Triad-DNA: a model for trinucleotide repeats

Sir — Trinucleotides reiterated many times within and outside coding regions are genetic manifestations associated with more than 10 human diseases[1,2], and the repertoire of triplets is growing[3,4]. The predominant sequences are GC-rich with repeating units consisting primarily of $(CGG)_n$ (more precisely $(GGC)_n$ (ref. 4) and $(CAG)_n$. Correlated flanking sequences are a common accompanying feature, as for example in Huntington's disease, in which a $(CCG)_n$ repeat occurs 12 nucleotides downstream of $(CAG)_n$ (ref. 5). A CpG island is present in the vicinity of the large $(GGC)_n$ repeats[6]; there is a strong correlation between triplet repeat amplification, hypermethylation (of the CpG island and possibly the repeat[2]), and fragile site expression[2]. In the case of loci corresponding to 5'- or 3'-untranslated genetic regions, the pathological consequences of the repeated sequences appear to arise from changes in transcription (levels, compositions, and lengths of the corresponding transcripts). Trinucleotide expansion of open reading frames affects translation and function of the peptides expressed (efficiency of translation, loss or gain of function). Other as yet undefined *cis*- and *trans*-acting features and factors may be responsible for the neurodegenerative diseases that are characteristic of this category of dynamic mutations.

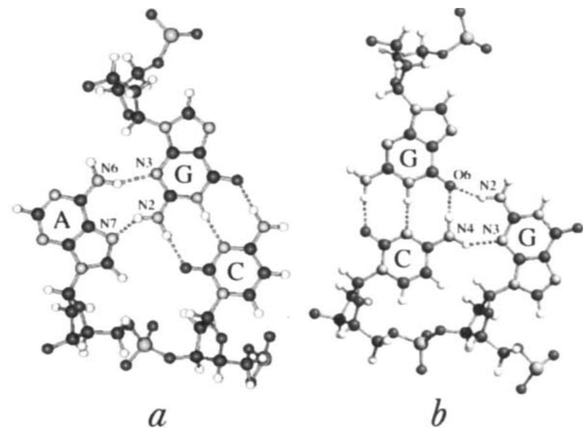

Fig. 1 Base triads. *a*, CA•G; *b*, GC•G. The central dot separates the contributions from the two strands.

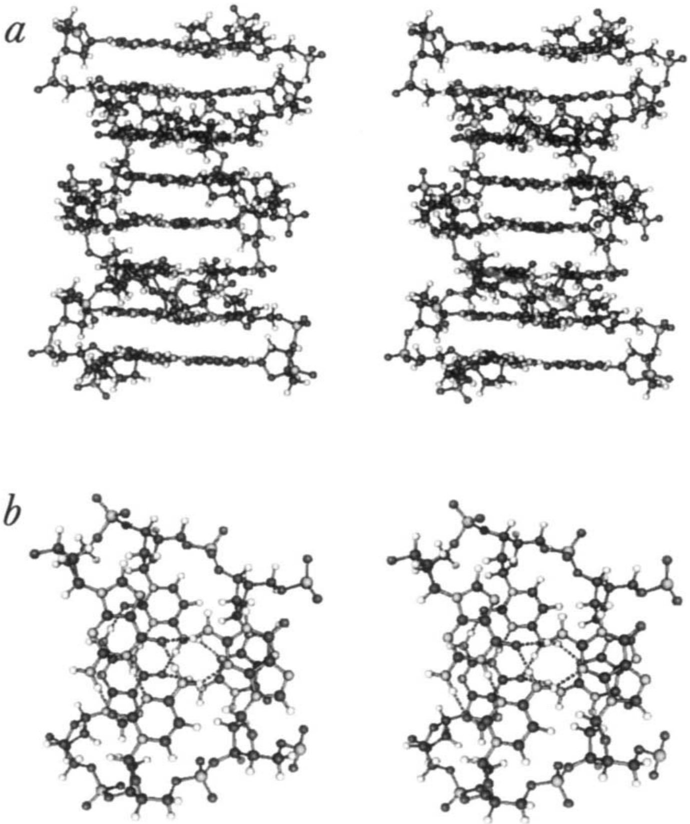

Fig. 2 Stereo representations of $(GC{\bullet}G)_8$ triad-DNA: *a*, side projection; *b*, top view of two successive GC•G triads.

Certain genetic and biochemical mechanisms have been proposed for explaining the origin, inheritance patterns, and associated pathogenicity of the trinucleotide repeats. Female meiotic transmission is presumably responsible for expansion of premutations to full mutations, accounting for the observation that daughters of normal transmitting males do not express the fragile X phenotype, whereas somatic mutation could lead to the mosaicism within the triplet repeats[7]. Schemes of DNA replication involving slippage structures[8] ('slippery' DNA (ref. 9)) provide a dynamic model[10] for reiterating the sequences. Tetraplex structures, stabilized by methylation and capable of inhibiting transcription, replication, and chromatin condensation, have been proposed for the $(GGC)_n$ tracts[11], and preferential nucleosomal positioning has been suggested as a mechanism for transcriptional inhibition in $(CAG)_n$ repeats[12]. However, as recently bemoaned by the editor of *Nature*[13], there have been no proposals based on specific molecular structures, that is, involving (un)usual DNA or RNA models, for rationalizing the key features of dynamic mutations: (1) the trinucleotide motif, (2) the length dependence of the pathogenicity, (3) the involvement of both coding and non-coding sequences, (4) the interrelation of neighbouring sequences, and (5) amplification with retention of the triplet unit.

We introduce here a novel DNA secondary structure, potentially also expressed in RNA, that can account for at least some of the features of the trinucleotide repeat syndromes. This structure, designated triad-DNA, consists of an antiparallel double helix of base triads instead of the basepairs of conventional B-DNA. The triads involve both hydrogen-bonded and covalent interactions between a Watson–Crick (WC) G•C nucleotide pair and a third nucleotide (Fig. 1*a*,*b*). In the CA•G triad (Fig. 1*a*) the third base (A) interacts with the minor groove side of a WC G•C basepair, forming a G•A basepair known to be very stable in tandem[14]. In the GC•G triad (Fig. 1*b*), the third base (G) interacts with the major groove side of the WC G•C

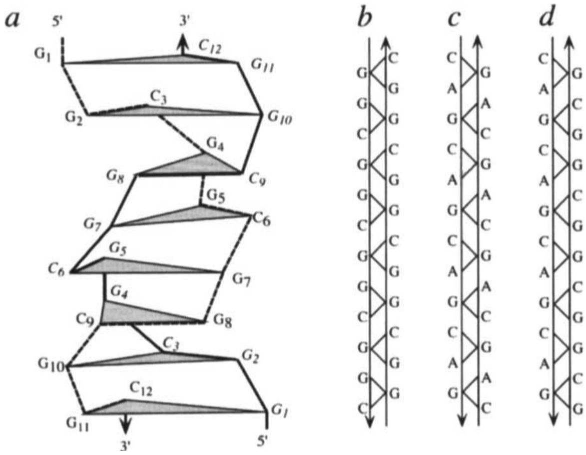

Fig. 3 Schematic representations of triad-DNA. *a*, Three-dimensional (3-D) contour of the $(GC{\bullet}G)_8$ helix shown in Fig. 2*a*. The base triads are shown as shaded triangles and the two backbones as solid or dashed lines. *b–d*, 2-D schemes of triad-DNAs formed from three combinations of $(CA{\bullet}G)_8$ and $(GC{\bullet}G)_8$ strands. The triangles between the straight lines (backbones) denote the base triads.

basepair, in analogy to the familiar triple-stranded helices[15]. The double-stranded configurations with base triads we have found so far consist of the same set of nucleotides observed in the trinucleotide repeat syndromes.

The triads can be connected in a symmetrical fashion so as to generate a double helix with a 2:1/2:1 repeat, in which two bases in one triplet are connected to a base from an adjacent triplet (Figs 2 and 3*a*). The backbone progression consists of a step and two rises, accounting for a net helical twist of 60° (Fig. 3 and Table 1). As a consequence, the helical repeat is 6 base triads (18 nucleotides) instead of the 10 basepairs (20 nucleotides) of B-DNA (Table 1). Using the same helical rise of 3.4 Å triad-DNA is more compact than B-DNA by a factor of 1/3. The interstrand basepairs connected to successive rise points are different. In the homoduplex GGC or CAG structures (Fig. 3*b*,*c*) one rise corresponds to a WC geometry, whereas the second rise consists of successive G–G or A–G pairs, respectively. That is, the triad structural motif is also reflected in a trinucleotide repeating unit along the helix. Triad-DNA can also be formed between $(GGC)_n$ and $(CAG)_n$ strands (Fig. 3*d*); in this case CA·G and GC•G triads alternate.

The helical configuration of triad-DNA is best perceived in the stereo representations of a GC•G homoduplex shown in Fig. 2. The stacking between the two non-WC Gs of successive triads (Fig. 2*b*) is interstrand and even more pronounced than between the Gs and Cs involved in the WC basepairs, thereby providing the potential for strong stabilization of triad-DNA. The uniform base triads constitute a second potential advantage of triad-DNA as a model for the trinucleotide repeat sequences. Alternative double-stranded basepaired structures necessarily incorporate mismatch basepairs (G•G, A•A or A•G), which disturb the stacking interactions and regularity of the backbone. We have performed molecular-mechanics calculations demonstrating that GC•G triad-DNA (Fig. 2*a*) is indeed more stable than duplexes including G•G mismatches in either of two known geometries[16,17].

In the context of the trinucleotide repeat syndromes, we note that triad-DNA could provide a stable self-structured conformation for the extruded strand in slippage-based mechanisms of replication, particularly if topological relaxation and compaction of chromatin are involved and required for stabilization of intermediates in the pathological expansion of such sequences. The complementary pyrimidine-rich strand would be unpaired and thus could serve as a template while maintaining the trinucleotide repeat motif. Formation of the triad-DNA structure might also result in both retardation of replication, and transcription regulation in the regions with expanded trinucleotides, known to occur in the fragile X chromosome. A triad RNA structure might influence processing and translation efficiency of corresponding transcripts. The $(CAG)_n$ repeat of Huntington's disease can be combined in a single triad-DNA structure (Fig. 3*d*) with a nearby $(GGC)_n$ segment from the other strand of the original genomic DNA, providing a possible rationale for the cited functional correlation of the flanking sequence. One can also speculate that the fundamental similarity (Figs 2 and 3) but sequence-dependent distinctions between different triad-DNA constructs explored to date could be reflected in the polymorphism (variations in threshold copy number) exhibited in patients afflicted with the trinucleotide repeat diseases.

The existence of triad-DNA remains to be demonstrated by the biochemical and biophysical techniques applied in structural studies of nucleic acids. High resolution nuclear magnetic resonance (NMR) should serve to detect the short intra- and interstrand distances distinctive for triad-DNA. One would expect strong intrastrand sequential nuclear Overhauser effects (NOEs) between C1'H and C2'H of the deoxyguanosine residue and C5H of deoxycytidine, respectively, for the GC•G triad-DNA(Fig. 1*b*), and a strong NOE between two interstrand C1'H protons of deoxyadenosine residues for the CA•G triad-DNA. Features of the secondary structure can also be probed by circular dichroism (CD), absorption, and vibrational spectroscopy, by chemical modification, and by systematic base substitutions. The compaction afforded by the triad-DNA conformation would be compatible with existing electrophoretic data for non-methylated triplet repeats, although the published interpretations have invoked other structural forms such as G4-like quadruplexes[11]. The latter are difficult to reconcile with the fact that the generating triplet sequences reside in

Table 1 Comparison of triad-DNA and B-DNA

ds DNA	Unit of secondary structure	Helical rise (Å)	Helical repeat (nucleotides/ turn)	Helical twist (deg)	Backbone progression
B-DNA	Basepair	3.4	10 (20)	36	Rise rise
Triad-DNA	Base triad	3.4	6 (18)	60	Step-rise rise

a single strand of the genomic DNA and with a parallel orientation of the four segments constituting the tetraplex. Furthermore, a quadruplex offers no obvious rationale for the ubiquitous triplet nature of the repeating sequences involved in dynamic mutations. Finally, as in the case of all proposed alternative DNA conformations, one must search for specific proteins recognizing triad-DNA.

Vitaly V. Kuryavyi
Thomas M. Jovin
Department of Molecular Biology, Max Planck Institute for Biophysical Chemistry, P.O. Box 2841 D-37018 Göttingen, Germany

1. Richards, R.I. & Sutherland, G.R. *Nature Genet.* **6**, 114–116 (1994).
2. Willems, P.J. *Nature Genet.* **8**, 213–215 (1994).
3. Lindblad, K. *et al. Nature Genet.* **7**, 124 (1994).
4. Han, J. *et al. Nucl. Acids Res.* **22**, 1735–1740 (1994).
5. Andrew, S.E. *et al. Hum. molec. Genet.* **3**, 65–67 (1994).
6. Hansen, R.S. *et al. Hum. molec. Genet.* **1**, 571–578 (1992).
7. Oberle, I., *et al. Science* **252**, 1097–1102 (1991).
8. Strand, M. *et al. Nature* **365**, 274–276 (1993).
9. Kunkel, T.A. *Nature* **365**, 207-208 (1993).
10. Richards, R. & Sutherland, G.R. *Cell* **70**, 709–712 (1992).
11. Fry, M. & Loeb, L.A. *Proc. natn. Acad. Sci. U.S.A.* **91**, 4950–4954 (1994).
12. Wang, Y.-H. *et al. Science* **265**, 669–671 (1994).
13. Maddox, J. *Nature* **368**, 685 (1994).
14. Li, Y., Zon, G. & Wilson, W.D. *Proc. natn. Acad. Sci. U.S.A.* **88**, 26–30 (1991).
15. Rao, B.J. *et al. J. molec. Biol.* **229**, 328–343 (1993).
16. Cognet, J.A.H. *et al. Nucl. Acids Res.* **19**, 6771–6779 (1991).
17. Skelly, J.V. *et al. Proc. natn. Acad. Sci. U.S.A.* **90**, 804–808 (1993).

Towards fully automated genome-wide polymorphism screening

Sir—Large-scale screening for known polymorphisms will require techniques with a minimal number of steps and the ability to automate each one. In this regard, the 5' nuclease PCR assay first described by Holland *et al.*[1,2] and refined by Lee *et al.*[3] is especially attractive because it virtually eliminates post-PCR processing. For this assay, a fluorogenic probe, labelled with both a fluorescent reporter dye and a quencher dye, is included in a typical PCR amplification. During PCR, this probe is cleaved by the 5' nuclease activity of *Taq* polymerase[1,4], only if it hybridizes to the segment being amplified. Cleavage of the probe generates an increase in the fluorescence intensity of the reporter dye. Thus, amplification of a specific product is detected by simply measuring fluorescence after PCR. Furthermore, by using different reporter dyes, cleavage of multiple probes can be detected in a single PCR. Lee *et al.*[3] used probes labelled with the reporter dyes FAM and TET to distinguish the *ΔF508* and normal alleles of the human cystic fibrosis gene.

The –23 A/T diallelic polymorphism of the human insulin gene (*INS*) is associated with susceptibility to type 1 diabetes[5]. A is the type 1 diabetes-associated or '+' allele; B or '–' is the non-associated allele. Here we report the successful application of the 5' nuclease PCR assay to discriminate these two *INS* alleles, which differ by a single A–T base substitution. This is noteworthy for three reasons. First, we used a simplified and general method of probe design. Second, these probes retain enough specificity to distinguish a single base difference. Third, our standard method of analysis enables automated genotype determination.

Lee *et al.*[3] used probes with the reporter dyes on the 5' end and the quencher dye TAMRA on the seventh nucleotide from the 5' end. There were concerns that restrictions on the placement and spacing of dyes would limit the applicability of this type of probe system. The design of fluorogenic probes has been greatly simplified by the recent discovery that probes with a reporter dye on the 5' end and a quencher dye on the 3' end exhibit adequate quenching for the probe to perform in the 5' nuclease PCR assay[6]. Thus, placement of the reporter and quencher dyes puts no constraints on the selection of a probe sequence, enabling application of this technology to any PCR system.

We included probes specific for alleles A or B in reactions amplifying a segment of *INS* (Fig. 1). After PCR, fluorescence intensities at three wavelengths were measured in a 96-well format. The fluorescence plot (Fig. 1*a*) shows four distinct clusters of data points. The no template controls have the lowest fluorescence intensities for both TET and FAM because the emissions of these dyes are quenched in the intact probes. Based on the relative fluorescence intensities, the other three clusters can be identified as AA, AB, and BB individuals. The separation of the clusters can be enhanced by applying multicomponent analysis (Fig. 1). By normalizing for the extent of the reaction, the tight clusters of data points shown in Fig. 1*b* are obtained. Based on this plot, the individuals are clearly identified as AA, AB, or BB. The genotypes of the 18 individuals analysed in Fig. 1 are completely concordant with the genotypes determined previously using the PCR-*Hph*I digestion assay[5]. The typing of 57 additional individuals also produced tight clusters of normalized A and B values, permitting unambiguous determination of AA, AB, and BB genotypes (data not shown).

The principles of probe design and data analysis we have used for *INS* typing are generally applicable to any polymorphic site that can be amplified by PCR. The remarkable discrimination between genotypes (Fig. 1*b*) makes it easy to designate limits so that genotypes can be called automatically. We have written a spreadsheet macro that reads the fluorescence data from the spectrometer, performs the analyses, indicates reactions where specific amplification did not occur, and automatically makes genotype determinations. This macro can be used on fluorescence data from any PCR system as long as single-template and no template components are designated.

The hallmark of the 5' nuclease PCR assay is increased throughput. Reactions differ from typical amplification reactions only by the addition of fluorogenic probes. Unlike methods such as the oligonucleotide ligation assay (OLA) or primer-guided nucleotide incorporation assays[7], there are no post-PCR processing steps except