Non-B DNA Conformations, Genomic Rearrangements, and Human Disease*

Published, JBC Papers in Press, August 23, 2004, DOI 10.1074/jbc.R400028200

Albino Bacolla and Robert D. Wells‡
From the Institute of Biosciences and Technology, Center for Genome Research, Texas A&M University System Health Science Center, Texas Medical Center, Houston, Texas 77030

The history of investigations on non-B DNA conformations as related to genetic diseases dates back to the mid-1960s. Studies with high molecular weight DNA polymers of defined repeating nucleotide sequences demonstrated the role of sequence in their properties and conformations (1). Investigations with repeating homo-, di-, tri-, and tetrarnucleotide repeating motifs revealed the powerful role of sequence in molecular behaviors. At that time, this concept was heretical because numerous prior investigations with naturally occurring DNA sequences masked the effect of sequence (1). It may be noted that these studies in the 1960s predated DNA sequencing by at least a decade.

Early studies were followed by a number of innovative discoveries on DNA conformational features in synthetic oligomers, restriction fragments, and recombinant DNAs. The DNA polymorphisms were a function of sequence, topology (supercilie density),ionic conditions, protein binding, methylation, carcinogen binding, and other factors (2). A number of non-B DNA structures have been discovered (approximately one new conformation every 3 years for the past 35 years) and include the following: triplexes, left-handed DNA, bent DNA, cruciforms, nodule DNA, flexible and writhed DNA, G4 tetrad (tetraplexes), slipped structures, and sticky DNA (Fig. 1). From the outset, it was realized (1, 2) that these sequence effects probably have profound biological implications, and indeed their role in transcription (3) and in the maintenance of telomere ends (4) has recently been reviewed.

However, in the past few years dramatic advances from genomics, human genetics, medicine, and DNA structural biology have revealed the role of non-B conformations in the etiology of at least 46 human genetic diseases (Table I) that involve genomic rearrangements as well as other types of mutation events.

Non-B DNA Conformations

Segments of DNA are polymorphic. A large number of simple DNA repeat sequences can exist in at least two conformations. All of these sequences adopt the orthodox right-handed B form, probably for the majority of the time, with Watson-Crick (WC) A-T and G-C bp. However, at least 10 non-B conformations (5–7) are formed, perhaps transiently, at specific sequence motifs as a function of negative supercilie density, generated in part by transcription, protein binding, and other factors.

Non-B DNA structures (Fig. 1) include cruciforms and left-handed Z-DNA, which have bp in the WC conformation, whereas triplexes, tetraplexes, and slipped structures rely in part, or exclusively, on non-WC interactions. Cruciform DNA occurs at inverted repeats (IR), which are defined as sequences of identical composition on the complementary strands. Each strand folds at the IR center of symmetry and reconstitutes an intramolecular B-helix capped by a single-stranded loop, which may extend from a few bp to several kb. The term cruciform originates from the appearance of the duplex arms (5, 6), which at the 4-way junction adopts either an “open” configuration that allows strand migration (Fig. 1) or a “stacked” (locked) arrangement, where the helices stack upon each other. In both cases, the overall conformation and the intraduplex angles mimic the Holliday junction recombination intermediates (8). Cruciform DNA is often incorrectly referred to as a palindromic; a palindromic is a word that reads identically when spelled backwards, such as rotor or Hannah.

Triplex (triple helix, H-DNA) DNA (Fig. 1) occurs at oligo (purine-pyrimidine) ((R-Y)n) tracts (5, 6, 9, 10) and is favored by sequences containing a mirror repeat symmetry. The purine strand of the WC duplex engages the third strand through Hoogsteen hydrogen bonds in the major groove while maintaining the original duplex structure in a B-like conformation. Triplex DNA is polymorphic in the orientation and composition of the third strand; this may be either pyrimidine-rich and parallel to the complementary R-strand or purine-rich and antiparallel to the complementary R-strand (the latter case is shown in Fig. 1). In addition, the participating third strand (10) may originate from within a single (R-Y)n tract (intramolecular triplex) or from a separate tract (intermolecular triplex).

Direct repeats (DR) aligning out of register give rise to slipped structures with looped-out bases (Fig. 1). When DR involve several repeating motifs, like the telomeric or triplet repeat sequences, the looped-out bases may form duplexes stabilized by unusual pairs, such as T-T in (CTG)n (11) or shared paired motifs (G-A and G-C-AA), a specific arrangement characterized by interstrand, rather than intrastrand, stacking interactions (12).

A guanine tetrad, as found in telomeric DNA, is a square coplanar array of four guanines (4, 7). Each base donates two Hoogsteen hydrogen bonds to one neighbor and receives two hydrogen bonds from the other neighbor in a cyclic arrangement involving N-1, N-2, O-6, and N-7. Four-stranded DNA (tetraplex DNA, G-quadruplex) is composed of guanine tetrads stacked upon each other (Fig 1) and therefore is favored by four runs of 3 or more guanines. This conformation is also endowed with a high degree of polymorphism in terms of strand stoichiometry and polarity, glycosidic torsion angles, groove size, and connecting loops. Pairing schemes with other bases has also been observed, such as a guanine tetrad sandwiched between two G-C and A-T WC pairs (13).

Left-handed Z-DNA (Fig. 1) is adopted by alternating pyrimidine-purine (Y-R-Y-R) sequences such as (CG-GC)n and (CA-TG)n. The transition from the right-handed B- to the left-handed Z-form is accomplished by a 180° flip (upside-down) of the bp through a rotation of every other purine from the anti to the syn conformation and a corresponding change in the sugar-puckering mode from the C2'- to the C3'-endo conformation. Contrary to the B-DNA, which possesses a major and a minor groove, Z-DNA contains only one groove (3, 6). Other architectures (sticky DNA, nodule DNA, bent DNA, i-motifs, etc. (5–7)) are well known; their roles in genome instability are under current investigation.

Chromosomal Rearrangements

The involvement of double-strand breaks in genome instability is well documented (14), but the initiating events that lead to their formation are not fully understood. Recently, an increasing amount of information has revealed a major role for non-B DNA conformations. Indeed, analyses of large deletions stimulated by a 2.5-kb poly(Y-R) sequence from intron 21 of the polycystic kidney disease 1 gene (PKD1) or by long (CTG-CAG)n repeats in Escherichia coli indicated that the breakpoints occurred invariably at nucleotides.

* This minireview will be reprinted in the 2004 Minireview Compendium, which will be available in January, 2005. This research was supported by National Institutes of Health Grants NS37554 and ES11347 and by grants from the Robert A. Welch Foundation, Muscular Dystrophy Association, Friedreich’s Ataxia Research Alliance, and Seek-a-Miracle Foundation.
‡ To whom correspondence should be addressed. Tel.: 713-677-7651; Fax: 713-677-7689; E-mail: rwell@bt.tamuss.edu.

© 2004 by The American Society for Biochemistry and Molecular Biology, Inc.
abutting or within motifs capable of adopting non-B conformations (15). Because the same PKD1 tract also induced cell death in a supercoil-dependent manner (16), we concluded that the large deletions were mediated by supercoil-dependent non-B conformations. In addition, the presence of short (direct or inverted) homologies at the breakpoint junctions suggested a model whereby two distantly located non-B DNA conformations induced double-strand break repair and hence demarcated the deletion end points. Similarly, a search of 222 rearrangement breakpoints from the human Gross Rearrangement Breakpoint Database (GRaBD) revealed that both the frequency and the proximity of \((R/Y)\), and \((B/Y)\), tracts to breakpoint junctions were significantly greater than expected by chance (15). A detailed analysis of the DNA sequences flanking the breakpoints in 11 such deletions showed that, as in the bacterial cases, the events were explicable in terms of non-B DNA formation. Hence, most (if not all) non-B conformations appear to play a widespread role in rearrangement mutagenesis.

A second line of evidence in establishing a role for DNA structural features in rearrangements was the recent identification of extremely large blocks of chromosome-specific repetitive sequences (termed low copy repeats (LCRs), amplicons, duplicons, or segmental duplications). Such blocks constitute a fertile substrate for recurrent rearrangements associated with >40 human genomic diseases (Table I) (17–19) because they may adopt conformations of unprecedented complexity and size. Chromosome 22q11 is characterized by four LCRs (LCR-A to -D) that encompass ~350 kb of DNA over a ~3-Mb region (20, 21). Sequence analyses of constitutional translocation breakpoints between chromosome 22 and chromosomes 1, 4, 11, or 17 show the hallmarks of cruciform DNA. For example, the breakpoints on LCR-B clustered in a narrow IR AT-rich region that, in conjunction with flanking NP-1-like sequences (17, 22), is predicted to form a ~600-bp-long cruciform. Furthermore, the clustering occurred within 15 bp from the IR center and involved the symmetric deletion of a few base pairs on either side of the IR apex (23), consistent with the prediction that cruciform loops are susceptible to cleavage and hence to repair (24). The breakpoints on chromosomes 1p21.2, 4q35.1, 11q23, and 17q11.2 also occurred at, or near, the centers of IR AT-rich sequences (17). Interestingly, a case of ependymoma associated with the t(1;22) translocation revealed that the breakpoint on chromosome 1 was flanked by six copies of a 33-bp-long motif with IR symmetry (17). Conversely, the human genome assembly clone indicated the presence of only two such repeats, suggesting that expansion may have predisposed that individual to instability, possibly by increasing the propensity (and/or stability) for cruciform formation. Recombination events between highly homologous (97–98%) sequences within LCR-A to -D are associated with recurrent deletions and duplications (~1:3,000 live births), leading to the DiGeorge, velo-cardiofacial, conotruncal anomaly face and cat-eye syndromes (20, 21, 25) (Table I), suggesting that non-B DNA conformations may also be involved in these conditions.

Intergenic deletions in male-specific regions of the Y chromosome (Yq) are a major source of infertility (AZF) (26, 27) (Table I). Of the commonly affected regions (AZFa, AZFb, and AZFc), AZFb and AZFc are located within an ~8-Mb interval and are mostly comprised of five large IR (P1–P5), each consisting of a complex array of direct and inverted blocks harboring male-specific genes. Extensive (>90 kb) regions of near perfect identity exist among these IRs; nevertheless, the breakpoints strongly clustered toward the IR spacers where, in some cases, they took place between sequences that shared no homology. These features are consistent with a cruciform-mediated mechanism for the deletions (18, 27).

Isochromosomes are abnormal chromosomes that contain a duplication in a single chromosome arm (25). Isochromosome 17q (i(17q)) is the most common isochromosome associated with human neoplasia, including neuroectodermal tumors and medulloblastomas, representing ~3% of single or additional chromosomal abnormalities in carcinomas and hematologic malignancies, and occurs

---

**FIG. 1.** Non-B DNA conformations involved in rearrangements.

| Name                  | Conformation            | General Seq. Requirements | Sequence |
|-----------------------|-------------------------|---------------------------|----------|
| Cruciform             |                         |                           | TCCTGAGA CACCTGACCT |
| Triplex               | (R+Y)\(_n\) Mirror      |                           | AAGAGGGTGAAGTTCCCTT |
|                      | Repeats                 |                           | TCTCCTCAGCAAG |
| Slipped (Hairpin)     | Direct Repeats          |                           | TCGTTTCGCT AGCAAGCA |
| Structure             |                         |                           | AG3(T2AG3)3 single strand |
| Tetraplex             |                         |                           | CGCGTGCGCTGGCACGACACAC |
| Left-handed Z-DNA     |                         |                           | B-Z Junctions |

---

*2 M. Wojciechowska and R. D. Wells, unpublished data.*
Table 1
Selected genetic alterations, diseases, and DNA repeat motifs

| Genetic alteration | Syndrome or metabolic event | DNA motifs | Ref. |
|--------------------|-----------------------------|------------|------|
| t(11;22)(q23;q11.2) | Supernumerary der (22) | IR | 17, 22, 23 |
| t(17;22)(q11.2;q11.2) | Neurofibromatosis type 1 | IR | 17 |
| t(1;22)(p13.1;q11.2) | Epidermodysplasia | IR | 17 |
| t(4;22)(q35.1;q11.12) | Velocardiofacial | IR | 17 |
| i(17q) | Hematologic malignancies (chronic myeloid leukemia) | IR, DR | 18, 26, 27 |
| del(Yq) | Spermatogenic failure | IR, DR | 20, 21 |
| del(17q)(p11.2)p11.2) | DiGeorge, velocardiofacial, conotruncal anomaly face | IR, DR | 20, 21 |
| del(17q)(p11.2)p11.2) | Cat-eye | IR, DR | 25, 28, 29 |
| del(17q)(p11.2)p11.2) | Smith-Magenis | IR, DR | 25, 28, 29 |
| del(17q)(p11.2)p11.2) | Charcot-Marie-Tooth type 1 | IR, DR | 25, 28, 29 |
| del(17q)(p11.2)p11.2) | Hereditary neuropathy with liability to pressure palsies | IR, DR | 46 |
| del(17q)(p11.2)p11.2) | Williams-Beuren | IR, DR | 46 |
| inv(7)(q11.23) | Predisposition to Williams-Beuren | IR | 48 |
| del(15)(q11.23) | Prader-Willi and Angelman | IR, DR | 47 |
| t(X;22)(q13;q11) | Meloschizis and lumbosacral spina bifida | IR | 48 |
| Expanded (CTG-CAG)\_n | Expanded polyglutamine diseases (Haw River) | DR | 30, 31 |
| Expanded (CTG-CAG)\_n | Huntington, Huntington’s disease-like 2, spinobulbar muscular atrophy (Kennedy), spino-cerebellar ataxia (SCA) 1 | DR | 30, 31 |
| Expanded (CTG-CAG)\_n | Myotonic dystrophy type 1, SCA5, SCA12 | DR | 31 |
| Expanded (CTG-CAG)\_n | Expanded polyaniline diseases (infantile spasm, cleidocranial dysplasia, blepharophimosis/potiosis/epicanthus inversus type B, hand-foot-genital, synpolydactyly, oculopharyngeal muscular dystrophy, holoprosencephaly, oculopharyngeal muscular dystrophy) | DR | 31, 49 |
| Expanded (CTG-CAG)\_n | Fragile XA, fragile XE, fragile XF, Jacobsen (FRA11B) | DR | 30, 31 |
| Expanded (GGG-CCG)\_n | Friedreich’s ataxia | DR with (R-Y) | 30, 31 |
| Expanded (GGG-CCG)\_n | Pseudoachondroplasia, multiple epiphyseal dysplasia | DR | 31 |
| Expanded (GGG-CCG)\_n | Myotonic dystrophy type 2 | DR | 31 |
| Expanded (GGG-CCG)\_n | Progressive myoclonus epilepsy type 1 | DR | 30, 31 |
| Expanded (GGG-CCG)\_n | Creutzfeldt-Jacob | DR | 50 |
| Expanded (GGG-CCG)\_n | Fociascapulohumeral muscular dystrophy | DR | 51 |
| Expanded (GGG-CCG)\_n | IgH isotype switching | IR, G-rich | 37-41 |
| Expanded (GGG-CCG)\_n | t(X;22)(q12;q11) | Scott’s syndrome | IR | 44 |
| Expanded (GGG-CCG)\_n | Expanded polyglutamine diseases | DR | 45 |
| Expanded (GGG-CCG)\_n | Expanded (TCTG)\_n | DR | 46 |
| Expanded (GGG-CCG)\_n | Expanded (CAG-TCT)n | DR | 47 |
| Expanded (GGG-CCG)\_n | Expanded (GAA-TTC)n | DR | 48 |
| Expanded (GGG-CCG)\_n | Expanded (GAC-GTC)n | DR | 49 |
| Expanded (GGG-CCG)\_n | Expanded (CCTG-CAG)n | DR | 50 |
| Expanded (GGG-CCG)\_n | Expanded (ATTTCT-AAGAT)n | DR | 51 |
| Expanded (GGG-CCG)\_n | Expanded (CCCCCGCCCGCG)n | DR | 52 |
| Expanded (GGG-CCG)\_n | Expanded (CTG-CAG)n | DR | 53 |
| Expanded (GGG-CCG)\_n | Expanded (CAG-TCT)n | DR | 54 |
| Expanded (GGG-CCG)\_n | Expanded (ATTTCT-AAGAT)n | DR | 55 |
| Expanded (GGG-CCG)\_n | Expanded (CCTG-CAG)n | DR | 56 |
| Contracted 3.3-kb D4Z4 repeat | Class switch recombination | IgH isotype switching | IR | 57 |
| (coding) | (coding) | (coding) | (coding) |

Triplet Repeat Diseases

At least 20 hereditary neurological diseases (Table I) are caused by the expansion of simple triplet repeat sequences in either coding or non-coding regions. These expansions are mediated by DNA replication, repair, and recombination (30–33). DNA slipped structures (Fig. 1) play a prominent role in all of these mechanisms, and the preferential formation of hairpin loops with long repeating CTG, GGC, and GAA strands, compared with their complementary strands, is critical to the expansion and deletion behaviors (genetic instabilities). Furthermore, the flexible and writhed conformations (34) adopted by long CTG-CAG and CCG-CCG tracts promote these behaviors.

Friedreich’s ataxia (Table I) is the only hereditary neurological disease caused by an expanded GAA-TTC sequence. This tract, which occurs in intron 1, is well known to adopt the triplex conformation (Fig. 1) as well as the more complex sticky DNA structure between two long GAA-TTC tracts at distant locations of the same molecule (35, 36). All workers in this field agree (30) that the triplex and/or sticky DNA conformation adopted by this repeating sequence is involved in the inhibition of transcription of the Friedreich’s ataxia gene to give a reduction in the amount of the frataxin protein. Accordingly, the in vitro discovery of sticky DNA along with its molecular biological implications may be one of the more convincing cases of a non-B DNA structure involved in disease pathology.

Class Switch Recombination and Somatic Hypermutation

B lymphocytes stimulated by antigens undergo two main genetic alterations: class switch recombination (CSR), which deletes varying portions of the CH gene cluster, and somatic hypermutation (SHM), which generates point mutations in the variable regions of both the L and H chains (37). Both processes require transcription through the target sequences and depend on the function of the activation-induced cytidine deaminase (AID), a single-stranded DNA-specific deC deaminase that converts cytidine to uracyl (38).
Non-B DNA Conformations

The switch (S) regions are composed of 1–10-kb-long sequences 5’ to each Cγ gene; they are highly G-rich on the non-template strand and abundant in tandem repeats with IR symmetry (37). In vitro, the S regions adopt B-loop structures, whereby the transcripts form stable RNA-DNA hybrids leaving the G-rich non-template strand single-stranded (39, 40). Synthetic G-rich templates deprived of inverted repeats (39) or G-rich templates containing other types (AT-rich) of inverted repeats (41) support CSR. These results (37–41) suggest that transcription-induced extrusion of the non-template strand stabilizes non-B DNA conformations (G4 tetraplexes and/or cruciforms), which expose unpaired bases that target AID activity to generate GU mismatches. The resulting DU-DNA is then processed by uracil DNA glycosylase or AP endonuclease to generate double strand breaks which, when occurring on two separate S regions, serve as CSR intermediates.

In SHM, the AID-dependent nicking activity at or near the transcription-stabilized stem-loop conformations is proposed to lead to hypermutation through the stimulation of error-prone DNA synthesis (37). Interestingly, the transcription- and lymphocyte-dependent SHM has been activated in fibroblasts in culture upon infection with AID-producing retroviruses (42), suggesting that the enzymes leading to enhanced mutagenesis are normally present in non-lymphoid cells. In addition, computational analyses of these AID-induced mutations (43) revealed a strong correlation between the “mutability” of the nucleotide and the likelihood of being single-stranded as a result of stem-loop formation during transcription. Similar studies on single point mutations of the p53 gene associated with cancer (44) support the overall conclusion that transcription alone is sufficient to enhance the background mutation rate through stem-loop formation, a conformational feature frequently found at sites of small deletions (45).

Future Challenges

Dramatic advances were realized from the combined efforts of biophysics, human genetics, genomics, and molecular biological studies in our comprehension of the role of non-B DNA conformations in mutagenesis involved in genetic diseases. The thesis of this minireview is that, in general, all segments of DNA are usually in the orthodox right-handed B-conformation; however, certain repeat tracts occasionally adopt a non-B structure as a function of supercoiled density or protein binding, etc. These non-B conformations (cruciforms, triplehex, tetraplexes, etc.) are recognized by DNA repair systems and serve as end points for several types of genome rearrangements, which are the genetic basis for a number of diseases. Future emphases will attempt to fine-tune the adoption of the non-B conformations at specific loci to correlate more rigorously this behavior with the generation of rearrangement end points. In addition, analyses may be performed to identify the types of non-B DNA structures that elicit certain kinds of mutations and the enzyme systems involved. Many more diseases will be recognized as due to mutations at non-B DNA structures. Also, strategies will be focused toward the development of therapeutics to ameliorate the devastating consequences of these syndromes.
