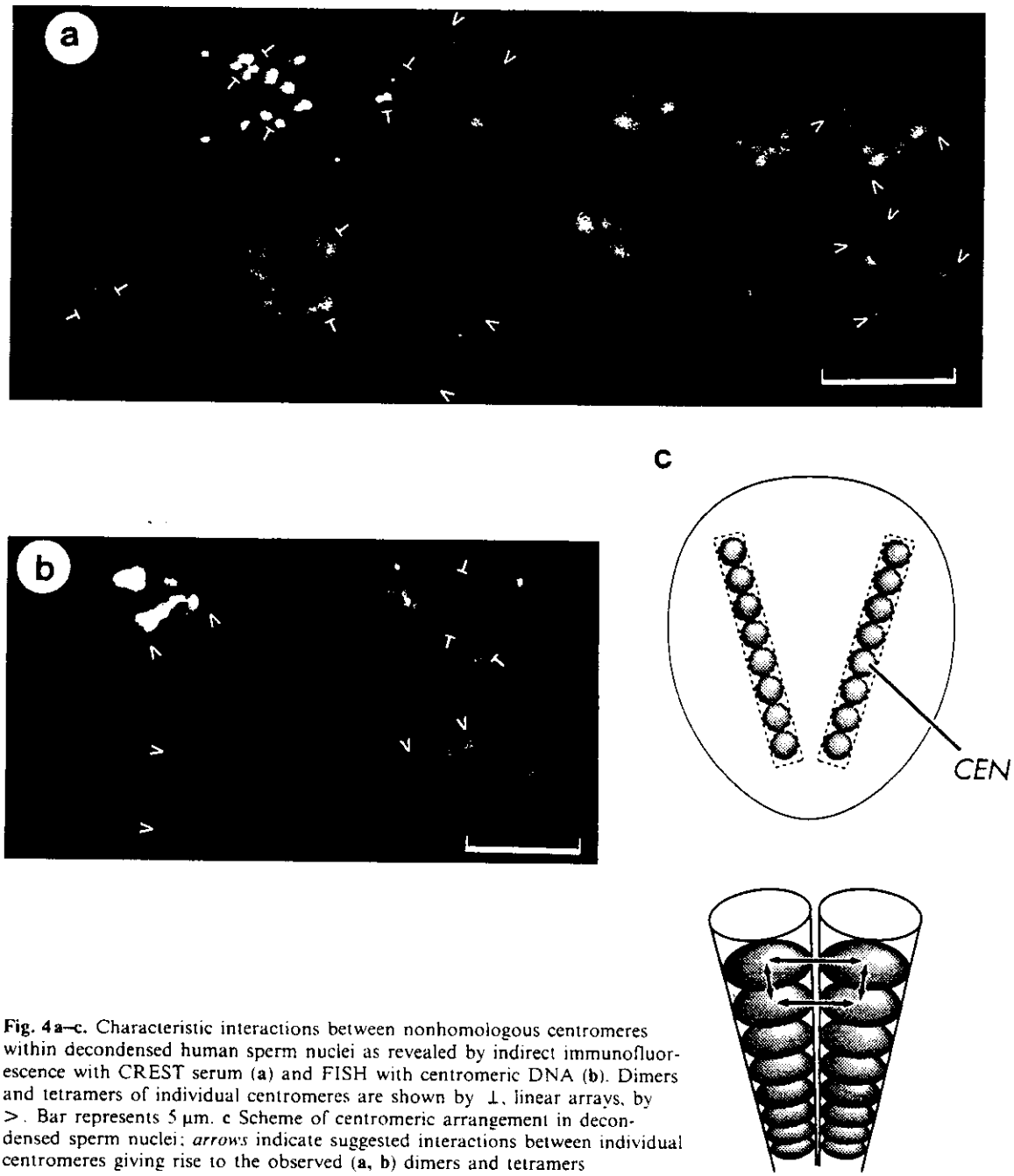


Fig. 4a–c. Characteristic interactions between nonhomologous centromeres within decondensed human sperm nuclei as revealed by indirect immunofluorescence with CREST serum (**a**) and FISH with centromeric DNA (**b**). Dimers and tetramers of individual centromeres are shown by ⊥, linear arrays, by >. Bar represents 5 μm. **c** Scheme of centromeric arrangement in decondensed sperm nuclei; *arrows* indicate suggested interactions between individual centromeres giving rise to the observed (**a**, **b**) dimers and tetramers

ing are the very similar patterns obtained by immunofluorescent labeling of CENP-A and by FISH. Their identity suggests that centromeric proteins remain associated with their specific DNA counterpart in mature human sperm.

We propose the dimer interactions between centromeres of individual chromosomes to be essential and elementary ones; they are responsible for maintaining the observed tetramers and linear arrays (scheme Fig. 4c). Probably the same interactions support the compact chromocenter in the nondecondensed sperm nucleus. The existence of dimeric interactions between nonhomologous centromeres has also been demonstrated for amphibian sperm (Haaf et al. 1990).

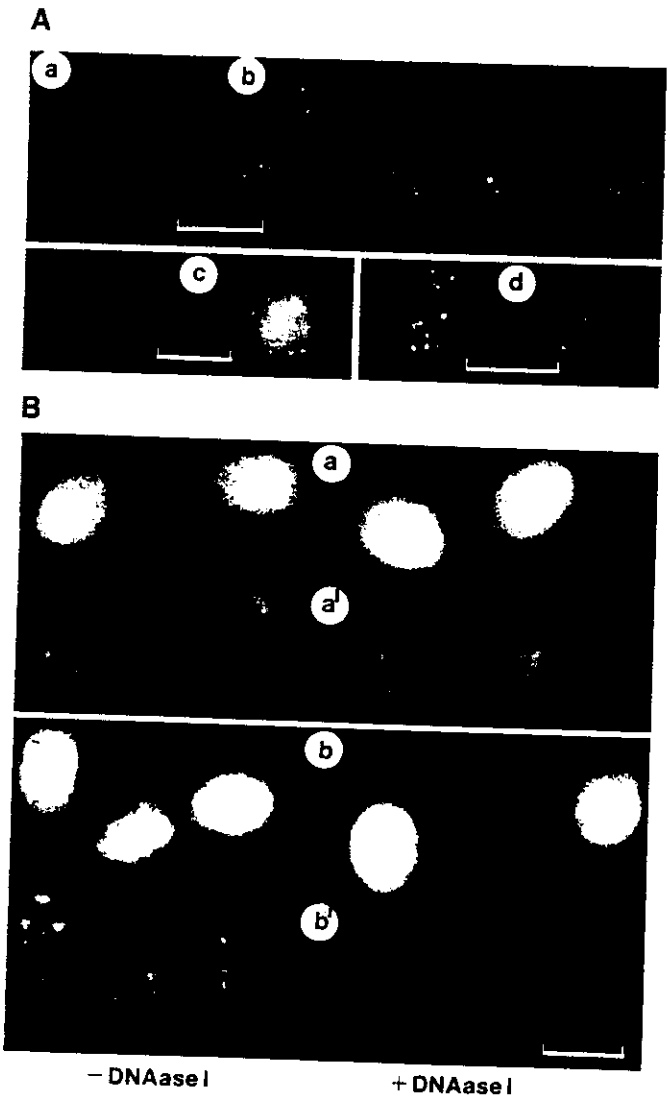


Fig. 5A, B. Localization of telomeric sequences in human sperm nuclei and different accessibility of telomeric and centromeric DNA to DNAase I digestion. **A** FISH with telomeric probe; total DNA counterstained with propidium iodide (panels *a–c*) or DAPI (*d*). Panel *a*, nondecondensed nuclei; *b–d*, nuclei decondensed by 0.05 mg/ml heparin. **B** Effect of pretreatment with DNAase I of sperm nuclei on hybridization with centromeric (panels *a*, *a'*) and telomeric (*b*, *b'*) sequences. Panels *a*, *b*, total DNA stained with DAPI; *a'*, *b'*, FISH patterns in the same nuclei. Bar represents 5 μm

Differences in localization of centromeric and telomeric sequences

We were also interested in how other distinctive chromosomal domains, namely telomeres, are localized in comparison with centromeres. To answer this we used the same approach – FISH with the biotinylated human telomere DNA sequence as a probe (Fig. 5). Telomeric DNAs were not accessible for hybridization in condensed human sperm nuclei (panel a, Fig. 5A). A positive result appeared only after nuclei had been decondensed with heparin (panel b, Fig. 5A). In this respect the behavior of telomeric and centromeric sequences was the same. Otherwise, the telomeric probe produced a

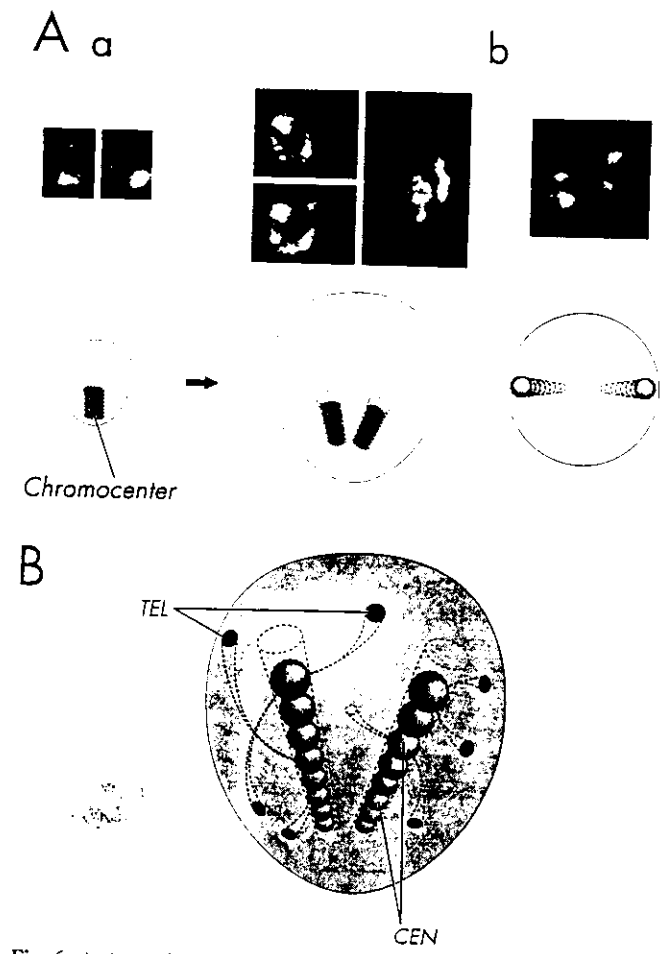


Fig. 6. A A model of the chromocenter in the nucleus of mature human sperm based on FISH data. Panel *a*, compact chromocenter; nuclei decondensed to a low extent (heparin < 0.02 mg/ml). Panel *b*, partially dispersed chromocenter; nuclei decondensed to a higher extent (0.05 mg ml heparin). Arrangement of nonhomologous centromeres in linear arrays (two-dot patterns when viewed end on) is shown. The two FISH patterns in the left part of the central panel are obtained from the same nucleus but in different focal planes. Examples of FISH in this figure are at different magnifications and the model is not to scale. **B** Hypothetical model of nuclear architecture in decondensed human sperm (not to scale). Two of several possible centromeric arrays are shown and an external localization of telomeres relative to centromeres is emphasized. *Dashed lines* hypothetical paths of chromosomes, which do not take into consideration packaging of DNA into nucleoprotamine structures

strikingly different pattern of hybridization in comparison with centromeres. Telomeric DNAs were localized in many small and dispersed spots.

The hybridization signal for the telomeric probe was weaker than for the centromeric probe, partially due to the scattered character of telomeres distribution, which made it more difficult to register and document their position. To overcome this problem an additional round of amplification with biotinylated anti-avidin antibodies was used in some experiments, and nuclei were counterstained with DAPI instead of propidium iodide, which allowed FITC fluorescence to be cut out using an appropriate filter (compare panels c and d, Fig. 5A). The second type of experimental problem to be mentioned is connected with the localization of (comparatively weak and small) telomeric spots in several different focal planes, so that the overall picture of telomeric distribution is difficult to obtain without confocal microscopy.

In some nuclei the telomeric hybridization signals were concentrated toward the nuclear edge (Fig. 5A, panels c, d). This gave us some indication of a probable peripheral localization of the telomeres. Supporting evidence for this comes from differential accessibility of telomeric and centromeric sequences to the action of endogenous nucleases. Sperm nuclei were treated with an excess of DNAase I and afterward were hybridized with the telomeric and centromeric probes (Fig. 5B). Centromeric DNA hybridization signals remain identical both quantitatively and qualitatively with the nuclease untreated control (Fig. 5B, panel a'), whereas the telomeric hybridization was completely eliminated (Fig. 5B, panel b'). The same result was obtained using micrococcal nuclease pretreatment (data not shown). DNAase I and micrococcal nuclease possess low (and different) nucleotide sequence specificity, therefore we assume the contrasting solubility rates reflect different sterical hindrances of telomeric and centromeric regions of chromosomes in the human sperm nucleus. Based on these data and patterns of hybridization we propose an external localization of telomeres opposite to the internal one of the centromeric complexes.

Discussion

We have established a method to uniformly decondense human sperm nuclei by treatment with heparin in the presence of DTT, while maintaining their characteristic ellipsoidal form and without noticeable loss of proteins. This method is based on the documented effects of heparin and other polyanions on mammalian sperm nuclei (Reyes et al. 1989; Jager et al. 1990). The exact mechanism of heparin action is unknown but it is proposed (Jager et al. 1990; Villepoteau 1992) that it weakens DNA-basic protein interactions causing DNA to stretch, which results in nuclear volume increase. It is also speculated that heparin-like compounds (i.e. glucosaminoglucans of follicular fluid) may play a role in sperm nuclear decondensation within the oocyte. It is important that the method allows both indirect immunofluorescent labeling of sperm nuclear proteins and FISH of DNA. Owing to the highly condensed state of mammalian sperm chromatin and the in situ hybridization requirement that all DNA sequences be equally accessible for probing, there have been few reports of successful applications of FISH to sperm chromosomes (Wyrobek et al. 1990; Powell et al. 1990; Coonen et al. 1991). Weak unreliable signals or their absence have also been reported.

We have demonstrated that centromeric DNA and CENP-A are similarly localized in discrete areas within decondensed human sperm nuclei. The number of centromeric loci in highly decondensed nuclei approaches the value expected for the haploid human genome. Thus centromeres are preserved in mature, ejaculated sperm from man and not formed de novo following fertilization. There are contradictory results concerning the persistence of centromeres in vertebrate spermatozoa obtained by indirect immunolabeling of CENP-A. Characteristic dot-like patterns were shown in spermatozoa of amphibians (Haaf et al. 1990) and the bull (Palmer et al. 1990), but not in rodents (Brinkley et al. 1986; Haaf et al. 1990; Courtens et al. 1992) and primates (Haaf et al. 1990). There are both positive (Sumner 1987) and negative (Haaf et al. 1990) results for human sperm. We propose that centromeres are preserved during spermatogenesis of all vertebrates and are not an evolutionary mosaic. The negative results, especially those obtained with extremely compact mammalian sperm nuclei, probably are the consequence of inadequate methods of decondensation. As was demonstrated in our study most of the commonly used procedures with the exception of heparin treatment did not result (at least in our hands) in subsequent immunolocalization of CENP-A.

We have shown also the retention of the Mr 19000 acid soluble CENP-A in mature human sperm nuclei by biochemical methods. Preservation of the homologous protein in boar sperm was previously demonstrated by Palmer et al. (1990). It is important to know whether other proteins ascribed to the centromere/kinetochore complex (CENP-B, CENP-D etc) are preserved in human sperm as well. The answer could be obtained using the corresponding antibodies and the approach established in this study.

In situ hybridization data indicate that the centromeric DNA sequences belonging to all chromosomes are packed together into a compact chromocenter in mature sperm nuclei that are decondensed to a very low extent (Fig. 6A, panel a; see also data of Fig. 3). Clustering of individual nonhomologous centromeres into a conspicuous chromocenter has been shown by indirect immunofluorescence in mouse midspermatid nuclei (Brinkley et al. 1986). Upon increased sperm nuclei swelling the chromocenter partially decondenses producing characteristic FISH patterns of linear arrays (or double dots when viewed end on) (Fig. 6A, panel b). At this initial step of nuclear swelling centromeres cannot be localized by immunofluorescent techniques. We think this is because CENP-A epitopes are still blocked, while higher affinity DNA probes to corresponding target sequences (as well as an additional step of formamide denaturation) allow the successful application of FISH. The chro-

mocenter has a complex three-dimensional organization. Its existence is indicated by different FISH patterns registered by conventional microscopy from the same nucleus but in different focal planes (panel b, Fig. 6A). The three-dimensional structure of the chromocenter is now under investigation by confocal microscopy.

Upon further decondensation of the nucleus (0.05–1.0 mg/ml of heparin) an internal structural organization of centromeric complexes has been demonstrated. Both DNA in situ hybridization and protein immunolabeling in decondensed nuclei show similar distinct arrangements of nonhomologous centromeres in dimers, tetramers and linear arrays (Fig. 4). The paired arrangement of nonhomologous centromeres probably reflects an essential interaction previously observed in the mature sperm from several amphibians (Haaf et al. 1990) and during some stages of spermatogenesis in the mouse (Brinkley et al. 1986). Although in the latter work the staining of centromeres in mature sperm was not shown, the authors postulated a fusion of nonhomologous centromeres into pairs as a characteristic feature of postmeiotic cells. The tetramer interactions and linear arrays are newly described elements of structural organization. It is to be noted that similar symmetrical linear structures have been observed (Hadlaczky et al. 1986) in the interphase nuclei of rat-kangaroo and Indian muntjac cells.

Our data show that centromeric DNA sequences are localized in particular regions inside the sperm head. This implies that specific constraints exist for the packing of chromosomes in the limited volume of the sperm nucleus. DNA condensation during spermatogenesis is nonrandom and probably important for selective unpackaging of the paternal genome in the egg cytoplasm. The nonrandom spatial organization of the genome is characteristic of bovine sperm nuclei, where there is a preferential localization of centromeric sequences in the equatorial region of the head (Powell et al. 1990).

The existence of long range spatial order was further shown by localization of telomeric sequences. In contrast to the compact and most probably internal localization of centromeric DNA, DNA of the individual telomeres appears to be distributed in dispersed dots presumably placed closer to the nuclear periphery. More external positioning of telomeres in comparison with centromeres was supported by their higher susceptibility to DNAase I and micrococcal nuclease action (Fig. 5B). Resistance of centromeric DNA to DNAase I digestion has previously been shown for mature rat sperm heads but not for spermatocytes and spermatogenic cells (Moens and Pearlman 1989).

Finally, we propose a model of the gross architecture in the decondensed human sperm nucleus (Fig. 6B). This model is based on the experimental data obtained by FISH with uniformly swollen nuclei. It suggests the existence of a compact chromocenter(s) in intact nondecondensed nuclei and in first approximation describes its formation from interacting centromeres of nonhomologous chromosomes. We think that these two features are basically proven by the data presented in this manuscript. The other feature of the proposed model is the relative positioning of centromeric and telomeric regions of chromosomes with the more external location of the latter. We have given some evidence supporting such a localization. Presently we are working on a three-dimensional reconstruction of the hybridization, which allows us to judge the model using more direct data. The model suggests the long range order of DNA packaging in sperm including the existence of certain spatial domains for individual chromosomes. These proposed features might be critical for the manner of chromosomal release upon fertilization, formation of the male pronucleus and even for the regulation of gene expression during the very early stages of development. And most importantly many of these questions could be experimentally tested.

Acknowledgements. This work was supported by the fellowship from the Center of Molecular Cytometry to A.O. Zalensky, DOE DEFG03-88ER-60673 grant to E.M. Bradbury and NIH CA-41424 grant to B.R. Brinkley. The authors thank R. Swank for critically reading the manuscript.

References

Balczon RD, Brinkley BR (1990) The kinetochore and its roles during cell division. In: Adolph KW (ed) Chromosomes: Eukaryotic, prokaryotic and viral, vol 1. CRC Press, Boca Raton, Florida, pp 167–189

Bedford JM, Bent MJ, Calvin H (1973) Variations in the structure, character and stability of the nuclear chromatin in morphologically normal human spermatozoa. J Reprod Fertil 33:19–29

Breneman JW, Yau P, Teplitz RL, Bradbury EM (1993) A light microscope study of linker histone distribution in rat metaphase chromosomes and interphase nuclei. Exp Cell Res 206:16–26

Brinkley BR (1990) Towards a structural and molecular definition of the kinetochore. Cell Motil Cytoskeleton 16:104–109

Brinkley BR, Brenner SL, Hall JM, Tousson A, Balczon RD, Valdivia MM (1986) Arrangement of kinetochores in mouse cells during meiosis and spermatogenesis. Chromosoma 94:309–317

Coonen E, Pieters MHEC, Dumoulin JCM, Meyer H, Evers JLH, Ramaekers FCS, Geraedts JPM (1991) Nonisotopic in situ hybridization as a method for nondisjunction studies in human spermatozoa. Mol Reprod Dev 28:18–22

Courtens J-L, Biggiogera M, Rothfield NF, Burnier M, Fakan S (1992) Migration of centromere proteins in rabbit spermatids. Mol Reprod Dev 32:369–377

Cox JW, Schenk EA, Olmsted JB (1983) Human anticentromere antibodies: distribution, characterization of antigens and effect on microtubule organization. Cell 35:331–339

Earnshaw WC (1991) When is a centromere not a kinetochore. J Cell Sci 99:1–4

Earnshaw WC, Rothfield N (1985) Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma 91:313–321

Guldner HH, Lakomek H-J, Bautz FA (1984) Human anti-centromere sera recognise a 19.5 kD non-histone chromosomal protein from HeLa cells. Cln Exp Immunol 58:13–20

Gusse M, Sautierre P, Belaiche D, Martinage A, Roux C, Dadoune J-P, Chevaillier P (1986) Purification and characterization of nuclear basic proteins of human sperm. Biochim Biophys Acta 884:124–134

Haaf T, Grunenberg H, Schmid M (1990) Paired arrangements of nonhomologous centromeres during vertebrate spermatogenesis. Expt Cell Res 187:157–161

Hadlaczky GY, Went M, Ringertz NR (1986) Direct evidence for non-random localization of mammalian chromosomes in the interphase nucleus. Exp Cell Res 167:1–15

Jager S, Wijchman J, Kremer JJ (1990) Studies on decondensation of human, mouse and bull sperm nuclei by heparin and other polyanions. Exp Zool 256:315–322

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 226:680–685

del Mazo J, Kremer L, Avila J (1987) Centromeric proteins recognized by CREST sera and meiotic chromosome segregation. Chromosoma 96:55–59

Moens P, Pearlman RE (1989) Satellite DNA I in chromatin loops of rat pachytene chromosomes and in spermatids. Chromosoma 98:287–294

Palmer DK, O'Day K, Wener MH, Andrews BS, Margolis RL (1987) A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. J Cell Biol 104:805–815

Palmer DK, O'Day K, Margolis RL (1990) The centromere specific histone CENP-A is selectively retained in discrete foci in mammalian sperm nuclei. Chromosoma 100:32–36

Paulson JR, Laemmli UK (1977) The structure of histone depleted metaphase chromosomes. Cell 12:817–828

Powell D, Cran DG, Jennings C, Jones R (1990) Spatial organization of repetitive DNA sequences in the bovine sperm nucleus. J Cell Sci 96:185–191

Reyes R, Rosado A, Hernandez O, Delgado NM (1989) Heparin and glutation: physiological decondensing agents of human sperm nuclei. Gamete Res 23:39–47

Schulman I, Bloom KS (1991) Centromeres: an integrated protein/DNA complex required for chromosome movement. Annu Rev Cell Biol 7:311–336

Sumner AT (1987) Immunofluorescent demonstration of kinetochores in human sperm. Exp Cell Res 171:250–253

Towbin H, Stahelin T, Gordon J (1979) Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose filters. Proc Natl Acad Sci USA 76:4350–4354

Valdivia NM, Brinkley BR (1985) Fractionation and initial characterization of the kinetochore from mammalian metaphase chromosomes. J Cell Biol 101:1124–1134

Villepoteau B (1992) Heparin increases chromatin accessibility by binding to the trypsin sensitive basic residues in histones. Biochem J 288:953–958

Ward WS, Coffey DS (1991) DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells. Biol Reprod 44:569–574

Wyrobek AJ, Alhborn T, Balhorn R, Stanker L, Pinkel D (1990) Fluorescence in situ hybridization to Y chromosomes in decondensed human sperm nuclei. Mol Reprod Dev 27:200–208

Zalensky AO, Yau P, Breneman JW, Bradbury EM (1993) The abundant 19-KDa protein associated with human sperm nuclei that is related to seminal plasma inhibins. Mol Reprod Dev, in press

Zinkowski RP, Meyne J, Brinkley BR (1991) The centromere-kinetochore complex: a repeated subunit model. J Cell Biol 113:1091–1110

Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 5199–5203, May 1995
Biophysics

Hairpins are formed by the single DNA strands of the fragile X triplet repeats: Structure and biological implications

(fragile X triplets/structure determination by two-dimensional NMR/hairpins of the C-rich strands/asymmetric expansion/*de novo* methylation)

XIAN CHEN*†, S. V. SANTHANA MARIAPPAN*, PAOLO CATASTI*†, ROBERT RATLIFF‡, ROBERT K. MOYZIS‡, ALI LAAYOUN§, STEVEN S. SMITH§, E. MORTON BRADBURY†¶, AND GOUTAM GUPTA*‖

*Theoretical Biology and Biophysics, T-10, MS-K710, †Life Sciences Division, LS-2, MS 880, ‡Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545; §Department of Cell and Tumor Biology, City of Hope National Medical Center, 1500 East Duarte Road, Duarte, CA 91010; and ¶Department of Biological Chemistry, School of Medicine, University of California, Davis, CA 95616

Communicated by Alfred G. Redfield, Brandeis University, Waltham, MA, December 22, 1994 (received for review August 9, 1994)

ABSTRACT **Inordinate expansion and hypermethylation of the fragile X DNA triplet repeat, $(GGC)_n$·$(GCC)_n$, are correlated with the ability of the individual G- and C-rich single strands to form hairpin structures. Two-dimensional NMR and gel electrophoresis studies show that both the G- and C-rich single strands form hairpins under physiological conditions. This propensity of hairpin formation is more pronounced for the C-rich strand than for the G-rich strand. This observation suggests that the C-rich strand is more likely to form hairpin or "slippage" structure and show asymmetric strand expansion during replication. NMR data also show that the hairpins formed by the C-rich strands fold in such a way that the cytosine at the CpG step of the stem is C·C paired. The presence of a C·C mismatch at the CpG site generates local flexibility, thereby providing analogs of the transition to the methyltransferase. In other words, the hairpins of the C-rich strand act as better substrates for the human methyltransferase than the Watson–Crick duplex or the G-rich strand. Therefore, hairpin formation could account for the specific methylation of the CpG island in the fragile X repeat that occurs during inactivation of the *FMR1* gene during the onset of the disease.**

Simple tandemly repeated DNA sequences are interspersed in both transcribed and nontranscribed regions of chromosomes (1–3). The hypothesis (4) that the unusual DNA structures adopted by these repeats principally determine their specific functions is gaining strength. We have previously described the unusual hairpin structures (5, 6) adopted by a variety of repetitive DNA sequences. Here, we show by NMR and gel electrophoresis that the individual strands from the fragile X triplet repeats, $(GGC)_n$·$(GCC)_n$, form intramolecular hairpins under physiological conditions. In these hairpins, the number of Watson–Crick G·C pairs is maximized in the stem through the formation of G·G or C·C mispairs flanked by G·C pairs (Fig. 1). As shown below, these hairpins provide structural basis for three major phenomena associated with the fragile X syndrome (3, 4): (*i*) the site-specific fragility, (*ii*) the amplification of the repeat (especially the preferential expansion of the C-rich strand), and (*iii*) the hypermethylation of the CpG island adjacent to the fragile X gene, *FMR1*.

MATERIALS AND METHODS

Gel Electrophoresis. Oligonucleotides were fully denatured by heating at 95°C for 2 min in 5 mM Tris/1 mM EDTA buffer, pH 7.5, containing 5 mM or 200 mM NaCl, followed by incubation at 55°C for 10 min and gradual cooling to room temperature. The samples were equilibrated at room temperature for 10 min and loaded on a preequilibrated gel [20% polyacrylamide in 0.6× TBE buffer (1× TBE buffer = 0.09 M Tris borate, pH 8.3/2 mM EDTA)] and then electrophoresed at 4°C at 75 V. The gel was stained with ethidium bromide (10 μg/ml).

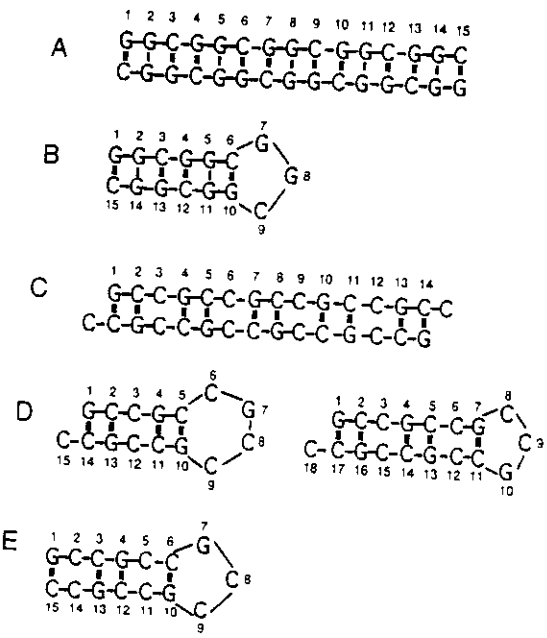


FIG. 1. Schematic representations of the self-assembled structure of $(GGC)_5$ duplex (*A*), $(GGC)_5$ hairpin (*B*), $(GCC)_5$ slipped duplex (*C*), $(GCC)_5$ and $(GCC)_6$ hairpins (*D*), and blunt $(GCC)_5$ hairpin (*E*). Note that in the hairpins of the G- and C-rich strands, the central mismatched G·G or C·C pair in the stem is surrounded by two Watson–Crick G·C pairs.

Methylation Assay. DNA methylation involving tritium incorporation was carried out, following the procedure of Smith *et al.* (6). A mixture containing 8 μM DNA and 8 μM tritiated *S*-adenosylmethionine was preincubated in the reaction buffer [50 mM Hepes, pH 7.0/50 mM NaCl/2 mM dithiothreitol/75 μM spermine/10% (vol/vol) glycerol] for 30 min at 37°C. The reaction was initiated with DNA methyltransferase and the reaction product was then recovered by precipitation with trichloroacetic acid (TCA). The rate of methylation was determined by measuring tritium incorporation into the TCA-insoluble DNA.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: NOE, nuclear Overhauser enhancement; NOESY, NOE spectroscopy; DQF-COSY, double-quantum filtered correlated spectroscopy.

¶To whom reprint requests should be addressed.

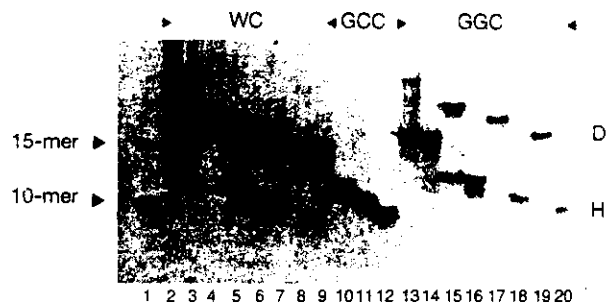


FIG. 2. Electrophoretic mobilities in nondenaturing polyacrylamide gel of single-stranded repeats $(GGC)_n$, $(GCC)_n$, and Watson–Crick (WC) heteroduplex in presence of NaCl at low and high concentrations (5 mM and 200 mM, respectively). Lane 1 shows the 15- and 10-mer DNA duplex markers. The positions of unimolecular hairpin conformation (i.e., hairpin H) and the self-complementary duplexes (D) with 15 and 10 Watson–Crick base pairs are also marked in the gel as 15- and 10-mer, respectively. For $(GGC)_n$ [lanes 14 (n = 11), 16 (n = 7), 18 (n = 6), and 20 (n = 5)], the hairpin is the predominant species in 5 mM NaCl. With increasing salt concentration (200 mM NaCl), the equilibrium shifts toward the $(GGC)_n$ duplex [lanes 13 (n = 11), 15 (n = 7), 17 (n = 6), and 19 (n = 5)]. The population of hairpin increases with increasing repeat number, n. As can be seen, the hairpin is still the dominant conformer for n = 11 even in 200 mM NaCl. Also note that the odd repeat number gives higher percentage of hairpin than the next even number $[(GGC)_5 > (GGC)_6]$. Lanes 10, 11, and 12 show the hairpin conformations of $(GCC)_n$ for n = 7, 6, and 5, respectively. Because $(GCC)_n$ remains in the hairpin conformation at low and high salt, the gel data are shown only for the low salt. Lanes 2, 4, 6, and 8 show the gel data for Watson–Crick duplexes, $(GGC)_n$·$(GCC)_n$, for n = 11, 7, 6, and 5, respectively, in 200 mM NaCl. Lanes 3, 5, 7, and 9 show the gel data for Watson–Crick duplexes, $(GGC)_n$·$(GCC)_n$, for n = 11, 7, 6, and 5, respectively, in 5 mM NaCl. Note that lanes 2–9 tend to show a small population of $(GCC)_n$ hairpins.

Gel assay for cytosine 5-methylation of the G-rich or C-rich strand and Watson–Crick duplex was carried out with the repeat number n = 7. For the bacterial methyltransferase *Sss* I, DNA methylation reactions were performed according to the instructions of the supplier, New England Biolabs. The DNA samples (8 μM) were preequilibrated and annealed in 1× NEB (50 mM NaCl/10 mM MgCl2/10 mM Tris·HCl, pH 7.5/1 mM dithiothreitol). Four units of *Sss* I and 160 μM *S*-adenosylmethionine were added and the mixture was incubated at 37°C for 2 hr. The same treatment was used for DNA substrates (8 μM) of the human methyltransferase in 4× Hepes buffer (50 mM NaCl/10 mM dithiothreitol/400 mM Hepes, pH 7.5). Equimolar concentrations of the complementary strands of G- or C-rich strands were added to the solution after the reaction was complete. This was followed by treatment with protease K to inactivate and digest the methyltransferase. The solution then was heated to 95°C and slowly cooled to anneal the Watson–Crick duplex. The proteins were removed by extraction with phenol and the DNA duplex was recovered by precipitation with ethanol. To determine the relative methylation on the target strands (G- or C-rich strands and Watson–Crick duplexes), the DNA samples were resuspended in a 20-μl solution containing 1× NEB and the methyl-sensitive restriction enzyme *Bso*FI (8 units) and were incubated at 55°C for 2 hr. The restriction enzyme *Bso*FI cuts double-stranded DNA with a recognition site of 5′-GCNGC-3′, but not the methylated form of the site. The digested products were analyzed by gel electrophoresis.

RESULTS

Identification of the Hairpins 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 (Fig. 1). The duplex and the stem of the corresponding hairpin for G- and C-rich strands involve G·G and C·C pairs, respectively. Note that the hairpin of a given sequence [i.e., $(GGC)_n$ or $(GCC)_n$] should have half the length of but approximately the same cross-section as the duplex of length n (Fig. 1). Therefore, the duplex is expected to have slower gel mobility than the corresponding hairpin. Fig. 2 shows the electrophoretic mobilities of G- and C-rich single strands along with the marker duplex sequences of defined lengths in a nondenaturing polyacrylamide gel. The gel data show the presence of two structural forms of the G-rich strands—hairpin and duplex. For a given sequence, the relative population of the hairpin or the duplex depends on the repeat number (n) and the DNA and salt concentrations. For low n, the hairpin of the G-rich strands is the predominant conformation at low salt (5 mM NaCl) and the duplex is the major conformer at high salt (200 mM NaCl). With increasing n, the equilibrium gradually shifts toward the hairpin structure within the salt concentration range of 5–200 mM NaCl. For n = 11, the hairpin is the predominant conformation at all salt concentrations. The gel data of C-rich strands show only hairpins at both 5 and 200 mM NaCl. Thus the C-rich strands more readily form hairpins than the G-rich strands. The thermodynamic preference for the individual G- and C-rich strands to form hairpins as a function of n is determined by how easily the fragile X duplex, $(GGC)_n$·$(GCC)_n$, can undergo transitions to two unimolecular hairpins (Fig. 1). This tendency of the G-rich strand not to form a hairpin for low repeat numbers at physiological salt concentrations implies that the C-rich strand does not dissociate from the Watson–Crick paired duplex to form a hairpin at low n. This is primarily because the formation of hairpins by the individual G- and

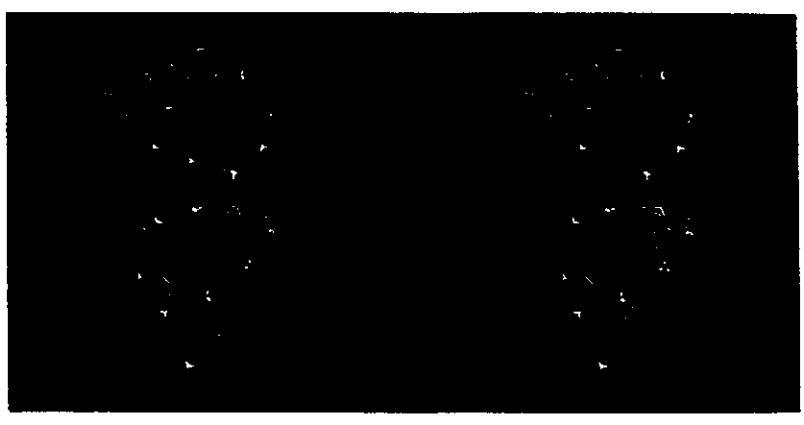


FIG. 3. Stereo pair of a representative energy-minimized structure of the $(GCC)_5$ hairpin consistent with the NMR data. Analysis of NOESY data (200 and 500 ms of mixing time) of $(GCC)_5$ with the aid of full-matrix NOESY simulation resulted in a set of 135 distance constraints.

FIG. 4. A stereo pair of the hairpin model of $(GGC)_9$ in which the stem satisfies the NMR constraints of the $[(GGC)_4]_2$ duplex. Analysis of NOESY data (200 and 500 ms mixing time) of $[(GGC)_4]_2$ with the aid of full-matrix NOESY simulation resulted in a set of 163 average interproton distances for the 12-base-pair-long duplex.

C-rich strands is energetically less favorable than formation of the Watson–Crick duplex $(GGC)_n{\cdot}(GCC)_n$. The biological occurrence of repeats which are all longer than $n = 11$ relieves this constraint. Nonetheless, the observation that for appropriate n the Watson–Crick duplex $(GGC)_n{\cdot}(GCC)_n$ can undergo a transition to the hairpins formed by the individual single strands is of great importance, since these hairpins provide molecular descriptions of the "slippage" structures proposed to be replication intermediates required for strand expansion (see Fig. 7).

Structural Characterization of the Hairpins by NMR. Knowledge of the precise stereochemistry of the hairpins, including the exact base-base interactions in the stem and loop, is important in understanding their biological role. Extensive one- and two-dimensional (1D/2D) NMR experiments were carried out to obtain the stereochemical details of the hairpin structures formed by the individual G- and C-rich strands of the fragile X repeat. The nature of base-pairing patterns was identified by monitoring (*i*) temperature-dependent imino proton profiles and (*ii*) nuclear Overhauser enhancements (NOEs) from the imino protons in 1D/2D NOE experiments in ^1H_2O:^2H_2O (9:1, vol/vol). Glycosidic torsions and sugar puckers were deduced by interpreting the NOE intensities and the coupling constants derived from double-quantum filtered correlated spectroscopy (DQF-COSY). A set of average interproton distances for pairwise interactions was obtained by performing full-relaxation-matrix simulation with the NOE spectroscopy (NOESY) intensities at 500 and 200 ms of mixing time in conjunction with the linked-atom least squares refinement (7). The starting structure was constructed from these average interproton distances. The interproton distance constraints as well as the base-pairing constraints that were consistent with the 2D NMR data were included as harmonic potentials in restrained molecular dynamics and energy minimization (5, 7). Detailed NMR studies will be published elsewhere.

C-Rich Strand. Fig. 3 shows the stereo pair of a representative energy-minimized structure of the $(GCC)_5$ hairpin. All nucleotides in the hairpin adopt C2′-*endo*, *anti* conformation as evidenced by the NOESY and DQF-COSY data. In this structure, the 5′-C in the CpG step of the stem is C·C paired and not G·C paired. The C·C pair most likely involves a single hydrogen bond between N4-H and N3. This leads to two possibilities in which either of the two cytosines can act as a proton donor (participation of N4-H) or an acceptor (participation of N3). In a 400-ps unrestrained molecular dynamics simulation, we observed that the C3·C12 pair in the $(GCC)_5$ hairpin (Fig. 1) can undergo local periodic sliding motions between the two degenerate hydrogen-bonding states without violating local or distant NOE constraints. Such a sliding motion makes cytosines in the C·C pairs intrinsically more flexible than cytosines in the G·C pairs. Previously Gueron *et al.* (8) showed that mismatch pairs are more susceptible to base-pair open-closure dynamics than are the canonical Watson–Crick 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.

G-Rich Strands. The structural details of the hairpins formed by the G-rich strand are also important for understanding the expansion of this strand during replication. However, complete structure determination of these hairpins by NMR was not possible, since we found that the G-rich strand exists predominantly in the duplex state for a range of DNA concentrations (0.5–2 mM in strand) for $n \geq 11$. Since the stem of the hairpin and the duplex are conformationally similar, it is noteworthy to mention some of the salient features of the duplex: (*i*) $(GGC)_4$, $(GGC)_5$, and $(GGC)_6$ form a six-base-pair-long structural repeat

$$G1^{anti}\text{-}G2^{anti}\text{-}C3^{anti}\text{-}G4^{anti}\text{-}G5^{syn}\text{-}C6^{anti}\text{-}G7^{anti}$$
$$C1^{anti}\text{-}G2^{syn}\text{-}G3^{anti}\text{-}C4^{anti}\text{-}G5^{anti}\text{-}G6^{anti}\text{-}C7^{anti}.$$

(*ii*) Two symmetric O6···H-N1 hydrogen bonds are present in the $G^{anti}{\cdot}G^{syn}$ pairing. (*iii*) The cytosines in the two CpG sequences of this repeat are distinguished by the presence of $G^{anti}{\cdot}G^{syn}$ on the 5′ end of C. The (5′-3′) sequential H2″(C)–H8(G) NOE at the CpG fragment is weak when G^{anti} is the 5′ neighbor of C (i.e., C3pG4) and the same NOE is totally absent when G^{syn} is the 5′ neighbor of C (i.e., in C6pG7). This (5′-3′) sequential H1′–H8 and H6–H8 NOEs are also absent at the two CpG segments of the repeat. This type of intrastrand NOE discontinuity suggests a local structural discontinuity at the CpG step due to the formation of adjacent $G^{anti}{\cdot}G^{syn}$ pairs. We used a molecular modeling approach to construct the hairpin

Table 1. Rates of methylation of the cytosines in the CpG step of triplet repeats

Methyl-transferase	Repeat (*n*)	Methylation rate, fmol/min		
		GCC	GGC	WC
Sss I	5	—	—	—
	6	—	2.6	2.0
	7	—	24	23
	11	7.1	30	21
Human	5	3.9	1.1	1.6
	6	12	1.7	4.9
	7	23	9.5	14
	11	290	27	54

All values are averages of four measurements. A dash indicates too low to measure. Hairpins contain half the potential methylation sites present in the Watson–Crick (WC) duplex.

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 energetically stable loop segments. Fig. 4 shows the stereo pair of 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.

In summary, the NMR and gel data show that both $(GCC)_n$ and $(GGC)_n$ strands can form hairpin structures under physiological conditions. Once formed, these monomeric hairpins may be susceptible to nuclease attack. Nucleases with apparent specificity for single-stranded loop regions have been reported in yeast (9). Hence, this tandemly repeated DNA sequence may become fragile simply due to its propensity to form hairpins.

Site-Specific Methylation of the Hairpins by Human and Bacterial Methyltransferases. Apart from being a site prone to cleavage, the fragile X repeat is also associated with the *FMR1* gene (3). In fragile X syndrome, the cytosines at the CpG sites become hypermethylated. This methylation may play a role in inactivating the *FMR1* gene (10). We used the bacterial methyltransferase (*Sss* I) and the methyltransferase from human placenta to study the methylation of the G-rich and C-rich strands under conditions in which unimolecular hairpins were the primary species.

The bacterial methyltransferase, *Sss* I, appears to have a strict requirement for a Watson–Crick paired CpG as found in the Watson–Crick duplex and the G-rich hairpin (structure **1** in Fig. 6). These two forms are methylated by *Sss* I at similar rates (Table 1). The substrate efficiency of the G-rich strand with repeat number 11 is 1.5 times higher than that of the Watson–Crick duplex. However, in the C-rich strand the CpG dinucleotides are disrupted by a C·C mispair. The substrate efficiency of the C-rich hairpins for *Sss* I is 1/4 that of the Watson–Crick duplex (Table 1). Fig. 5*A* shows (in a gel assay) the protection against *Bso*FI subsequent to methylation by *Sss* I. The restriction enzyme *Bso*FI has the duplex recognition site GCNGC, where N can be any nucleotide and the cytosines are unmethylated. Upon methylation of the cytosines, the same sequence becomes resistant to *Bso*FI. Note that the G-rich strand shows the highest protection against *Bso*FI (Fig. 5*A*) because the recognition site in this strand, GCGGC, gets methylated at the cytosines by *Sss* I.

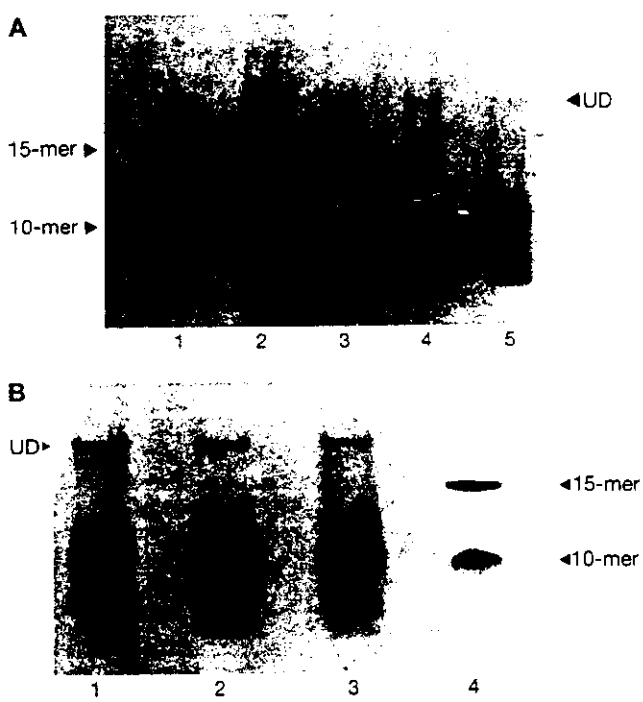


FIG. 5. Gel assay for detecting methylation in the CpG step of the G-rich and C-rich single strands and the Watson–Crick heteroduplex. (*A*) DNA methylation was catalyzed by bacterial methyltransferase *Sss* I. Lanes: 1, DNA markers, 10-mer and 15-mer duplex; 2, *Bso*FI digest of methylated G-rich strands; 3, *Bso*FI digest of methylated C-rich strands; 4, *Bso*FI digest of methylated Watson–Crick duplexes; 5, *Bso*FI digest of unmethylated Watson–Crick duplexes (restriction control). UD, undigested. (*B*) DNA methylation was catalyzed by human methyltransferase. Lanes: 1, DNA markers; 2, *Bso*FI digest of methylated C-rich strands in presence of the unmethylated complementary G-rich strand; 3, *Bso*FI digest of G-rich strands in presence of the unmethylated complementary C-rich strand; 4, *Bso*FI digest of methylated Watson–Crick duplexes.

In contrast, as shown in Table 1, compared with the Watson–Crick duplex the enzymatic activity of the human methyltransferase is 6-fold stimulated by the presence of the C·C mispair at the CpG site (11, 12); see structure **2** in Fig. 6. As shown in Fig. 5*B*, a weak protection against *Bso*FI cleavage is observed in the C-rich strand subsequent to methylation by the human methyltransferase even when tritium incorporation is quite significant. This is because the human enzyme methylates the C-rich strand in such a way that the *Bso*FI recognition site is altered at the nonessential nucleotide (i.e., GCmCGC).

Structural Basis of CpG Methylation Inside the Fragile X Repeat. Bacterial and human methyltransferases operate by the same mechanism (13, 14). For each enzyme, initial sequence-specific recognition is followed by nucleophilic attack at C6 to form a nonplanar dihydrocytosine intermediate which cannot be accommodated by the structure of a B-DNA helix (11). In the case of the bacterial enzyme *Hha* I, the crystal structure of the intermediate (15) is known. This crystal structure suggests that the enzymes have evolved to accommodate the nonplanar intermediate by rotating it to an extrahelical position within the active site of the enzyme. Therefore,

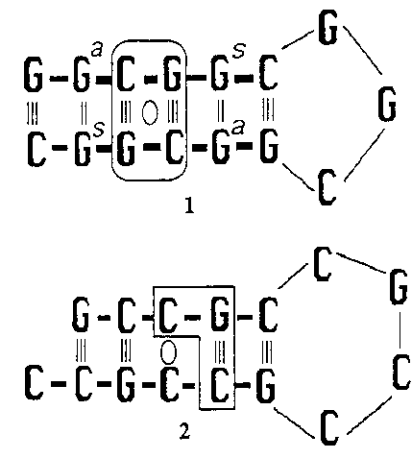


FIG. 6. The structural features of the recognition elements of bacterial and human methyltransferases. The hairpin of the G-rich strand (structure **1**) has the recognition sequence

$$5'\text{-}G^{anti}\text{-}C^{anti}\text{-}G^{anti}\text{-}G^{syn}\text{-}3'$$
$$3'\text{-}G^{syn}\text{-}G^{anti}\text{-}C^{anti}\text{-}G^{anti}\text{-}5'$$

for the bacterial enzyme *Sss* I (two potential methylation sites in the sequence are shown in bold). Note that the Watson–Crick paired CpG site is flanked by G^{anti}·G^{syn} mismatches which introduce structural discontinuity at the CpG step and flexibility at the CpG link. The recognition element for *Sss* I is boxed. The C-rich hairpin (structure **2**) provides the recognition sequence

$$5'\text{-}C^{anti}\text{-}C^{anti}\text{-}G^{anti}\text{-}3'$$
$$3'\text{-}G^{syn}\text{-}C^{anti}\text{-}C^{anti}\text{-}5'$$

for the human enzyme. The recognition element for the human enzyme is boxed. Note that the cytosines in the CpG step are involved in C·C pairs which offer a great deal of flexibility to the cytosines in the CpG step. The twofold symmetry in the recognition sequences of the bacterial and human enzymes is marked with an ellipse.

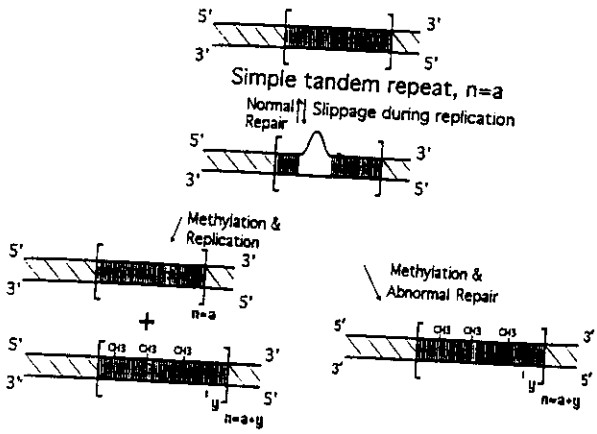


FIG. 7. Slippage during replication is proposed. The slippage is due to the formation of the hairpins by the G- and C-rich strands shown respectively as structures 1 and 2 (Fig. 6). The amount of expansion which can occur due to abnormal repair of a second round of replication depends on the size of the hairpin—i.e., the longer the size of hairpin (or the slippage), the longer the size (y) of expansion. From our studies, it appears that the C-rich strand of the fragile X repeat is most likely to get methylated first, and then in the next round of replication the complementary G-rich strand will also be methylated by the mechanism of maintenance or methyl-directed methylation (6, 16, 17).

any structure that satisfies the sequence recognition requirement and also possesses sufficient flexibility at the CpG site to allow the C to be rotated out would be a better substrate for the methyltransferase than the Watson–Crick duplex. Note that the NMR-derived structure of the C-rich strand hairpin (Figs. 3 and 6) satisfies this requirement because of the presence of the flexible C·C pair at the CpG that allows considerable freedom for the C to flip in and out of the structure. Similarly, in the stem of the hairpins of G-rich strands (Figs. 4 and 6), the presence of G^{syn}·G^{anti} pair adjacent to the CpG site causes local flexibility such that the C can flip in and out of the structure with relative ease.

DISCUSSION

On the basis of the structural and enzymological data described above, we propose a simple mechanism to explain the molecular events associated with fragile X syndrome. Fig. 7 describes a slippage model suggested by the strong tendency of the triplet repeats to form hairpins. In this scheme, the G- and the C-rich strands independently form hairpin structures. During replication, the growing chain can easily slip or slip and slide to form a hairpin which makes a part of the already replicated template again available for replication. This would result in expansion of the repeat in the next replication cycle. Our studies also predict that the C-rich strand is more likely to form slippage structure and get preferentially expanded during *in vitro* replication. Ordinarily repair through loop excision would preserve the sites intact although this might promote chromosomal breaks. Unrepaired loops would be substrates for *de novo* methylation (Fig. 7). Erroneous repair against the longer strand in the heteroduplex could also result in repeat expansion. The tendency for expansion is dependent on repeat number, since the tendency for hairpin formation is also dependent on repeat number. This tendency has been referred to as a dynamic mutation (18). Stability of repeats shorter than $n = 50$ and instability of repeats longer than $n = 50$ may reflect the tendency of the repair systems to become saturated when loop formation in the region occurs at high frequency. Finally, our studies suggest that methylation of the C-rich strands is likely to be a consequence of hairpin formation. However, we cannot rule out the possibility that methylation of loops might cause site-specific fragility (10). Moreover, it should be added that the maintenance of methylation as shown in Fig. 7 is associated with late replication of the fragile X repeat (16). This might lead to delayed chromosome condensation or gap formation resulting in site-specific fragility.

We thank Dr. Angel E. Garcia for his help during various stages of this work. We thank Drs. Neville Kallenbach of New York University and Christian Burks of Los Alamos National Laboratory for suggestions and corrections on the manuscript. We are grateful to Dr. Cliff Unkefer for giving us access to the 500-MHz Bruker-AMX NMR spectrometer at Chemical Science and Technology 4 (CST-4). This work was supported by Los Alamos National Laboratory Grant XL-77 and the Human Genome Project of the Office of Health and Environmental Research of the Department of Energy and by Grant 0388 from the Smokeless Tobacco Research Council, Inc.

1. Moyzis, R. K., Torney, D. C., Meyne, J., Buckingham, J. M., Wu, J. R., Burks, C., Sirotkin, K. M. & Goad W. B. (1989) *Genomics* **4**, 273–289.
2. Grady, D. L., Ratliff, R. L., Robinson, D. L., McCanlies, E. C., Meyne, J. & Moyzis, R. K. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 1695–1699.
3. Oberié, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boué, J., Bertheas, M. F. & Mandel, J. L. (1991) *Science* **252**, 1097–1102.
4. Sinden, R. R. & Wells, R. D. (1992) *Curr. Opin. Biotechnol.* **3**, 612–622.
5. Catasti, P., Gupta, G., Garcia, A. E., Ratliff, R., Hong, L., Yau, P., Moyzis, R. K. & Bradbury, E. M. (1993) *Biochemistry* **33**, 3819–3830.
6. Smith, S. S., Laayoun, A., Lingerman, R. G., Baker, D. J. & Riley, J. (1994) *J. Mol. Biol.* **243**, 143–151.
7. Gupta, G., Garcia, A. E. & Hiriyanna, K. T. (1993) *Biochemistry* **32**, 948–960.
8. Gueron, M., Kochoyan, M. & Leroy, J. (1987) *Nature (London)* **328**, 89–92.
9. Symington, L. S. & Kolodner, R. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 7247–7251.
10. Nancacarrow, J., Kramer, E., Holman, K., Eyre, H., Doggett, N. A., LaPeaslier, D., Callen, D. F., Sutherland, G. R. & Richards, R. I. (1994) *Science* **264**, 1938–1941.
11. Smith, S. S., Kan, J. L. C., Baker, D. J., Kaplan, B. E. & Dembek, P. (1991) *J. Mol. Biol.* **217**, 39–51.
12. Smith, S. S., Hard, T. A. & Baker, D. J. (1987) *Nucleic Acids Res.* **15**, 6899–6916.
13. Osterman, D. G., Depillis, G. D., Wu, J. C., Matsuda, A. & Santi, D. V. (1988) *Biochemistry* **27**, 5204–5210.
14. Smith, S. S., Kaplan, B. E., Sowers, L. C. & Newman, E. M. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4744–4748.
15. Klimasanskas, S., Kumar, S., Roberts, R. J. & Cheng, X. (1994) *Cell* **76**, 357–369.
16. Laird, C. D., Jaffe, E., Karpen, G., Lamb, M. & Nelson, R. (1987) *Trends Genet.* **3**, 274–281.
17. Riggs, A. D. (1975) *Cytogenet. Cell Genet.* **14**, 9–25.
18. Richards, R. I. & Sutherland, G. R. (1994) *Nat. Genet.* **8**, 114–116.

Unusual Structures of the Tandem Repetitive DNA Sequences Located at Human Centromeres†

Paolo Catasti,‡ Goutam Gupta,§ Angel E. Garcia,§ Robert Ratliff,∥ Lin Hong,‡ Peter Yau,‡ Robert K. Moyzis,∥ and E. Morton Bradbury*,⊥

Theoretical Biology and Biophysics Group, Division T-10, M/S K710, Life Sciences Division, M/S M881, P.O. Box 1663, and Center for Human Genome Studies, 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

Received April 19, 1993; Revised Manuscript Received December 13, 1993•

ABSTRACT: The presence of the highly conserved repetitive DNA sequence d(AATGG)$_n$·d(CCATT)$_n$ in human centromeres argues for a special role for this sequence in recognition, most probably through the formation of an unusual structure during mitosis. Quantitative one- and two-dimensional nuclear magnetic resonance (1D/2D NMR) spectroscopic studies reveal that the Watson–Crick duplex d(AATGG)$_n$·d(CCATT)$_n$ adopts the usual B-DNA conformation as illustrated by taking d(AATGG)$_3$·d(CCATT)$_3$ as an example, whereas the d(CCATT)$_n$ strand is essentially a random coil. In contrast, the d(AATGG)$_n$ strand adopts an unusual stem–loop motif for repeat lengths n = 2, 3, 4, and 6. In addition to normal Watson–Crick A·T pairs, the stem–loop structures are stabilized by mismatched A·G and G·G pairs in the stem and G-G-A stacking in the loop. Stem–loop structures of d(AATGG)$_n$ are independently verified by gel electrophoresis and nuclease digestion studies and were also previously shown to be as stable as the corresponding Watson–Crick duplex d(AATGG)$_n$·d(CCATT)$_n$ [Grady et al. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 1695–1699]. Therefore, the sequence d(AATGG)$_n$ can, indeed, nucleate a stem–loop structure at little free energy cost, and if, during mitosis, it is located on the chromosome surface, it can provide specific recognition sites for kinetochore function.

Genomes of all higher organisms contain subsets of DNA sequences in multiple copies. Whereas many classes of these repetitive DNA sequences are interspersed throughout the genome (Moyzis et al., 1989), other classes are localized, e.g., in extremely long tracts at the telomeric or centromeric regions of the chromosome (Zakian et al., 1989). Tracts of d(TTAGGG)$_n$ [where n refers to the number of repeats] sequences are found at the termini of all human chromosomes and represent the functional human telomere (Moyzis et al., 1988). The ability of telomeric DNA sequences to adopt unusual DNA structures in which the G-rich tracts form a tetraplex stem–loop structure (Kang et al., 1992) is thought to relate to their biological function (Zakian et al., 1989). Other unusual DNA structures of possible biological function include a triple helix through the formation of a stem–loop motif in the purine-rich regions of the eukaryotic gene promoters (Durland et al., 1991) and Z-DNA, which may be involved in transcription and supercoiling (Rich et al., 1984).

The highly conserved repetitive DNA sequences d(AATGG)$_n$ present in human centromeres were shown to have unusually high thermal stabilities, particularly for a sequence lacking perfect Watson–Crick duplex complementarity (Grady et al., 1992). UV melting studies on the synthetic analogs of centromeric DNA repeats suggested that the high stability of this sequence results from the utilization of A·G and possible G·G base pairing (Grady et al., 1992) in an unusual DNA structure. One- and two-dimensional (1D/2D) NMR spectroscopy has been used to determine the three-dimensional structure of d(AATGG)$_n$ and the Watson–Crick duplex d(AATGG)$_n$·d(CCATT)$_n$. The duplex d(AATGG)$_n$·d(CCATT)$_n$ was shown to adopt the usual B-DNA double helix with Watson–Crick G·C and A·T pairs. However, whereas d(CCATT)$_n$ was found to exist in the random-coil state (Grady et al., 1992), the complementary d(AATGG)$_n$ strands displayed stable ordered structures for various repeat lengths n = 2, 3, 4, and 6. Stabilities of d(AATGG)$_n$ structures varied with the repeat length (n) and the ionic conditions. Previously, we have reported structural studies using 1D/2D NMR spectroscopy (Gupta et al., 1987; Garcia et al., 1990) for other DNA sequences and shown that a monomeric hairpin is favored at low DNA concentrations and low salt concentrations, whereas at higher DNA concentrations and higher salt concentrations the duplex is typically favored. A stem–loop structure is stabilized by specific base-pairing schemes in the stem and specific base stacking in the loop. The accurate determination of the stem–loop structure of repetitive DNA sequences requires the characterization of the base-pairing scheme in the stem region, the base stacking arrangement in the loop region, and, most importantly, the conformations of the individual nucleotides. All these structural features of a stem–loop motif can be determined in solution by one- and two-dimensional nuclear magnetic resonance (1D/2D NMR) experiments. Here we report the determination of the solution structures of the highly conserved human centromeric DNA repeats d(AATGG)$_n$ of lengths n = 2, 3, 4, and 6.

In the text below, the single-stranded loop-folded structures (the hairpins or the stem–loop motifs) are denoted as d(AATGG)$_n$, whereas the end-stacked hairpin dimers (also called the stem–loop motifs) are denoted as [d(AATGG)$_n$]$_2$.

† Support for this work from the U.S. Department of Energy to the Human Genome Center and the Life Sciences and Theoretical Divisions of LANL and to E.M.B. (Grant DE-FG03-88ER60673) is greatly appreciated.

* Corresponding author (telephone, 505-667-2690; FAX, 505-665-3024).

‡ Department of Biological Chemistry, University of California at Davis.

§ Theoretical Biology and Biophysics Group, Los Alamos National Laboratory.

∥ Center for Human Genome Studies, Los Alamos National Laboratory.

⊥ Life Sciences Division, Los Alamos National Laboratory.

• Abstract published in *Advance ACS Abstracts*, February 15, 1994.

0006-2960/94/0433-3819$04.50/0

MATERIALS AND METHODS

NMR Experiments. NMR spectra were recorded on a GE-Omega 500 spectrometer. 1D NMR experiments in H_2O were conducted using the 11-echo pulse sequence of Sklenar and Bax (1987). The acquisition parameters for phase-sensitive 2D NOESY/COSY experiments were as follows: sweep width = 5000 Hz, complex data points in t_2 = 2048, complex FIDs in t_1 = 256, number of transients = 32, and relaxation delay = 1.5 s. The mixing times, τ_m, for NOESY experiments were 100 and 250 ms, respectively. The data in t_1 was zero-filled to 1024 before Fourier transformation of the 2048 × 1024 data matrix. The data were not symmetrized.

Sequential Assignment. First, the sequential assignment of the spin system H8/H6, H1′, H2′, H2″ was obtained from the NOESY cross sections H8/H6 vs H2′, H2″ at various mixing times. Second, the spin system H1′, H2′, H2″, H3′, H4′ was sequentially assigned by monitoring the intranucleotide interactions (NOE or *J*-coupling) involving H1′–H2′, H1′–H2″, H2′–H3′, H2″–H3′, H3′–H4′ in the NOESY/COSY cross sections.

Structural Analyses. The following steps were adopted to interpret the 1D/2D NMR data. First, the nature of H-bonding in the structure was characterized by monitoring the temperature dependence and the solvent-exchange properties of the exchangeable imino signals and by performing 1D NOE experiments. Second, a set of interproton distances (i.e., average values and associated dispersions) was extracted for various pairwise interactions by performing full-matrix NOESY simulation and associated *R*-factor tests by comparing the corresponding calculated and observed NOESY intensities (Gupta et al., 1988). The sugar puckers of different residues were estimated by monitoring the corresponding *J*-coupling parameters of the H1′–H2′, H1′–H2″, H2′–H3′, H2″–H3′, etc., interactions in the corresponding phase-sensitive COSY cross sections. Third, these interproton distances were used as structural constraints for constant high-temperature (400 K) 200-ps molecular dynamics (MD) simulations after 3 ps of temperature equilibration. The starting configuration for MD simulation is an energy-minimized structure that satisfies the NOE distance constraints and the observed base-pairing scheme. Fourth, 200 snapshots (one after every 1 ps) were extracted from the MD trajectory, and constrained energy minimization on each snapshot was used to map local minima on the sampled energy surface; this is the *temperature quenching step*. Fifth, 200 energy-minimized structures were assigned to different disjoint clusters such that conformationally similar hairpins belong to the same cluster while conformationally distinct stem–loop structures belong to different clusters (Gupta et al., 1993). Finally, full-matrix NOESY simulation and the associated *R*-factor tests were performed on the representative structures of different clusters to check the agreement with the NOESY data (Gupta et al., 1988). Steps 3–5 are collectively referred to as "high-temperature MD simulation followed by rapid temperature quenching, HTMD/RTQ" (Stillinger & Weber, 1983). During HTMD/RTQ all NOE-derived distance constraints were imposed by using appropriate constraint energy functions. Therefore, all the final 200 energy-minimized structures are in agreement with the NMR data. In order to distinguish local and global rearrangements of atoms or groups among different structures, we defined a hierarchy of structures by progressively dividing structures among different clusters (Gupta et al., 1993). The mean-square distance between all pairs of structures is used as a discriminating parameter for this purpose.

MD and energy minimization were performed using the all-atom force field of Weiner et al. (1986) in AMBER 3.0. All calculations were done *in vacuo* with a constant dielectric coefficient of 78.4 (Garcia & Soumpasis, 1989; Garcia et al., 1990) and without any nonbonding cutoff. High-temperature (400 K) simulations were performed with a set of strong H-bonding constraints (k = 100 kcal mol^{-1} Å^{-2} for the A·T, A·G, and G·G pairs in the structure). Strong constraints are also imposed for interproton distances ≤2.5 Å. The A·G base-pairing constraints in the G-G-A loop segments were not imposed in the calculations.

Gel Electrophoresis and Nuclease Digestion. Electrophoretic patterns of $d(AATGG)_{2,3,4,6}$ were monitored in a nondenaturing gel: 12% polyacrylamide and 0.5 × TBE buffer. Oligonucleotides were labeled by ^{32}P. Samples were heated to 80 °C and then gradually cooled down to 4 °C. Gels were run in a cold room with an ambient temperature of 4 °C. To keep the gel cool, the gel plates were kept in direct contact with the cold circulating buffer. About 12 μL of the sample containing 0.1 mg of DNA was loaded in each lane. No mobility difference was found when the DNA concentration was changed.

The mung bean nuclease (a probe for single-stranded regions in DNA) was used to map the single-stranded regions expected in the stem–loop structures. Oligonucleotides were labeled by ^{32}P before digestion. Reaction conditions were 30 mM sodium acetate, 50 mM NaCl, and 15 μM $ZnCl_2$, pH 5.0. Reactions containing about 20 ng of oligonucleotides with 11 units of mung bean nuclease were run at 0 °C for 2.5, 5.0, and 10.0 min. The reaction was stopped at different times by addition of 50 mM EDTA. Denaturation DNA gels were used: 15% polyacrylamide (20:1) and 7 M urea in Tris buffer.

RESULTS

The structural details are based upon extensive 1D/2D NMR data that include (i) identification of A·T, A·G, and G·G pairs by 1D/2D NOE experiments involving H-bonding (exchangeable) protons (Gupta et al., 1988), (ii) determination of sugar puckers by phase-sensitive COSY and determination of interproton distances by full-matrix NOESY simulations and the associated *R*-factor calculations with respect to the 2D NOESY experiments at two mixing times (Gupta et al., 1988), and finally (iii) derivation of quantitative structures in agreement with the NMR data by HTMD/RTQ (Gupta et al., 1993).

The Hairpin Monomer of $d(AATGG)_2$ and the Stem–Loop Motif of $[d(AATGG)_2]_2$. The 1D NMR spectra of $d(AATGG)_2$ at different solution conditions are shown in Figure 1A. They show the imino, amino, and base proton signals. For low $d(AATGG)_2$ concentration (0.4 mM in strand) and low salt (25 mM NaCl, pH 7), spectrum 1 shows the imino proton signals characteristic of a monomeric hairpin structure. Note that (i) the two imino signals within 13.5–13.2 ppm correspond to two A·T pairs in the stem as confirmed by the irradiation of the imino signals (Figure 1A, spectrum 3), (ii) the broad signal at 11 ppm is due to the imino signals of G's in the loop, (iii) the imino signal from the terminal A·G pair is not clearly visible due to solvent exposure and dynamically rapid open ↔ close exchange, and (iv) the terminal unpaired G shows no imino signal. NOESY experiments (mixing time, τ_m = 250 ms) for this hairpin revealed that all the nucleotides in this hairpin belonged to the C2′-*endo*, *anti* conformation (data not shown) and weaker NOEs were found for the proton pairs in the loop compared to the proton pairs in the stem, as previously reported for other hairpin structures (Gupta et al., 1987; Blommers et al., 1989; Williamson et al., 1989). At the

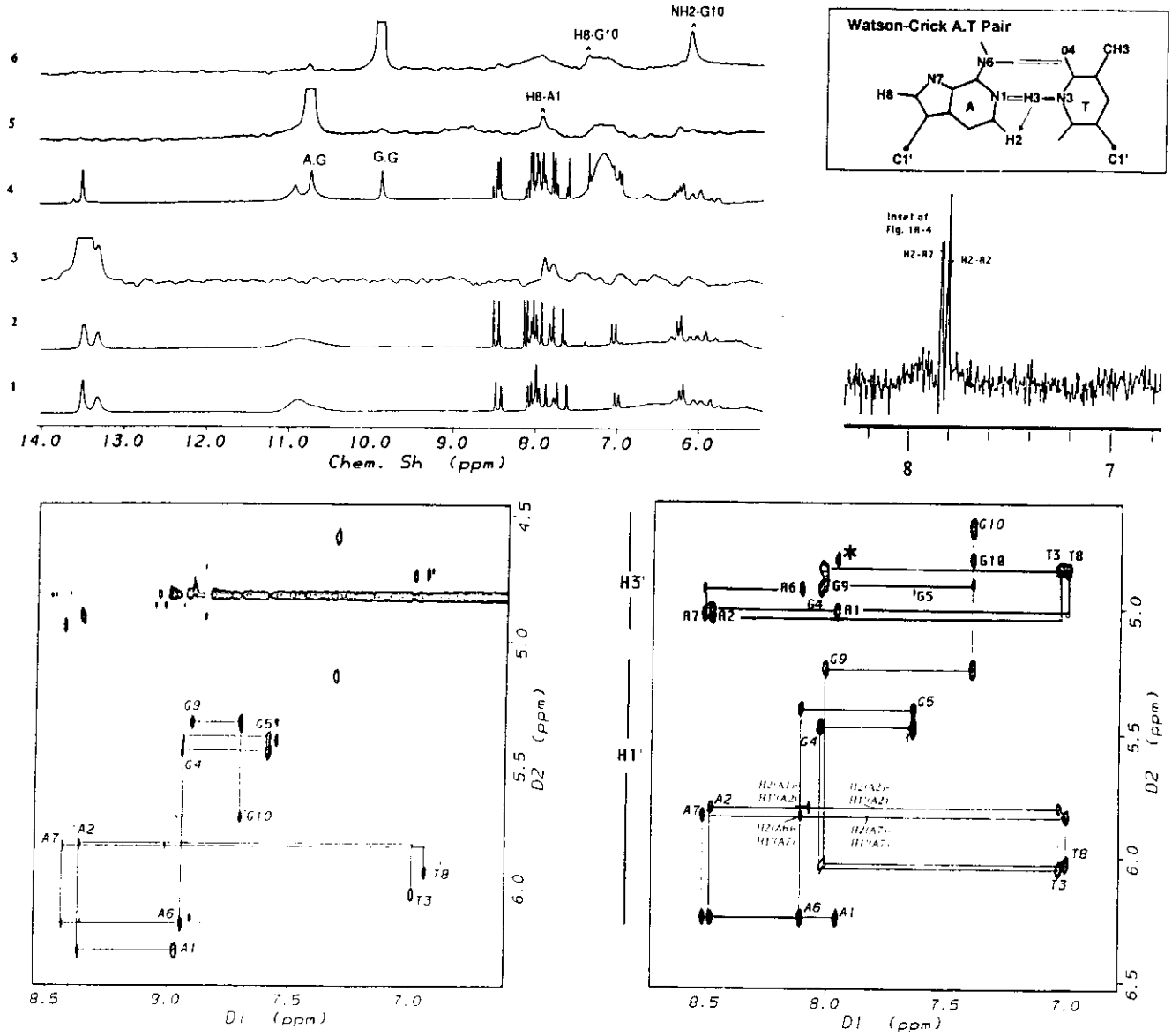


FIGURE 1: (A, top left) 500-MHz 1D proton NMR spectra of $d(AATGG)_2$ showing the imino, amino, and base protons under different conditions of DNA and salt concentrations. The spectra were recorded in a H_2O–D_2O (9:1) mixture using the 1–1 echo pulse sequence of Sklenar and Bax (1987). Spectrum 1 (0.4 mM in strand, 25 mM NaCl, pH 7, temperature 10 °C) shows the presence of two imino signals (13.5–13.2 ppm) due to two A·T pairs and a broad signal at 11 ppm due to the imino signals of G's in the loop. Spectrum 2 (2.4 mM in strand, 25 mM NaCl, pH 7, temperature 16 °C) shows the presence of two imino signals (13.5–13.2 ppm) due to two A·T pairs and a broad signal at 11 ppm due to the imino signals of G's in the loop. The similarity of spectra 1 and 2 indicates that $(AATGG)_2$ can form a monomeric hairpin over a wide range of DNA concentrations. Spectrum 3 shows a 1D NOE at H2's of A upon irradiation of the low-field imino signal due to one of the A·T pairs (presaturation time = 150 ms). Spectrum 4 (2.4 mM in strand, 1 M NaCl, pH 7, temperature 3 °C) shows the presence of two imino signals (13.5–13.2 ppm) due to two A·T pairs, a broad signal at 11 ppm due to the imino signals of G's in the loop, a signal at 10.8 ppm due to an A·G pair, and a signal at 9.9 ppm due to a G·G pair. Also shown as an inset (top right) are two H2's of A2 and A7 as the sites of NOE when the imino signals of the Watson–Crick A·T pairs are irradiated for 100 ms. The Watson–Crick A·T pair is also shown in the inset along with the expected NOE pathway (shown by an arrow) when N3-H is irradiated. Spectrum 5 shows a 1D NOE at H8 of A1 upon irradiation of the imino signal at 10.8 ppm (presaturation time = 400 ms); such an NOE pattern is consistent with the A1·G9 pair as shown in Figure 2B with an NOE pathway N1-H(G) → N2-H(G) → H8(A) (NOEs are denoted as arrows). Spectrum 6 shows 1D NOEs at the NH_2 region and H8 of G10 upon irradiation of the imino signal at 10.8 ppm (presaturation time = 400 ms); such an NOE pattern is consistent with the G10·G10 pair as shown in Figure 2C, where N1-H(G) is close to N2-H(G) of the same base and H8(G) of the pairing partner (NOEs are denoted as arrows). Therefore, the imino signal and 1D NOE pattern in spectra 4–6 suggest the presence of a stem–loop motif (as shown in panel C) under conditions of high DNA concentration and high salt concentration. The H atom and the acceptor atom involved in H-bonding for the A·T, A·G, and G·G base pairs are joined by an open bar. (B, bottom left) 2D NOESY (τ_m = 250 ms) spectrum of $d(AATGG)_2$ in D_2O for the H1′ vs H8/H6 cross section (2.4 mM in DNA strand, 25 mM NaCl, pH 7, temperature 16 °C). The internucleotide connectivity pattern is indicative of a monomeric hairpin structure (Gupta et al., 1987). Full-matrix NOESY simulations with respect to the observed data at τ_m = 250 and 100 ms reveal that all nucleotides adopt a C2′-*endo*, *anti* conformation. Cross-peaks not on the H1′–H8/H6 connectivity route are due to the minor population of a stem–loop motif. (C, bottom right) 2D NOESY (τ_m = 250 ms) spectrum of $[d(AATGG)_2]_2$ in D_2O for the H1′,H3′ vs H8/H6 cross section (2.4 mM in DNA strand, 1 M NaCl, pH 7, temperature 3 °C). The intra- and internucleotide NOEs reveal that nine nucleotides (A1 through G9) exist predominantly in *C2′-endo*, *anti* conformations while one of two G10's shows an *anti* to *syn* conversion to facilitate the G10·G10 pair. Internucleotide NOEs involving H1′(A2/A7)–H2(A1/A6) (weak NOE) and H1′(G9)–H8(G10) (strong NOE) are indicative of special stacking patterns at the T·G and G·G steps, as discussed later in Figure 4. Note the high-field shift of H8,H1′(G10). Full-matrix NOESY simulations with respect to the observed data at τ_m = 250 and 100 ms allow us to extract 100 independent interproton distances as independent constraints for structure derivation. Intranucleotide H3′–H8/H6 NOEs are shown. Internucleotide N3′(i–1)-H8/H6(i) NOEs are also observed, but the connectivity pattern is not shown to preserve the clarity of the diagram. H3′ and H1′ chemical shift regions are nonoverlapping except for G10. Note the presence of the intermolecular NOESY cross-peak (marked *) between A1 and G10.

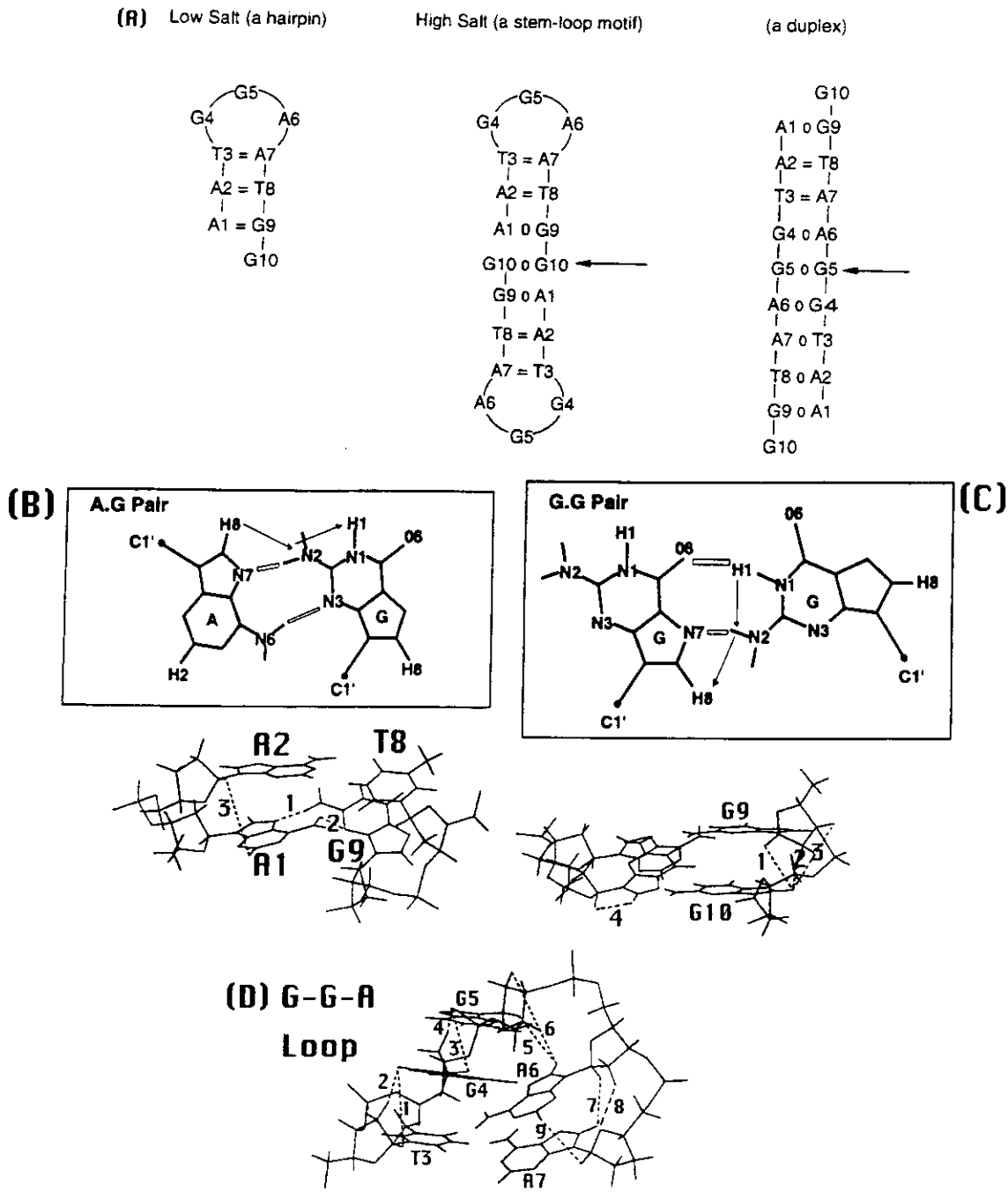


FIGURE 2: Schematic representations of (A) a monomeric hairpin, a stem–loop motif (in which two hairpins anneal with an internal 2-fold symmetry), and a duplex for $d(AATGG)_2$. (B) A·G pairing scheme and the crucial H-bonding and interproton contacts at the A1pA2·T8pG9 step of the representative model of $[d(AATGG)_2]_2$ (Figure 7): distance 1 (=2.02 Å), N7(A1)–HN2(G9); the corresponding H-bond angle N7(A1)–HN2(G9)–N2(G9) = 177°; distance 2 (=2.22 Å), N3(G9)–HN6(A1); the corresponding H-bond angle N3(G9)–HN6(A1)–N6-(A1) = 132°; distance 3 (=4.11 Å), H2(A1)–H1′(A2). The H-bond lengths and angles show acceptable H-bonding geometry. (C) G·G pairing scheme consistent with the imino proton profile and the NOE pattern at the G9pG10-G10-A1 step of the representative model of $[d(AATGG)_2]_2$ (Figure 7). Both G10's can be in *syn* or *anti* conformation (but not simultaneously) in order to preserve the G(*syn*)·G(*anti*) pairing scheme. The following interproton contacts are predicted in the G9(*anti*)pG10(*anti*) fragment from our model of Figure 7: distance 1 (=2.11 Å), H2″(G9)–H8(G10); distance 2 (=2.97 Å), H1′(G9)–H8(G10); distance 3 (=4.46 Å), H3′(G9)–H8(G10). In the G9(*anti*)pG10(*syn*) fragment the only prominent distance is 4 (=2.2 Å), H1′(G10)–H8(G10). Because of the *syn*/*anti* flip-flop of G10 the corresponding observed NOE should reflect the averages of the four distances. (D) Expected NOE pattern in the G-G-A loop of the representative model of $[d(AATGG)_2]_2$ (Figure 7): distance 1 (=4.44 Å), H1′(T3)–H8(G4); distance 2 (=2.28 Å), H2″(T3)–H8(G4); distance 3 (=3.73 Å), H1′(G4)–H8(G5); distance 4 (=2.42 Å), H2″(4)–H8(G5); distance 5 (=3.35 Å), H1′(G5)–H8(A6); distance 6 (=5.52 Å), H2″(G5)–H8(A6); distance 7 (=3.88 Å), H1′(A6)–H8(A7); distance 8 (=2.54 Å), H2″(A6)–H8(A7); distance 9 (=3.96 Å), H2(A6)–H1′(A7). Note that G5 is the most flexible part of the G-G-A loop. However, a sampling of different stacking conformations of G5 results in the effective (averaged) unbroken NOE pathway.

higher $d(AATGG)_2$ concentration (2.4 mM in strand), the same hairpin conformation was retained, as evidenced by imino proton spectrum 2 of Figure 1A. Under the conditions of high DNA concentration and low salt concentration, NOESY experiments at τ_m = 250 and 100 ms also confirmed a hairpin structure. Figure 1B shows the H1′ vs H8/H6 NOESY cross section (τ_m = 250 ms) typical of a monomeric hairpin (Gupta et al., 1987; Blommers et al., 1989; Williamson et al., 1989). The hairpin motif that is fully consistent with the NMR data is shown in Figure 2A. When the salt concentration is increased to 1 M NaCl (pH 7), two hairpins of $d(AATGG)_2$ form an end-stacked dimer (i.e., $[d(AATGG)_2]_2$) as shown

in Figure 2A. The NMR evidence of such a structure is discussed below.

(*i*) *The Presence of Watson–Crick A·T Pairs.* In our model the number of Watson–Crick A·T pairs equals the number of T's present in the sequence. For example, in the stem–loop structure of $[d(AATGG)_2]_2$ (Figure 2A), only two A·T pairs are expected, and this is exactly what we observe by 1D NOE experiment. In this experiment, the imino proton of T is irradiated (inset of Figure 1A, spectrum 4), and a strong NOE is observed at H2 of A even at 100 ms of presaturation time—a characteristic feature of a Watson–Crick A·T pair. NOEs are observed only at two H2's of A, suggesting that there are two Watson–Crick A·T pairs. And out of the four H2's belonging to four A's in $[d(AATGG)_2]_2$, H2's belonging to H-bonded A·T pairs show high-field shifts as expected. By thermal melting and chemical substitution studies, it was also shown that A·T pairs are crucial to the stability of the centromeric structures (Grady et al., 1992).

(*ii*) *The Nature of A·G Pairs.* Once the imino signals above 12 ppm are accounted for by the two A·T pairs, the signals below 11 ppm remain to be identified. As discussed below, they belong to G's in the A·G and G·G pairs and to G's in the loop. The sharp signal at 10.8 ppm belongs to the A·G pairs. It is clear that the A·G pairing is not through the imino proton of G because that should give the imino G signals above 12 ppm. This rules out the possibility of A(*syn*)·G(*anti*) and A(*anti*)·G(*anti*) pairing as observed in the single crystals (Kennard, 1988; Prive et al., 1988). The A(*syn*)·G(*anti*) pairing is also inconsistent with the NMR data because no A was observed in a *syn* conformation. The A(*anti*)·G(*anti*) pairing is also ruled out because the irradiation of the exchangeable signals below 12 ppm did not show any strong NOE at H2 of A (Kan et al., 1983). This leaves two other types of A·G pairings that involve amino protons of G instead of the imino protons (Li et al., 1991). One such pairing that is consistent with our NOE data is shown in Figure 2B. In this pairing, as observed in Figure 1A, spectrum 5, the irradiation of the imino proton of G at 10.8 ppm should show a secondary NOE at H8 of A via NH2 of G. Another additional feature of such an A·G pairing (as shown in Figure 2B) is the NOE between the H2 of A of the A·G pair and the H1′ of the neighboring 3′ A·T pair. Observation of such an NOE is shown in Figure 1C. Even though such an H-bonding, as shown below, has propeller-twisted A·G pairs, it has acceptable geometry and is free of any short sugar–base contacts. In this pairing scheme the imino protons of G do not participate in H-bonding; however, because of A-G-G stacking, the imino proton of the central G involved in the A·G pairing is excluded from solvent and hence not exchanged. Li et al. (1991) also demonstrated such an A·G pairing in a DNA duplex where the imino protons of G were not readily exchanged and located within 10–11 ppm.

(*iii*) *The Nature of G·G Pairing in the Stem–Loop Motif.* If the G·G pairing involved two imino protons of G's, then these two protons are expected to be located at two distinct chemical shifts and strong NOEs are expected between them. For the stem–loop structure of $[(AATGG)_2]_2$ (Figure 2A), the irradiations of the signals at 10.8 ppm did not produce any NOE at 9.9 ppm or vice versa (Figure 1A, spectra 5 and 6) as observed by Cognet et al. (1991) for a DNA duplex with G·G pairs. However, the G·G pairing scheme shown in Figure 2C is consistent with the NOE pattern shown in Figure 1A, spectrum 6, in which irradiation of the G imino signal at 9.9 ppm results in a primary NOE at NH2 of G and secondary NOE at H8 of G. In this pairing scheme G10 is in the *syn* conformation. In view of the fact that a single G10 signal is observed in panels A and C of Figure 1, it appears that there is a rapid *syn*/*anti* flip-flop without exposing the imino proton to the solvent. The observed intranucleotide H1′(G10)–H8(G10) NOE, though strong, is only the average of *syn*/*anti* conformations. The strong internucleotide H1′(G9)–H8(G10) contact is also consistent with the pairing shown in Figure 2C. The chemical shifts of the imino protons of G·G pairs as high-field-shifted as 9.9 ppm are not uncommon in the literature (Cognet et al., 1992; Smith & Feigon, 1992; Gupta et al., 1993). The G·G part happens to be the most flexible region of the stem–loop structures of $[d(AATGG)_2]_2$. Thermal melting and base substitution studies also confirm this observation (Grady et al., 1992); i.e., substitutions of the G·G pairs by any other mismatches have little effect on the melting temperature.

(*iv*) *The NOE Pattern of the G-G-A Loop.* Even under conditions of 1 M NaCl the imino signals of G's in the loop were still present, and these signals, as expected for a hairpin, did not show any specific NOE. Therefore, under the conditions of high DNA concentration and high salt concentration, two hairpins most probably anneal to form a stem–loop motif as shown in Figure 2A. The H1′ vs H8/H6 NOESY cross section (τ_m = 250 ms) of the stem–loop motif is shown in Figure 1C. Comparison of panels B and C of Figure 1 shows the following high salt induced changes: (i) a high-field shift of H8,H1′(G10) and strong H1′(G9)–H8(G10) NOEs (with a distance of 2.5–3.0 Å) providing information regarding stacking at the G-G step in the stem, (ii) *anti* to *syn* conversion of one of the G10's (as evidenced by the presence of a strong intranucleotide H1′–H8 NOE) consistent with the G·G pairing shown in Figure 2C, and (iii) emergence of weak but observable internucleotide H1′(A2/A7)–H2(A1/A6) NOEs (with distances of 3.8–4.2 Å) in agreement with the A·G pairing scheme shown in Figure 2B and also providing information regarding stacking at the T-G steps. Comparison of the base pairing in the duplex with the stem–loop motif (Figure 2A) shows a relatively small difference. Whereas the duplex has four internal A·T pairs, two internal A·G pairs, one internal G·G pair, two terminal A·G pairs, and two unpaired G's, the stem–loop motif has the same number of internal A·T, A·G, and G·G pairs. The expected difference in the two structure lies in the fact that two terminal A·G pairs and two unpaired G's in the duplex are replaced in the stem–loop motif by two G-G-A loops. However, as will be discussed later, the NOE constraints and energetic considerations can lead to an A·G pairing involving the first and the third base in the G-G-**A** loop (paired bases are shown in bold). Therefore, a stem–loop motif and a duplex can have the same number of base pairs and the same type of base stacking for all internal base pairs (Figure 2A). In this case, the energy difference in stacking between G in the loop of the stem–loop motif and the terminal G in the duplex essentially determines the difference in stability between these two structures. In the case of a stem–loop structure with three nucleotides (G-G-A) in the loop, the relative stabilities of a duplex and the corresponding end-stacked dimer (from now on referred to as the stem–loop structure) are determined by the difference in energy between *base stacking and base pairing in the duplex and base stacking and loop entropy in the stem–loop structure.*

In summary, the NMR data described above show that the centromeric repeat, $d(AATGG)_2$, at a low salt concentration forms a monomeric hairpin while at a high salt concentration two such hairpins anneal to form a stem–loop structure of $[d(AATGG)_2]_2$ that is stabilized by A·T, A·G, and G·G pairs (Figure 2A). The NOE contacts in the G-G-A loop as expected from our stem–loop structure of $[(AATGG)_2]_2$ are shown in

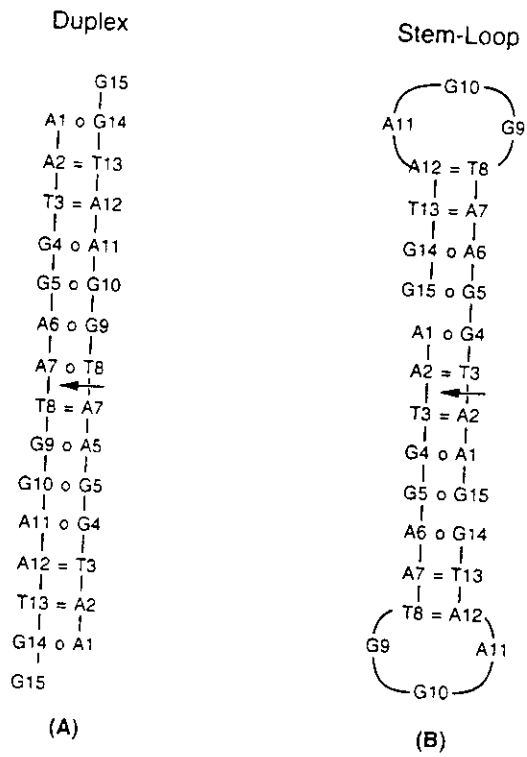


FIGURE 3: Schematic representations of (A) a duplex and (B) a stem–loop motif for $[d(AATGG)_3]_2$.

Figure 2D. Note the presence of continuous H1′(*i*)–H8(*i*+1) NOE contacts in the loop region and also the presence of the H2(A6)–H1′(A7) NOE. The NOE data shown in Figure 1C are in complete agreement with the structural model of Figure 2D. Although the NOE connectivities are often broken in the loop segment of a hairpin, such a criterion is never an absolute requirement of a loop structure. However, H2″(G*i*)–H8-(A*i*+1) NOEs in the loop segment are either weak or absent.

The Stem–Loop Motif of $[d(AATGG)_3]_2$. Results from the preceding NMR studies on $[d(AATGG)_2]_2$ show that the stem–loop motif utilizes all T's in A·T pairs. For $[d(AATGG)_3]_2$, utilization of all three T's in A·T pairing is achieved either in (A) a duplex or in (B) a stem–loop motif as shown in Figure 3. Spectra 1, 2, and 3 in Figure 4A show the imino signals of $d(AATGG)_3$ under different conditions: (i) for 1.8 mM $[d(AATGG)_3]_2$ concentration and low salt concentration (spectrum 1), (ii) for 1.8 mm $[d(AATGG)_3]_2$ and high salt concentration (spectrum 2), and (iii) in the presence of an equimolar complementary Watson–Crick strand, $[d(CCATT)_3]_2$ (spectrum 3). Spectra 1 and 2 show imino signals corresponding to three A·T pairs, two A·G pairs, one G·G pair, and G's in the loop, whereas spectrum 3 shows imino signals typical of Watson–Crick G·C and A·T pairs which were characterized by performing 2D NOESY experiments in (9:1) H_2O–D_2O (data not shown). The nature of the imino proton pattern in spectrum 3 and NOESY experiments in D_2O at various mixing times, τ_m = 150 and 50 ms (data not shown), unequivocally show that $d(AATGG)_3$·$d(CCATT)_3$ forms the usual Watson–Crick B-form duplex. Similar experiments on $d(AATGG)_2$·$d(CCATT)_2$ also indicate the presence of the usual Watson–Crick B-DNA conformation (data not shown). However, spectra 1 and 2 indicate that $d(AATGG)_3$ forms a stem–loop motif of $[d(AATGG)_3]_2$ under a wide range of salt concentrations. Figure 4B shows the H1′ vs H8/H6 NOESY cross section (τ_m = 250 ms) under conditions of high DNA concentration and low salt concentration observed for $[d(AATGG)_2]_2$ in Figure 1C, are also preserved here. These are (i) a high-field shift of H8,H1′(G5/G15) and strong H1′(G4/G14)–H8(G5/G15) NOEs (with distances of 2.5–2.8 Å) providing information regarding stacking at the G-G steps in the stem, (ii) *anti* to *syn* conversion of one of the G5/G15's (as evidenced by the presence of a strong intranucleotide H1′–H8 NOE) consistent with the G·G pairing shown in Figure 2C, and (iii) emergence of weak but observable internucleotide H1′(A2/A7/A11)–H2(A1/A6/A11) NOEs (with distances of 3.8–4.2 Å) in agreement with the A·G pairing scheme shown in Figure 2B and also providing information regarding stacking at the T-G steps. Analysis of NOESY data (τ_m = 250 and 100 ms) of $d(AATGG)_3$ under conditions of high salt also revealed the same stem–loop motif of $[d(AATGG)_3]_2$. However, as will be discussed, in addition to the three-nucleotide G-G-A loop, the NOE constraints and energetic considerations also indicate the possibility of an A·G pairing involving the first and the third base in the G-G-A loop (paired bases are shown in bold).

In summary, structural analyses of $[d(AATGG)_{2,3}]_2$ show that the stem–loop structure in each case (Figure 2A and Figure 3B) is stabilized by utilizing all T's in A·T pairs and formation of A·G and G·G pairs (which show enhanced stability at high salt). The melting profiles of $[d(AATGG)_{2,3}]_2$ show that the imino signals of the A·G and G·G pairs disappear at a much lower temperature than the imino signals of the A·T pairs. Therefore, it appears that the base pairing and base stacking at the A-T steps nucleate these structures.

The Stem–Loop Motifs of $[d(AATGG)_{4,6}]$*—an Extension of* $[d(AATGG)_{2,3}]_2$. Detailed analyses of 1D/2D NMR data of $d(AATGG)_{4,6}$ (data not shown) were unambiguous and revealed the presence of the stem–loop motifs shown in panels A and B of Figure 5. The stem–loop motif of $d(AATGG)_4$ (Figure 5A) is similar to the stem–loop motif of $[d(AATGG)_2]_2$ (under conditions of high salt, Figure 2A) except for an additional phosphodiester linkage. Also, the stem–loop motif of $d(AATGG)_6$ (Figure 5B) is similar to the stem–loop motif of $[d(AATGG)_3]_2$ (Figure 3B) except for an additional phosphodiester linkage.

The 1D NMR spectrum of $d(AATGG)_4$ showed the presence of imino signals belonging to four A·T pairs, two A·G pairs, one G·G pair, and the G's in the loop. Similarly, the stem–loop motif of $d(AATGG)_6$ (Figure 5B) is almost the same as the stem–loop motif of $[d(AATGG)_3]_2$ (Figure 3B) except for a phosphodiester covalent linkage. The 1D NMR spectrum of $d(AATGG)_6$ confirms the presence of the imino signals of six A·T pairs, four A·G pairs, two G·G pairs, and the G's in the two G-G-A loops (1D NMR spectra are included in Figure 8 of the supplementary material).

Comparisons of NOESY (τ_m = 250 and 100 ms) cross sections of $d(AATGG)_4$ and those of $[d(AATGG)_2]_2$ at high salt reveal a striking similarity: H8, H1′ of the nucleotides in the A-A, A-T, T-G, T-G, G-G steps in the stem region and H8, H1′ of the nucleotides in the loop G-G-A segment show similar chemical shifts and NOE patterns for the two repeat lengths. Further comparisons of NOESY (τ_m = 250 and 100 ms) cross sections of $d(AATGG)_6$ with those of $[d(AATGG)_3]_2$ show similarities in the chemical shift and NOE pattern (2D NMR spectra are included in Figures 9 and 10 of the supplementary material). Figure 5C shows the average stem–loop structures of $d(AATGG)_{4,6}$. The atoms are color coded (i.e., C = green, N = blue, O = red, P = yellow); the H-atoms are omitted to enhance the clarify of the diagram. In each case the average structure, taken over 200 local minima, is in agreement with the NMR data. In Figure 5C, the stem–loop

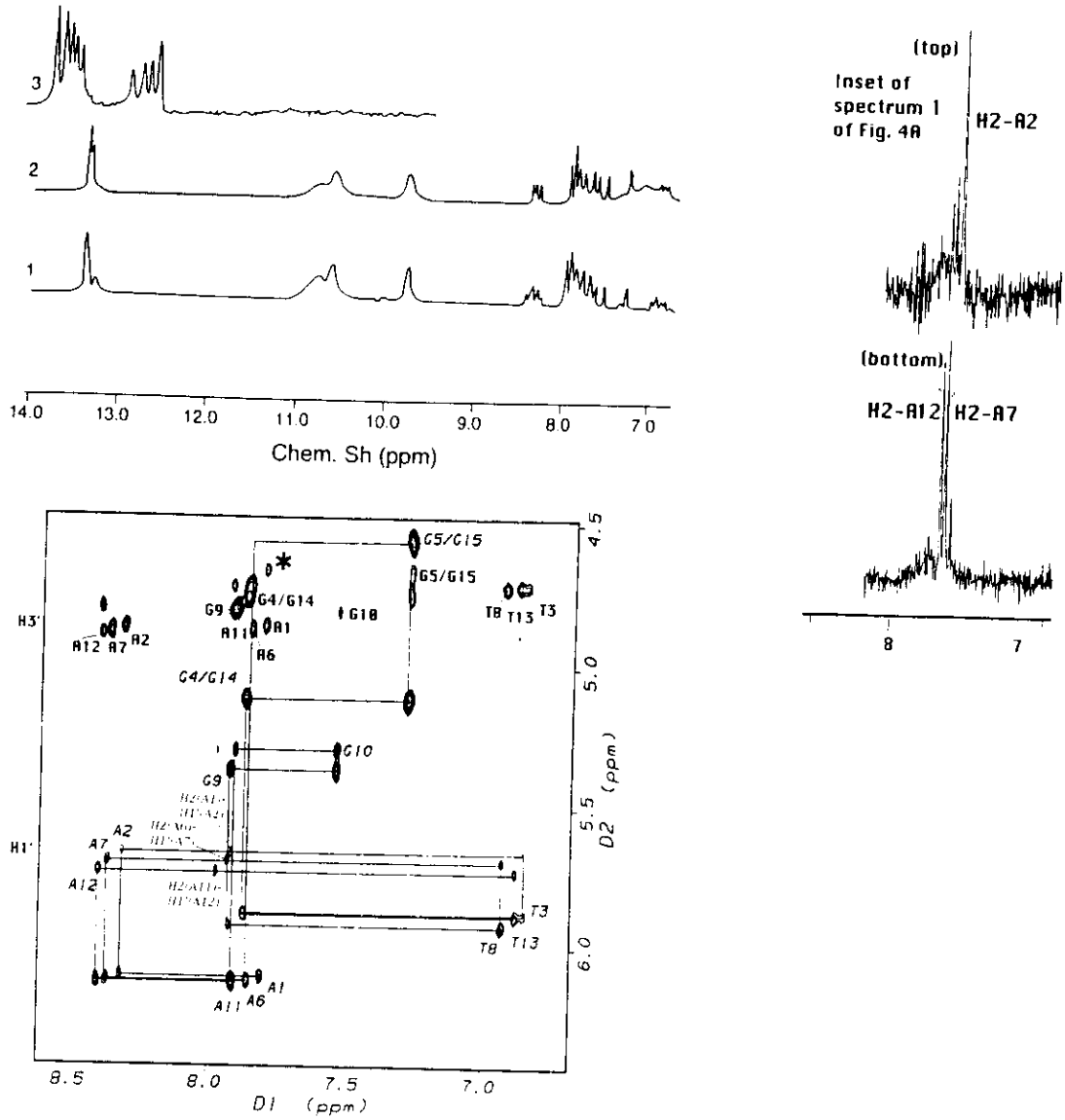


FIGURE 4: (A, top left) 1D NMR spectra of $d(AATGG)_3$ in a H_2O–D_2O (9:1) mixture under three different conditions: (spectrum 1) high DNA concentration (1.8 mM in strand) and low salt concentration (25 mM NaCl, pH 7) at 3 °C, (spectrum 2) high DNA concentration and high salt concentration (1 M NaCl, pH 7) at 3 °C, and (spectrum 3) in the presence of the complementary Watson–Crick strand $(CCATT)_3$ (1:1 molar ratio of two strands, 1 mM in duplex, 100 mM NaCl, pH 7, temperature 20 °C). 1D NOE data are shown in the inset (top right). When the smaller imino signal of A2·T8 in spectrum 1 is irradiated for 100 ms, a strong NOE is observed at H2 of A2 (the top NOE spectrum), and when the combined imino signals of T3·A7 and T13·A12 are irradiated, strong NOEs are observed at H2's of A7 and A12 (the bottom NOE spectrum). (B, bottom) 2D NOESY (τ_m = 250 ms) spectrum of $[d(AATGG)_3]_2$ in D_2O for the H1′,H3′ vs H8/H6 cross section (1.8 mM in DNA strand, 25 mM NaCl, pH 7, temperature 3 °C). The intra- and internucleotide NOEs reveal that 13 nucleotides (A1 through G4, A6 through G9, A11 through G14, and G10) exist predominantly in C2′-*endo*, *anti* conformations while either G5 or G15 shows an *anti* to *syn* conversion to facilitate the G5·G15 pairs. G10 resides in the loop segment. Internucleotide NOEs involving H1′(A2/A7/A12)–H2-(A1/A6/A11) (weak NOEs) and H1′(G4/G14)–H8(G5/G16) (strong NOEs) are indicative of special stacking patterns at the A-G and G-G steps, as discussed later. Note the high-field shift of H8,H1′(G5/G15). Full-matrix NOESY simulations with respect to the observed data at τ_m = 250 and 100 ms allow us to extract 150 independent interproton distances as structural constraints for molecular model building of a stem–loop motif. Intranucleotide H3′–H8/H6 NOEs are shown. Internucleotide H3′(i–1)–H8/H6(i) NOEs are also observed, but the connectivity pattern is not shown to preserve the clarity of the diagram. Note the presence of the intermolecular NOE (marked by *) between A1 and G15.

stem–loop structure of $d(AATGG)_6$ is shown on the right. Also shown are the approximate loop-folding axes drawn through the two central G's on the two loops. This clearly illustrates how the polynucleotide chain folds back onto itself to form a base-paired stem and two single-stranded loops. Stereoviews and different parts of the stem–loop structure are described in greater detail in Figure 7.

Additional Support for the Stem–Loop Structure. A stem–loop motif of $d(AATGG)_n$ differs from the corresponding non-Watson–Crick duplex (Figures 2, 3, and 5) in the following respects: (i) the overall size, e.g., as shown in Figure 5, the stem–loop motifs of $d(AATGG)_{4,6}$ are about half the length of the corresponding non-Watson–Crick duplex $d(AATGG)_{4,6}$·$d(AATGG)_{4,6}$, and (ii) the present of internal single-stranded loops in the stem–loop motif that are absent from the corresponding duplex. Therefore, the electrophoretic mobility of the stem–loop structure in a nondenaturing gel should be different from that of the duplex. Further, the internal single-stranded loops of the stem–loop structure should be susceptible to single-strand-specific *mung bean nuclease*, whereas the corresponding duplex should not be susceptible to single-strand scission.

Figure 6A shows the electrophoretic mobilities of $d(AATGG)_n$ in a nondenaturing gel. Note that $d(AATGG)_{4,6}$ migrate faster than the Watson–Crick duplexes $d(AATGG)_{4,6}$·$d(CCATT)_{4,6}$. This is consistent with the monomeric stem–loop

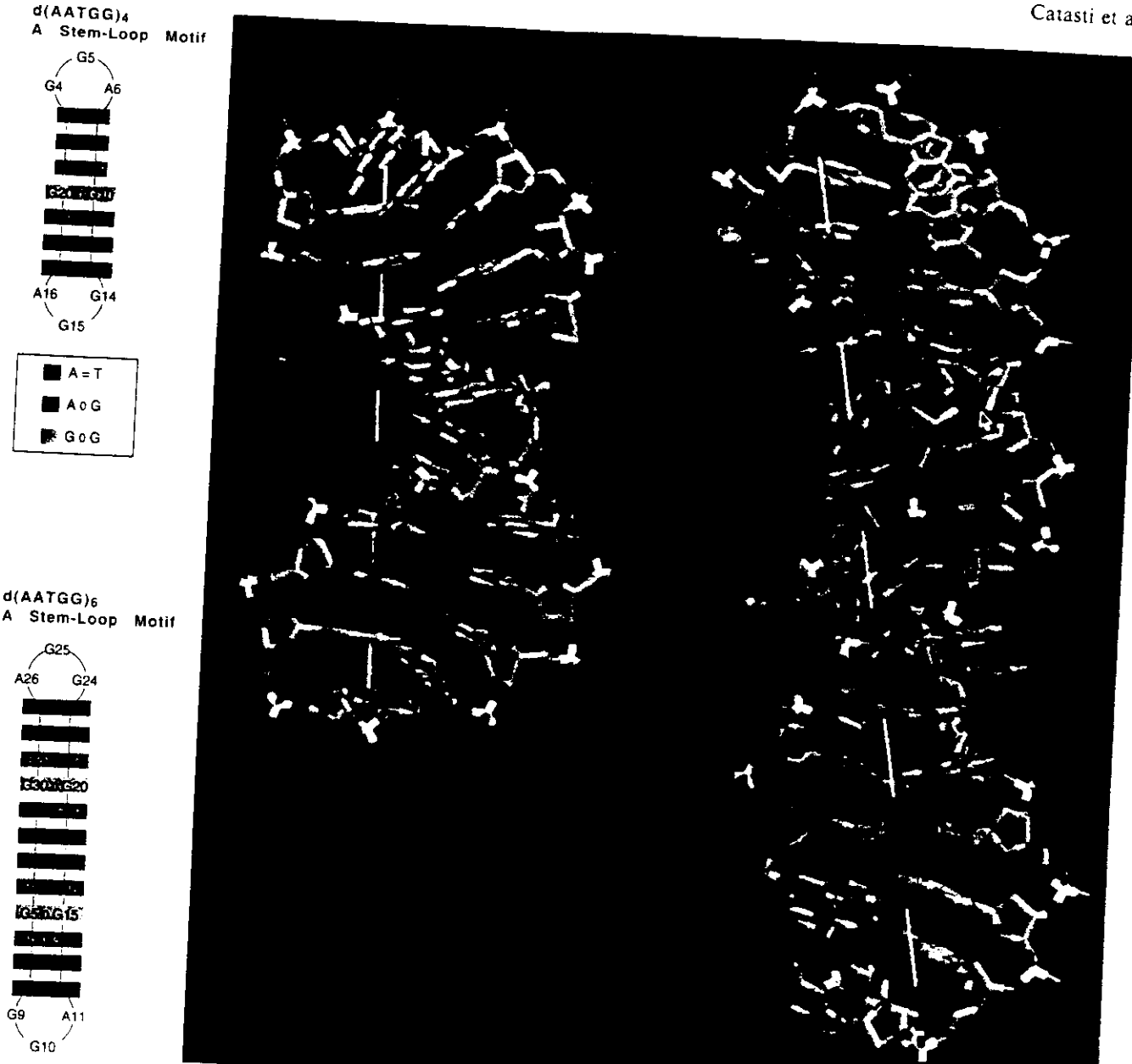


FIGURE 5: Schematic representations of stem–loop motifs for (A, top left) $d(AATGG)_4$ and (B, bottom left) $d(AATGG)_6$. (C, right) The skeletal model of the stem–loop structure of $d(AATGG)_4$ on the left and that of $d(AATGG)_6$ on the right. Only the non-hydrogen atoms are shown; C = green, N = blue, P = yellow, and O = red. The 5′-end of the structures is colored magenta while the 3′-end is colored cyan. In the stem–loop structure of $d(AATGG)_6$ the arrow is placed close to the 5′-end. The approximate axis of folding is also indicated by a dashed line. The structures represent the average of 200 sampled local minima obtained after simulated annealing subject to the NOE constraints.

structures of $d(AATGG)_{4,6}$ and not with the non-Watson–Crick duplexes $d(AATGG)_{4,6}$·$d(AATGG)_{4,6}$ because the latter is expected to migrate in a manner similar to that of the Watson–Crick duplex of the same size. Note that $d(AATGG)_6$ has the same gel mobility as the marker $d(CCATT)_4$, which is of shorter length and is a random coil under experimental conditions. Also note that $d(AATGG)_6$ migrates even faster than the Watson–Crick duplex $d(AATGG)_4$·$d(CCATT)_4$; this is consistent with the fact that the former is shorter in length than the latter. Similarly, $d(AATGG)_4$ migrates a little faster than the marker $d(CCATT)_3$ and much faster than the Watson–Crick duplex $d(AATGG)_4$·$d(CCATT)_4$. These observes support the presence of the stem–loop motifs for $d(AATGG)_{4,6}$. The gel electrophoresis patterns for $d(GGATT)_{4,6}$ are included in lanes 6 and 7 to show (without performing detailed NMR studies) that they also adopt similar stem–loop structures.

Figure 6B shows the digestion pattern of the stem–loop structures and different markers as produced by mung bean nuclease in a denaturing gel for different times of digestion. As shown in Figure 5B, two internal single-stranded loops are present in the stem–loop structure of $d(AATGG)_6$: one within nucleotides 8–12 and the other within nucleotides 23–27. Therefore, a single nick at any one of the loops is likely to produce DNA fragments of length greater than 20 but less than 25, whereas double nicks at both the loops are likely to produce DNA fragments of length greater than 10 but less than 15. Digestion of $d(AATGG)_6$ for 2.5 and 5 min produces such digestion products (lanes 2 and 3, Figure 6B) as expected for a stem–loop motif of $d(AATGG)_6$ (Figure 5B). Similarly, the single or the double nicks of $d(AATGG)_4$ (Figure 5A) are likely to produce DNA fragments of lengths greater than 10 but less than 15. Here also the nuclease treatment of $d(AATGG)_4$ for 2.5 and 5 min produces such digestion products (lanes 4 and 5, Figure 6B) as expected for a stem–loop motif of $d(AATGG)_4$ (Figure 5B).

The Structure of the Stem–Loop Motifs. Analyses of NOESY data (τ_m = 250 and 100 ms) of $[d(AATGG)_{2,3}]_2$ and

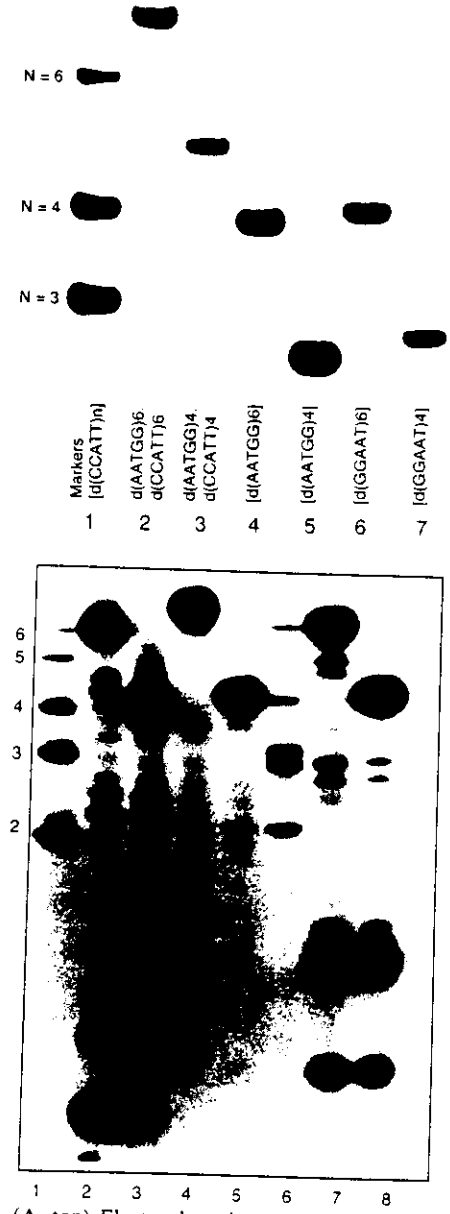


FIGURE 6: (A, top) Electrophoretic pattern of centromeric DNA repeats in nondenaturing gel. About 12 μL of the sample containing 0.1 mg of DNA was loaded in each lane. No mobility difference was found when the DNA concentration was changed. Lanes: 1, mixture of markers; 2, Watson–Crick duplex $d(AATGG)_6$·$d(CCATT)_6$; 3, Watson–Crick duplex $d(AATGG)_4$·$d(CCATT)_4$; 4, $d(AATGG)_6$; 5, $d(AATGG)_4$; 6, $d(GGAAT)_6$; 7, $d(GGAAT)_4$. (B, bottom) Digestion profiles of various centromeric repeats by mung bean nuclease (a probe for single-stranded regions in DNA). Lanes: 1, untreated $d(AATGG)_{2,3,4,5,6}$ used as markers; 2, digestion of $d(AATGG)_6$ for 2.5 min; 3, digestion of $d(AATGG)_6$ for 5 min; 4, digestion of $d(AATGG)_4$ for 2.5 min; 5, digestion of $d(AATGG)_4$ for 5 min; 6, untreated $d(GGAAT)_{2,3,4,5,6}$ and a 14-mer DNA used as markers; 7, digestion of $d(GGAAT)_6$ for 2.5 min; 8, digestion of $d(AATGG)_6$ for 5 min. Note that the duration of digestion does not alter the nature of cleavage.

$d(AATGG)_{4,6}$ with the aid of full-matrix NOESY simulations (Keepers et al., 1984; Gupta et al., 1988) resulted in a set of average interproton distances for various repeat lengths. The number of independent interproton distances was as follows: ~100 for $d(AATGG)_2$, ~150 for $d(AATGG)_3$, ~200 for $d(AATGG)_4$, and ~300 for $d(AATGG)_6$ (for illustration, chemical shift values and the NOESY data for the stem–loop structure of $[d(AATGG)_3]_2$ are included in Tables I and II of the supplementary material). Here we discuss the structures of $[d(AATGG)_2]_2$ and $d(AATGG)_4$ corresponding to the schematic models shown in Figures 2A and 5A, respectively. Using the interproton distances as structural constraints, we performed MD and energy minimization calculations (Weiner et al., 1986) in order to determine three-dimensional structures that satisfy the NMR data. A starting model of the stem–loop motif was constructed with a right-handed helical stem connecting two G-G-A loops. A left-handed helix that satisfied the observed NOEs could not be constructed. All structural parameters were taken close to B-DNA for the helical stem, except for the G·G base pair region where one of the G's in the pair adopted a *syn* conformation. For the central G·G base pair, two possibilities [i.e., G(*syn*)·G(*anti*) and G(*anti*)·G(*syn*)] were considered in our calculations. Constrained MD simulations were carried at 400 K, for 200 ps, with the purpose of sampling local and large-scale variations around a starting model (Garcia, 1992; Gupta et al., 1993). Snapshots of configurations along the MD trajectory were taken every 1 ps and subjected to constrained energy minimization (temperature quenching).

For $d(AATGG)_4$, an analysis of the rms distances among all pairs of quenched configurations followed by a hierarchical tree analysis (Gupta et al., 1993) revealed that two main families of stem–loop configurations were sampled. The main differences between both families of structure reside in the G-G-A loop region. The first family exhibits T3-G4-G5 stacking at the 5′-end and A6-A7 stacking at the 3′-end, with a two-hydrogen-bonded G4·A6 base pair similar to the base pairing found in the stem (Figure 5A). Thus, in this family two loops have the same conformation with only one unpaired base, and the G4·A6 base pair forms a part of the stem. The second family exhibits T3-G4 stacking at the 5′-end and G5-A6-A7 stacking at the 3′-end. This stacking is characteristic of DNA/RNA hairpin sequences with a Watson–Crick base-paired stem (Gupta et al., 1987; Cheong et al., 1990). This loop does not contain a G4·A6 base pair and therefore consists of three bases. In this family, only one loop adopts this conformation, whereas the other loop contains only one base for the first family of structures as described above. It is expected that each loop will independently exchange between the single- and three-base form, thus leading to four families of configurations: two symmetrical stem–loop motifs with two identical G-G-A loops (with or without a G·A base pair) and two nonsymmetrical stem–loop motifs with two different G-G-A loops (one with a G·A base pair and the other without a G·A base pair). The average energy of all minimized structures is 89.2 ± 2.3 kcal/mol, with 86.3 and 98.1 kcal/mol being the minimum and maximum energies, respectively. The average mean square distance among all structures is 1.48 Å^2. The energy differences among all structures (which account for all the conformational variants) are within 12.2 kcal/mol.

It may be noted that the energy difference between the loops with one and three unpaired bases is rather small. This becomes apparent by analyzing the relevant interaction energies in the loop segment of the stem–loop structures of $d(AATGG)_4$. For example, the interaction energy between G and A nucleosides in the loop with one unpaired base is approximately −5.0 kcal/mol of G·A, whereas the interaction energy between G and A nucleosides in the loop with three unpaired bases is approximately −3.0 kcal/mol. This difference is compensated by stacking interactions between bases in the loop and those in the stem such that the final stabilization energy difference is only 0.3 kcal/mol in favor of the conformer with loops of one unpaired base. It may also be noted that the energies in our HTMD/RTQ are only enthalpic contributions to the free energy. It is reasonable to expect that the loops with three unpaired bases will gain additional stability from the loop entropy by virtue of being inherently more

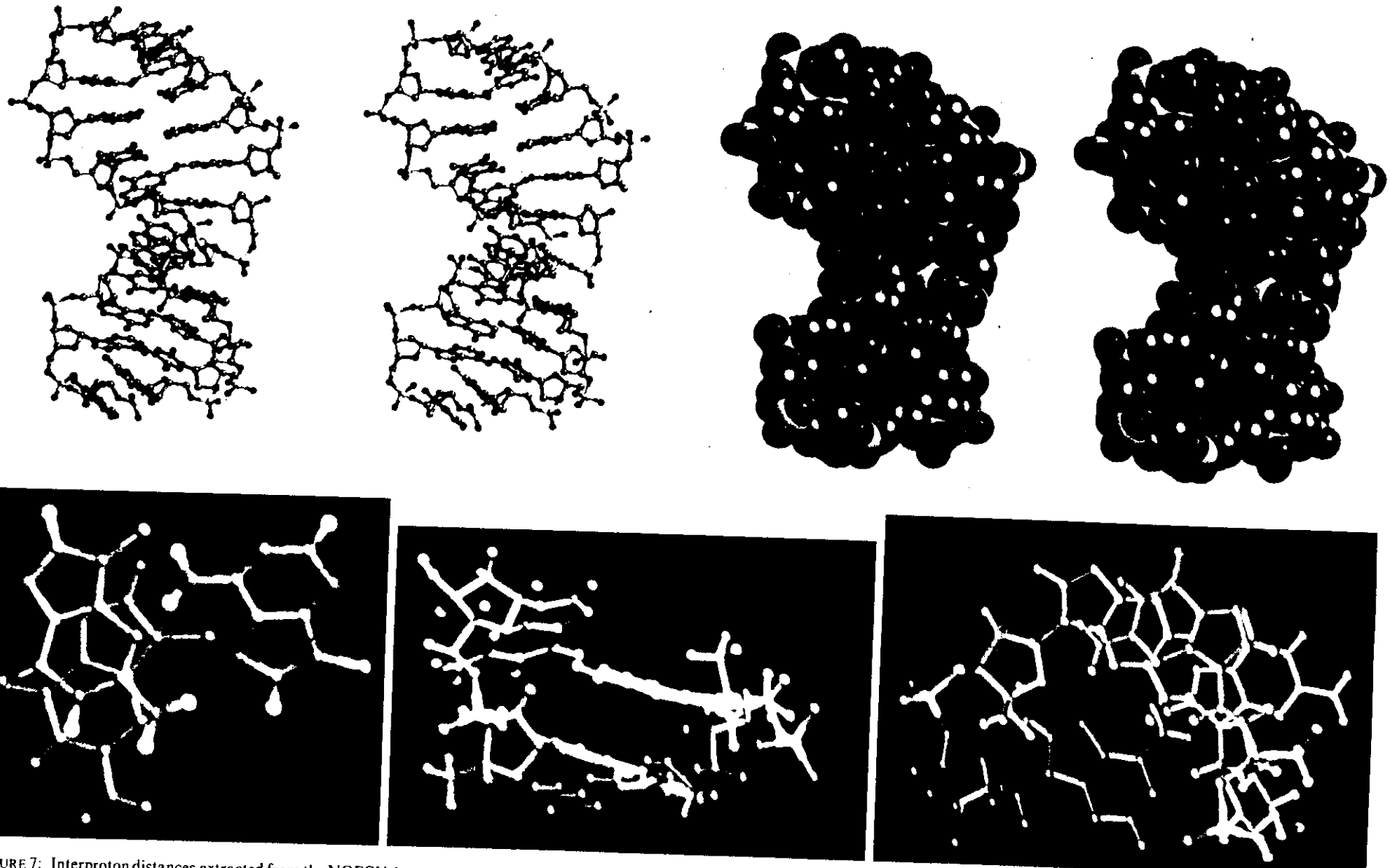


FIGURE 7: Interproton distances extracted from the NOESY data used to obtain energy-minimized models of stem–loop motifs for $d(AATGG)_4$. Only the non-hydrogen atoms are shown; C = gray, N = blue, P = yellow, and O = red. Note that in the model the stem regions contain A·T, A·G, and G·G pairs while the G-G-A constitutes the loop segments. (A, top left) Ball-and-stick model of the symmetric $d(AATGG)_4$ shown along the stem axis. (B, top right) CPK model of the symmetric $d(AATGG)_4$ shown along the stem axis. (C, bottom left) Horizontal view of the G-G-A stacking in this model. The unpaired G is shown in green. Note the presence of an A·G base pair involving the 5′G (left) and the 3′A (right). (D, bottom middle and right) Two views of stacking at the central G10–A11 step. A kink at this step leads to a localized bend of 31° for the stem. Note that the continuity of intrastrand stacking is broken at this base step. Because of unusual base pairing and stacking at the G-G-A segment of the stem, the helix describing the stem region is substantially unwound.

flexible. The activation barrier between two conformers (i.e., one with loops of three unpaired bases and the other with loops of a single unpaired base) is also small because such a transition can be locally achieved simply by moving or rotating away the 5′G and the 3′A in the loop. In view of the fact that the different loop configurations are almost equally stable and only a small barrier separates them, the 5′G and the 3′A are expected to show conformational equilibrium between the paired and unpaired states. The fact that the correspnoding G-NH signal is broad and sensitive to temperature change indicates a fast exchange of this proton within the NMR time scale, and hence this proton (and the corresponding A·G pair) is not locked only in the paired state. An RNA hairpin with an A·G pair between the 5′G and the 3′A has been reported (Heus & Pardi, 1991); in this hairpin two unpaired bases are present in the loop. Hairpins with loops containing a single nucleotide, though uncommon, are also observed in the single crystal structure of the human telomeric DNA d(GGGTTAGGG). In this structure, the 5′T and the 3′A (marked in bold) are involved in a Hoogsteen pair stacked on top of the G·G paired stem, thus leaving a single T in the loop (Alex Rich, personal communication, MIT).

Panels A and B of Figure 7 show the ball-and-stick and CPK models of the symmetrical stem–loop structure of $d(AATGG)_4$. The atoms are color coded (i.e., C = gray, N = blue, O = red, P = yellow). In this model both **G-G-A** loops have G·A base pairs between the 5′G and the 3′A in the loop (marked in bold). The base stacking arrangement in the loop is shown in Figure 7C where the unpaired G in the loop is colored uniformly in the green. The 5′G (on the right) and the 3′A (on the left) are colored by atoms (i.e., C = gray, N = blue, O = red, H = white).

The stem region of $d(AATGG)_4$ is a right-handed double helix unwound at the G·G pair in the stem; the unwinding also extends one base up and down the G·G pair in the stem. The stem pseudo-2-fold symmetry is broken by a G·G base pair, where one G is in a *syn* and the other is in an *anti* conformation. The stem regions separated by the G·G base pair are quite similar. A fitting of these two short helices to straight helices revealed that the two helix axes are kinked by 31°, resulting in a localized 31° bend at the central G·G base pair (Soumpasis et al., 1991). This feature is evident in Figure 7D, which shows the vertical and horizontal views of the central G10-A11 step. Note that the continuity of intrastrand stacking is broken at this base step.

As previously stated, the stem–loop motif of $[d(AATGG)_2]_2$ (Figure 2B) is similar to the $d(AATGG)_4$ model (Figure 5A) except for the lack of a phosphodiester linkage between G10 and A11. We therefore included the ball-and-stick (Figure 7E) and CPK (Figure 7F) stereoviews of $[d(AATGG)_2]_2$ in the supplementary material. As shown in Figure 7E,F, two identical hairpins are annealed in the stem–loop structure of $[d(AATGG)_2]_2$. In this model an A·G pair is present in the G-G-A loop with a stacking arrangement similar to that shown in Figure 7C.

Examination of the models in Figure 7 shows that Watson–Crick A·T pairs and stacking at A-T steps, A·G pairing and stacking at T-G steps, and G-G-A loop folding determine the stability of the stem–loop motifs of $d(AATGG)_{4,6}$ produced by two loop folding of the single strand. This is in agreement with the systematic thermal stability studies (Grady et al., 1992) that showed that any change in sequence that disrupted any one of these interactions reduced the stability of the structure.

DISCUSSION

A highly conserved repeated DNA sequence, $d(AATGG)_n$, has been identified in human centromeric regions, and for the reasons already discussed (Grady et al., 1992), it may represent a functional component of the human centromere. The high degree of sequence conservation exhibited by this sequence among diverse species (Grady et al., 1992), as great as or greater than that exhibited for animal telomeric DNA sequences (Meyne et al., 1989), suggests ongoing selective pressures to maintain this sequence. It is tempting to speculate that, like telomeric sequences (Zakian et al., 1989; Kang et al., 1992), an unusual DNA structure rather than an invarient nucleotide sequence may be the consensus recognition element for this sequence. As a non-self-complementary sequence, repeats of d(AATGG) exhibit unusual thermal stability, implying an intra- or interchain conformation with extensive base pairing (Grady et al., 1992).

Here we show by multi-dimensional NMR spectroscopy and nuclease digestion studies that repeats of d(AATGG) form a stable stem–loop structure. Formation of such a structure is expected to cost little free energy because the Watson–Crick complementary sequence d(CCATT) could easily fold back and form a triple helix with an enhanced stability at acidic pH to facilitate C^+·G·C triplets (Durland et al., 1991). Structural changes in the chromosome during the processes of chromosome condensation, particularly changes in DNA supercoiling, could induce a transition from a B-DNA duplex to the stem–loop or triple helix conformation, providing DNA binding sites for the proteins associated with the kinetochore (Pluta et al., 1990). The possibility of such a structural intermediate can be tested in two ways: (i) by using NMR spectroscopy to detect and determine the structure of a triple helix in isolated $d(AATGG)_n$·$d(CCATT)_n$ fragments at different pHs and (ii) by determining the ability of a $d(AATGG)_n$·$d(CCATT)_n$ insert in a plasmid DNA to induce negative superhelicity as a function of pH. A third approach is to search for proteins that recognize the stem–loop structures of the $d(AATGG)_n$ sequence under ionic conditions that stabilize them.

ACKNOWLEDGMENT

We thank Drs. Neville Kallenbach, Luis Marky, P. Reitemeier, C.-S. Tung, and C. Burks for their comments on the manuscript. The majority of the NMR work was done at the NMR Facility at the University of California, Davis, using the GE 500-MHz spectrometer (funded by NSF Grant DIR-88-04739 and USPHS Grant RR04795). Some of the NMR work was also done at Iowa State University, Ames, IA. The help and hospitality of Dr. A. Kintanar at the NMR facility of Iowa State University are greatly appreciated. We thank the Advanced Computing Laboratory for providing their facilities.

SUPPLEMENTARY MATERIAL AVAILABLE

Four figures showing the stem–loop structure of $[d(AATGG)_2]_2$ and 1D NMR spectra and 2D NOESY spectra of $d(AATGG)_4$ and $d(AATGG)_6$ and three tables giving chemical shift values of the stem–loop structures of $[d(AATGG)_2]_2$ and $[d(AATGG)_3]_2$ and NOESY data for $[d(AATGG)_3]_2$ (15 pages). Ordering information is given on any current masthead page.

REFERENCES

Blommers, M. J. J., Walters, J. A. L. I., Hassnoot, C. A. G., Aelen, J. M. A., van der Marel, G. A., van Boom, J. H., &

Hilbers, C. W. (1989) Effects of base sequence on the loop folding in DNA hairpins, *Biochemistry 28*, 7491–7498.

Cheong, C., Varini, G., & Tinoco, I., Jr. (1990) Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC, *Nature 346*, 680–682.

Cognet, J. A. H., Gabarro-Arpa, J., Bret, M. L., van der Marel, G. A., van Boom, J. H., & 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.

Durland, R. H., Kessler, D. J., Gunnel, S., Duvic, M., Pettitt, A., & Hogan, M. E. (1991) Binding of triple helix forming oligonucleotides to sites in gene promoters, *Biochemistry 30*, 9246–9255.

Garcia, A. E. (1992) Large-amplitude nonlinear motions in proteins, *Phys. Rev. Lett. 68*, 2696–2699.

Garcia, A. E., & Soumpasis, D. M. (1989) Harmonic vibrations and thermodynamic stability of a DNA oligomer in monovalent salt solution, *Proc. Natl. Acad. Sci. U.S.A. 86*, 3160–3164.

Garcia, A. E., Gupta, G., Soumpasis, D. M., & Tung, C. S. (1990) Energetics of the hairpin to mismatched duplex transition of d(GCCGCAGC) on NaCl solution, *J. Biomol. Struct. Dyn. 8*, 173–186.

Grady, D. I., Ratliff, R. L., Robinson, D. L., McCanlies, E. C., Meyne, J., & Moyzis, R. K. (1992) Highly conserved repetitive DNA sequences are present at human centromeres, *Proc. Natl. Acad. Sci. U.S.A. 89*, 1695–1699.

Gupta, G., Sarma, M. H., Sarma, R. H., Bald, R., Engelke, U., Oei, S. L., Gessner, R., & Erdmann, V. A. (1987) DNA hairpin structures in solution: 500-MHz two-dimensional ^{1}H NMR studies on d(CGCCGCAGC) and d(CGCCGTAGC), *Biochemistry 26*, 7715–7723.

Gupta, G., Sarma, M. H., & Sarma, R. H. (1988) On the question of DNA bending: two-dimensional NMR studies on d(GTTTAAAAC)$_2$ in solution, *Biochemistry 26*, 7909–7919.

Gupta, G., Garcia, A. E., & Hiriyanna, K. T. (1993) Sampling of the conformations of the d(CGCTGCGGC) hairpin in solution by two-dimensional nuclear magnetic resonance and theoretical methods, *Biochemistry 32*, 948–960.

Heus, H. A., & Pardi, A. (1991) Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops, *Nature 253*, 191–194.

Kan, L.-S., Chandrasegaran, S., Pulford, S. M., & Miller, P. S. (1983) Detection of a guanine·adenine base pair in a decadeoxyribonucleotide by proton magnetic resonance spectroscopy, *Proc. Natl. Acad. Sci. U.S.A. 80*, 4263–4265.

Kang, C. H., Zhang, X., Ratliff, R., Moyzis, R., & Rich, A. (1992) Crystal structure of four-stranded *Oxytricha* telomeric DNA, *Nature 356*, 126.

Keepers, J. W., & James, T. L. (1984) A theoretical study of distance determinations from NMR. Two-dimensional nuclear overhauser effect spectra, *J. Magn. Reson. 57*, 404–426.

Kennard, O. (1988) Structural studies of base pair mismatches and their relevance to theories of mismatch formation and repair, in *Structure and Expression: DNA and its Drug Complexes* (Olson, W. K., Sarma, R. H., Sarma, M., & Sundaralingam, M. S., Eds.) Vol. 2, pp 1–26, Adenine Press, Schenectady, NY.

Li, Y., Zon, G., & Wilson, W. D. (1991) NMR and molecular modeling evidence for a G·A mismatch base pair in a purine-rich DNA duplex, *Proc. Natl. Acad. Sci. U.S.A. 88*, 26–30.

Meyne, J., Ratliff, R. L., & Moyzis, R. K. (1989) Conservation of the human telomere sequence (TTAGGG)$_n$ among vertebrates, *Proc. Natl. Acad. Sci. U.S.A. 86*, 7049–7053.

Moyzis, R. K., Buckingham, J. M., Cram, L. S., Dani, M., Deaven, L. L., Jones, M. D., Meyne, J., Ratliff, R. L., & Wu, J.-R. (1988) A highly conserved repetitive DNA sequence, (TTAGGG)$_n$, present at the telomeres of human chromosomes, *Proc. Natl. Acad. Sci. U.S.A. 85*, 6622–6626.

Moyzis, R. K., Torney, D. C., Meyne, J., Buckingham, J. M., Wu, J.-R., Burks, C., Sirotkin, K. M., & Goad, W. B. (1989) The Distribution of interspersed repetitive DNA sequences in the human genome, *Genomics 4*, 273–289.

Pluta, A. F., Cooke, C. A., & Earnshaw, W. C. (1990) Structure of the human centomere at metaphase, *Trends Biochem. Sci. 15*, 181–185.

Prive, G. G., Heinemann, U., Chandrasegaran, S., Kan, L. S., Kopka, M., & Dickerson, R. E. (1988) Structural studies of base pair mismatches and their relevance to theories of mismatch formation and repair, in *Structure and Expression: DNA and its Drug Complexes* (Olson, W. K., Sarma, R. H., Sarma, M., & Sundaralingam, M. S., Eds.) Vol. 2, pp 27–48, Adenine Press, Schenectady, NY.

Rich, A., Nordhem, A., & Wang, A. H.-J. (1984) The Chemistry and Biology of Left-Handed Z-DNA, *Annu. Rev. Biochem. 53*, 791–846.

Sklenar, V., & Bax, A. (1987) Spin-echo water suppression for the generation of pure-phase two-dimensional NMR spectra, *J. Magn. Reson. 74*, 469–479.

Smith, F. W., & Feigon, J. (1992) Quadruplex structure of *Oxytricha* telomeric DNA oligonucleotides, *Nature 356*, 164–168.

Soumpasis, D. M., Tung, C.-S., & Garcia, A. E. (1991) Rigorous description of DNA structures. II. On the computation of best axes, planes, and helices from atomic coordinates, *J. Biomol. Struct. Dyn. 8*, 867–888.

Stillinger, F. H., & Weber, T. A. (1983) Dynamics of structural transitions in liquids, *Phys. Rev. A 28*, 2408–2416.

Weiner, S. J., Kollman, P. A., Nguyen, D. T., & Case, D. A. (1986) An all atom force for simulations of proteins and nucleic acids, *J. Comput. Chem. 7*, 230–245.

Williamson, J. R., & Boxer, S. G. (1989) Multinuclear NMR studies of DNA hairpins. 1. Structure and dynamics of d(CGCGTTGTTCGCG), *Biochemistry 28*, 2831–2836.

Zakian, V. A. (1989) Structure and function of telomeres, *Annu. Rev. Genet. 23*, 579–604.