Features in Short Guanine-Rich Sequences That Stimulate DNA Polymerization in Vitro[†]

Mitsu S. Reddy and Susan H. Hardin*

Department of Biology and Biochemistry, Institute of Molecular Biology, University of Houston, Houston, Texas 77204-5001

Received May 24, 2002; Revised Manuscript Received November 4, 2002

ABSTRACT: We have discovered that short guanine-rich oligonucleotides are able to self-associate into higher order structures that stimulate DNA synthesis in vitro without the addition of a conventional template [Ying, J., Bradley, R. K., Jones, L. B., Reddy, M. S., Colbert, D. T., Smalley, R. E., and Hardin, S. H. (1999) *Biochemistry 38*, 16461−16468]. Our initial analysis indicated the importance of the presence of three contiguous guanines (G) in an oligonucleotide that stimulates DNA polymerization. To gain insight into and to refine sequence requirements for the unexpected DNA synthesis, we analyzed a 231-member guanine-rich octamer library in a fluorescent nucleotide polymerization assay. We observe that, in addition to three contiguous Gs, the presence of a secondary G cluster within the octamer is essential. Furthermore, the location of the primary G cluster in the center of the molecule is most stimulatory. The majority of the octamers that form extended DNA products have a single non-G base separating the primary and secondary G clusters, the identity of which is predominantly thymine (T). Further, a T 5′ or 3′ of the primary G cluster positively influences the stimulatory function of the oligonucleotide. Overall, the occurrence of bases in the octamer is in the descending order of G > T > A > C. Our studies demonstrate that structures stabilized by noncanonical base pairings are recognized by a DNA polymerase in vitro, and these findings may have relevance within the cell. In particular, the features of these G-rich stimulatory sequences show striking similarities to telomeric sequences that form diverse G-quartet structures in vitro.

Our laboratory discovered that G-rich oligonucleotides (oligos) are able to self-associate into higher order structures that stimulate several thermostable DNA polymerases to synthesize extended DNA strands in vitro in the absence of a conventional DNA template (*1*). Analyses using base analogues demonstrated that the Gs are involved in forming structures stabilized by non-Watson−Crick base pairing, consistent with G·G Hoogsteen base pairing, and that these alternative base pairings are essential for the observed DNA polymerization, referred to as *h*igh *i*ntensity *d*ata [HID (*1*) and Figure 1]. Gel mobility shift assays provided additional evidence for the formation of higher order DNA structures in our polymerization assay conditions. Using scanning force microscopy, we demonstrated that the *Tetrahymena* telomeric sequence (Tet1.5, $G_4T_2G_4$) forms extended DNA structures in our reaction conditions [similar to G-wires (*2*)] and that these structures associate with the polymerase (*1*).

In this report, we present an analysis of the sequence features present in G-rich, eight-base oligos (octamers) that facilitate recognition by *Taq* DNA polymerase and stimulate DNA synthesis in vitro. To refine the sequence requirements of the G-rich stimulatory structures, we selected 198 oligos containing a minimum of three contiguous guanine bases (based upon our previous analysis) from a 1000-member, 75% GC-rich octamer library (*1*, *3*). Additionally, another 33 octamers were obtained to complete the analysis (231 oligos; Tables 1 and 2), since the oligos in certain classes were underrepresented in the 1000-member library [due to the original criteria used in designing the library for octamer-primed sequencing technology (*3*)].

Our previous study demonstrated the importance of Hoogsteen base pairing in structure formation (*1*). Hoogsteen interactions can stabilize various types of DNA associations: parallel-stranded duplexes, triplex, and G-quartets. Although DNA can adopt diverse and interesting structures, the data are most consistent with a G-quartet-stabilized DNA structure serving as the replication template.

The diverse structures formed by guanine-rich (G-rich) DNA sequences have been extensively investigated in recent years, and there is growing evidence for a biological significance for such structures. Sequences from the telomeres of *Tetrahymena*, *Oxytricha*, *Saccharomyces*, *Drosophila*, and humans (*4*−*11*), immunoglobin switch regions (*12*), gene promoter regions (*13*), and the fragile X region (*14*) form a variety of G-quadruplex structures in vitro. These structures can be formed from one, two, or four individual DNA strands (*9*, *10*, *12*, *15*, *16*). They are stabilized by eight cyclic G·G hydrogen bonds [Hoogsteen base pairing (*17*)] between four Gs in a coplanar arrangement (guanine quartet; Figure 2A).

G-rich sequences can form stable intermolecular parallel or antiparallel four-stranded structures (G4) (*12*, *18*), unimolecular (G4′) or bimolecular (G′2) antiparallel fold-back hairpin structures (*8*, *10*, *11*; Figure 2B). In general, the type of structure formed and its stability depend on several factors including DNA sequence, concentration, length, sequence

[†] This research was supported by the Welch Foundation (Grant E-1349).

* Corresponding author. Tel: 713-743-2686. Fax: 713-743-2636. E-mail: shardin@uh.edu.

10.1021/bi020380w CCC: $25.00
Published on Web 12/17/2002

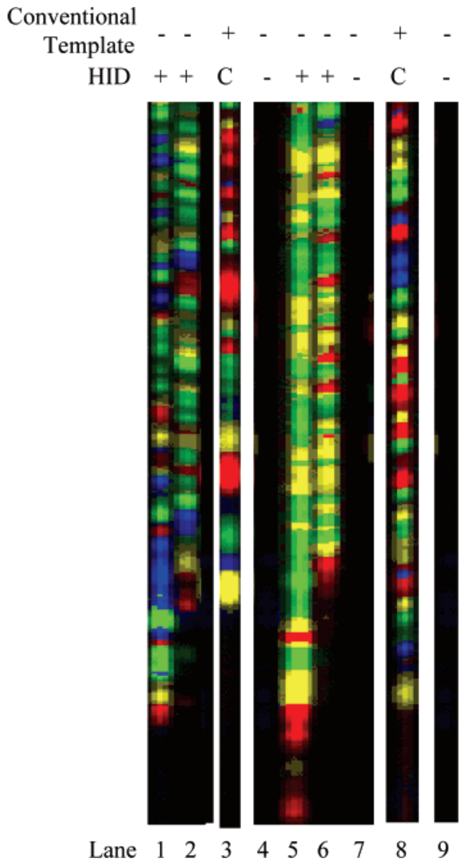


FIGURE 1: Gel picture of the products of the fluorescent nucleotide polymerization assay. Noncompressed view of a gel showing a sample of seven octamers tested for production of HID (without the addition of a conventional template). Lanes: 1, 384.214 (HID producing); 2, 384.352 (HID producing); 3, control (standard sequencing reaction); 4, 384.353 (non-HID producing); 5, 384.351 (HID producing); 6, 384.347 (HID producing); 7, 384.350 (non-HID producing); 8, control (standard sequencing reaction); 9, 384.336 (non-HID producing). The numbers 384 followed by a suffix refer to primer name/number.

context, number of contiguous G residues, number of G repeats per strand, identity of terminal residues, number of intervening hydrophilic residues, and nature and concentration of ions present (*9*, *19*−*21*). In this report analyses of the G-rich octamers provide clues about the stimulatory structures responsible for the unexpected DNA synthesis.

MATERIALS AND METHODS

Oligonucleotide Synthesis. DNA oligos were synthesized by Genosys Biotechnologies, Inc. (183 oligos with the prefix "384", 13 nonlibrary oligos, TAsub, and Perm2), Genemed Synthesis, Inc. (17 nonlibrary PS, SP, and Rev permutations of selected "384" library oligos), and MWG Biotech, Inc. (18 oligos with the prefix SH; Table 1). All oligos were desalted and purified following standard methods by their respective manufacturers.

Fluorescent Nucleotide Polymerization Assay/Octamer-Primed Automated DNA Sequencing. Fluorescent nucleotide polymerization/octamer-primed automated DNA sequencing reactions were performed with minor modifications to the published procedure (*22*). In brief, between 1 and 50 pmol of octamer primer, 2 μL of ABI PRISM dye-terminator cycle sequencing ready reaction rhodamine kit premix [containing Amplitaq FS, deoxynucleotides (dITP, dATP, dTTP, and dCTP), and fluorescently tagged dideoxynucleotides], 1.5 μL of 5× sequencing reaction buffer [80 mM Tris (pH 9), 2 mM $MgCl_2$], and the appropriate volume of water were added to assemble a 10 μL reaction (note that a conventional DNA template was not added into the reaction). The reaction mix was initially denatured at 96 °C for 2 min and cycled 99 times at 96 °C for 10 s (denaturing), 40 °C for 1 min (annealing), and 60 °C for 4 min [extension (*1*)]. Automated sequencing reaction controls were performed as per the manufacturer's instructions (Applied Biosystems). All samples were precipitated, resuspended in 3 μL of loading dye (5:1 formamide:blue dextran, consisting of 50 mg/mL blue dextran and 25 mM EDTA, pH 8.0), heated for 2 min at 96 °C, and loaded on a 5.28%, 36 cm Long Ranger sequencing gel. An ABI PRISM 377 DNA sequencer collected the data in 2× or 4× mode.

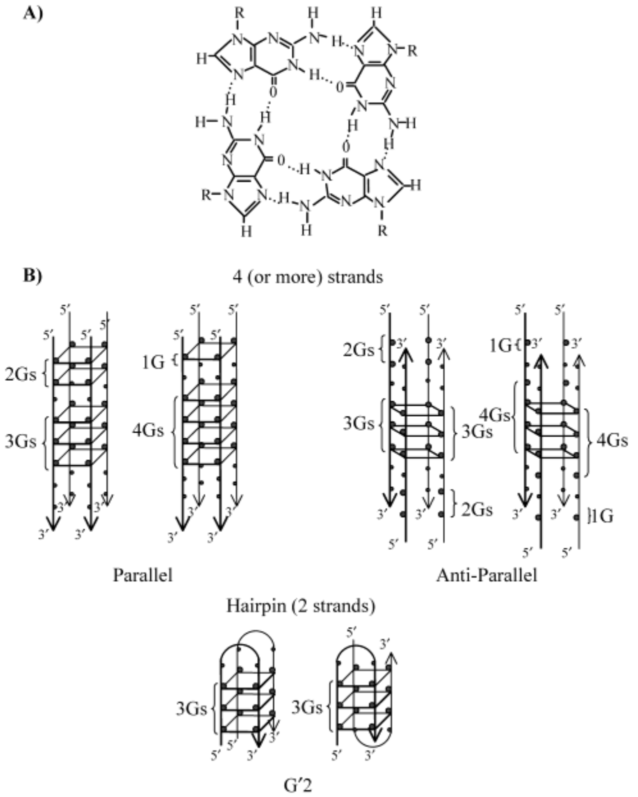


FIGURE 2: Guanine quartet and selected G-quadruplex structures. (A) Four guanines arranged in a single plane and hydrogen bonded in a cyclic manner, forming the building block of G-quadruplex structures. (B) Schematic drawing of selected intermolecular and intramolecular guanine quartet stabilized structures. Larger circles represent guanines, smaller circles represent non-guanine bases, and arrows represent strand directionality.

231-Member G-Rich Library Classification. We have classified the 231-member G-rich octamer library into three major classes on the basis of the length of the primary G cluster (P) present in the oligos (Figure 3A). Type I includes all oligos with three contiguous Gs (n = 166), type II includes all oligos with four contiguous Gs (n = 59), and type III includes all oligos with five contiguous Gs (n = 6). Each type is further classified into subclasses on the basis of the length of the secondary G cluster (S) in the oligo (Figure 3A). Type Ia includes all oligos with two separate clusters of three contiguous Gs, type Ib includes all oligos with two additional contiguous Gs, and type Ic includes all oligos with two additional, noncontiguous Gs. One or more

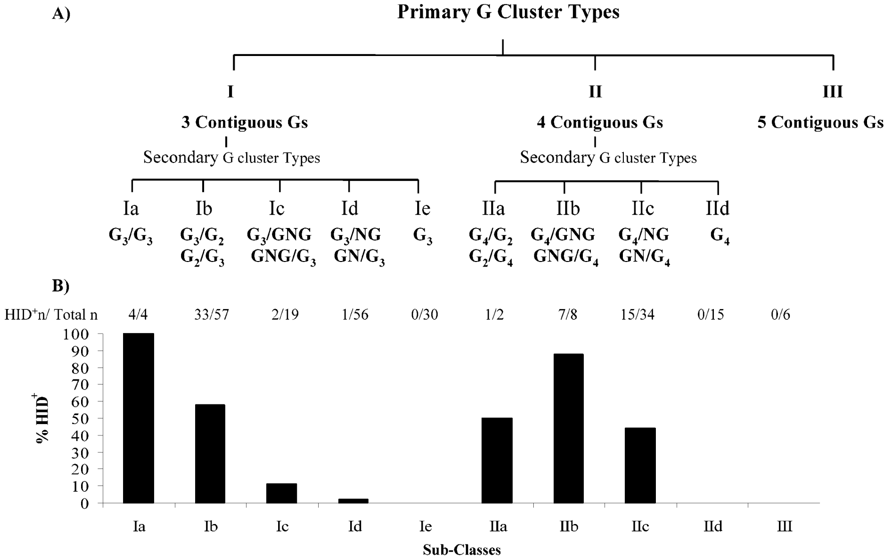


FIGURE 3: G-rich octamer library classification. (A) Classification of octamers based on type of primary (P) and secondary (S) G clusters observed in the oligonucleotides. (B) Percent HID-producing octamers within each subclass. Types: Ia, G_3/G_3 and G_3/G_3; Ib, G_3/G_2 and G_2/G_3; Ic, $G_3/\underline{GNG}$; Id, $G_3/\underline{NG}$; Ie, G_3; IIa, G_4/G_2 and G_2/G_4; IIb, $G_4/\underline{GNG}$ and $\underline{GNG}/G_4$; IIc, $G_4/\underline{NG}$ and $\underline{GN}$/G4; IId, G_4; III, G_5. The value above each bar is the number of HID-producing octamers over the total number of octamers in the subclass.

intervening non-G bases, i.e., G_3NGNG or G_3NNGNG, separate the G clusters in type Ic, and unless specified, the order of the primary and the secondary G clusters can vary within a category. Type Id includes all oligos with a single additional G (note that, for consistency in terminology, we refer to a single G as a secondary cluster), and type Ie includes all oligos lacking any additional G. Similarly, type II is further classified into four subclasses on the basis of the length of the secondary G cluster present in the oligos (Figure 3A). All HID- and non-HID-producing octamers of types Ia, Ib, IIa, IIb, and IIc were assayed for their ability to stimulate the nonconventional polymerization activity at least twice.

Data Analysis. All octamers tested were tabulated as HID producing or non-HID producing (Tables 1 and 2, respectively). They were further grouped on the basis of distinguishing features present in the oligos that impact the HID result. The total number of octamers in each category was considered, and the percentage of HID-producing octamers is graphically depicted for each category (Figures 3−8).

Preparation of Reaction Products for Cloning. Fifty microliter polymerase chain reactions were performed with 25 pmol of primer $(CGGG)_3$ or $G_4T_2G_4$, 200 μM each dNTP, 1× PCR buffer [20 mM Tris-HCl (pH 8.8), 2 mM $MgSO_4$, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 0.1% Triton X-100, and 0.1 mg of nuclease-free BSA], and 5 units of *Pfu* DNA polymerase (Stratagene, Inc.). Note that a conventional template is not added to the reaction. The cycling regimen is as follows: initial denaturation at 95 °C for 2 min, cycling 50 times between 95 °C for 45 s, 40 °C for 1 min, and 60 °C for 4 min, and a final incubation at 60 °C for 10 min.

Cloning and Sequencing of Products Produced by G-Rich Oligos. The amplified products were electrophoresed through a 1% agarose gel, stained with ethidium bromide, and visualized using a Kodak EDAS 120 imaging system. DNA in the 0.5−1.5 kb size range from each of the amplifications was excised and isolated in 1× TAE buffer by electroelution, extracted with phenol−chloroform−isoamyl alcohol (25:24:1), and precipitated. The purified DNA samples were A-tailed and used in ligation reactions with pGEM T-easy cloning vector as per the manufacturer's protocols (Promega). The ligation products were transformed into *Escherichia coli* JM109, and the cells were plated on Luria−Bertani (LB) plates containing ampicillin (100 μg/mL), IPTG (0.5 mM), and X-gal (80 μg/mL). White and faint blue colonies were selected and grown in 3 mL of LB broth, and plasmid DNA was isolated using the PERFECT prep DNA isolation kit (5′ to 3′, Inc.). DNA from each sample was analyzed for insert presence by *Eco*RI restriction digestion. Clones containing an insert were sequenced using T7 and SP6 vector primers. Automated DNA sequencing reactions were performed as per the manufacturer's instructions (Applied Biosystems, ABI Prism), precipitated, resuspended in 3 μL of loading dye, heated for 2 min at 96 °C, and loaded on a 5.28%, 36 cm Long Ranger sequencing gel. An ABI PRISM 377 DNA sequencer collected the data in 2× mode. All insert sequences were submitted to GenBank (accession numbers AY145505−AY145519) and analyzed via BLAST@NCBI during August 2002. Additionally, all insert sequences were analyzed for their GC contents, and a primer pattern search was performed using a 10 base sequence, $(CGGG)_2CG$ or $G_4T_2G_4$, with a threshold of 70% identity.

RESULTS

Previously, we proposed that G-richness and the presence of three contiguous Gs in an octamer are important requirements for stimulation of DNA synthesis in vitro (*1*). To gain insight into the phenomenon of *h*igh *i*ntensity *d*ata (HID) production, a sample library of 231 G-rich octamers containing, minimally, three contiguous Gs was individually tested for their ability to stimulate DNA synthesis in vitro (Tables 1 and 2). Twenty-seven percent of the octamers ($n = 63/231$) are able to stimulate production of the high-intensity fluorescent signal without the addition of a conventional sequencing template (see Materials and Methods, Figure 1, and ref *1*). Upon analyses of these 231 octamers (Tables 1 and 2), we further refine the sequences that affect HID production.

Table 1: 63 HID-Producing Octamers Classified According to Type of Primary and Secondary Clusters Present[a]

no.	name/number	sequence	CL	CO	IB	IB-RY
		Type Ia G_3NG_3N				
1	Perm2	AGGGAGGG	3	SP and PS	1	R
2	384.388/Perm3	GGGAGGGA	3	SP and PS	1	R
3	384.754/184	GGGTGGGA	3	SP and PS	1	Y
4	6−176X	GGGATGGG	3	SP and PS	2	RY
		Type Ib NG_3NG_2N				
5	454PS	GGGAGGTG	3	PS	1	R
6	384.566/900/Perm1	GAGGGAGG	3	PS	1	R
7	SH12	CGGGAGGA	3	PS	1	R
8	SH13	GGGAGGTC	3	PS	1	R
9	SH15	GAGGGAGG	3	PS	1	R
10	389SP/Asub/849	GGAGGGAG	3	SP	1	R
11	384.331	GTGGAGGG	3	SP	1	R
12	390SP/Asub3″C	GGAGGGAC	3	SP	1	R
13	384.454/Atsub	GGAGGGTG	3	SP	1	R
14	SH1	TGGAGGGC	3	SP	1	R
15	SH2	GGAGGGCT	3	SP	1	R
16	SH4	GGATCGGG	3	SP	3	RYY
17	384.216	GGGTGGTC	3	PS	1	Y
18	384.341	GGGTGGCA	3	PS	1	Y
19	384.351	GTGGGTGG	3	PS	1	Y
20	384.559	GAGGGTGG	3	PS	1	Y
21	346PS	CAGGGTGG	3	PS	1	Y
22	TASUB PS	GGGTGGAG	3	PS	1	Y
23	749PS	GGGTGGTG	3	PS	1	Y
24	SH14	CTGGGTGG	3	PS	1	Y
25	SH16	GGGCGGTT	3	PS	1	Y
26	SH17	GGGTGGCT	3	PS	1	Y
27	SH18	GGGCGGTA	3	PS	1	Y
28	384.346	CAGGTGGG	3	SP	1	Y
29	384.557	GAGGTGGG	3	SP	1	Y
30	TAsub	GGTGGGAG	3	SP	1	Y
31	384.214	GGTGGGTC	3	SP	1	Y
32	384.749/Tsub	GGTGGGTG	3	SP	1	Y
33	SH5	GGCGGGAA	3	SP	1	Y
34	SH6	GGCGGGAT	3	SP	1	Y
35	SH3	GGTACGGG	3	SP	3	YRY
36	384.352/175	GGGCTTGG	3	PS	3	YYY
37	384.344	GGTCTGGG	3	SP	3	YYY
		Type Ic G_3NNGNG				
38	384.562	GAGGGCTG	3	SP and PS		
39	384.745	GGGTGCTG	3	SP and PS		
		Type Id G_3NG				
40	384.565	GAGGGCAC	3	SP	1	R
		Type IIa G_4NG_2				
41	384.347	GGGGATGG	4	PS	2	RY
		Type IIb G_4NGNG				
42	384.746	GGGGAGTG	4	PS	1	R
43	384.579	GAGAGGGG	4	SP	1	R
44	384.338	GTGAGGGG	4	SP	1	R
45	384.561	GAGGGGAG	4	SP and PS	1	R
46	384.260	GGGGTGAG	4	PS	1	Y
47	384.750	GGGGTGTG	4	PS	1	Y
48	384.551	GAGTGGGG	4	SP	1	Y
		Type IIc $G_4NGNN/G_4NNGN/G_4NNNG$				
49	6−50X	GAGGGGCT	4	SP	1	R
50	551PS	GAGGGGTG	4	SP and PS	1	RY
51	384.135	GATGGGGC	4	SP	2	RY
52	384.229	GGGGACAG	4	PS	3	RYR
53	384.219	GGGGTGTC	4	PS	1	Y
54	SH8	CTGGGGTG	4	PS	1	Y
55	SH10	ACGGGGTG	4	PS	1	Y
56	384.376	GTGGGGAC	4	SP	1	Y
57	219SP	GTGGGGTC	4	SP	1	Y
58	SH9	GCGGGGTA	4	SP	1	Y
59	SH7	GTGGGGTG	4	SP and PS	1	Y
60	384.360	GGGGCAGA	4	PS	2	YR
61	SH11	CGGGGTAG	4	PS	2	YR
62	384.137	GTTGGGGC	4	SP	2	YY
63	384.339	GTCTGGGG	4	SP	3	YYY

[a] The name or number and sequence of the octamers are indicated. The octamers are sorted by type, primary cluster length (CL), primary (P) versus secondary (S) cluster order (CO), intervening base separating the primary and secondary clusters (IB), and intervening base classification (IB-RY; R = purine; Y = pyrimidine).

Presence of Three Contiguous Gs in an Octamer and G-Richness Alone Are Insufficient To Produce Nonconventionally Templated DNA Extension Products. Analysis of the 231 octamers (Tables 1 and 2; Figure 3B) that contain at least three contiguous Gs in their sequences reveals that G-richness and the presence of three contiguous Gs are insufficient for HID production. For example, all oligos tested of the types Ie (containing three contiguous Gs), IId (containing four contiguous Gs), and III (containing five contiguous Gs) are unable to stimulate DNA polymerization. Additionally, types Ib and IIc, the two subclasses in which the majority of the HID-producing oligos occur, also have a high percentage of non-HID-producing oligos. Thus, additional sequence features impact HID production.

To refine the sequence features in an octamer that affect HID production, the following analyses focus on the subclasses Ib and IIc, since these subclasses have a larger sample size of both HID-producing and non-HID-producing categories. Analysis of these subclasses in greater detail was performed to reveal distinguishing features in these oligos that either enable or prevent HID production. All HID-producing oligos examined thus far have a secondary G cluster. Therefore, types Ie, IId, and III have been excluded from the following analyses. In addition, types Ia, Ic, Id, IIa, and IIb do not have significant representation in both categories (i.e., HID-producing versus non-HID-producing oligos), making identification of any additional features that influence the ability of an octamer to stimulate DNA polymerization difficult. Thus, these groups have also been excluded from further analysis.

Analyzed features include (i) position of the primary G cluster (i.e., occurrence of the primary cluster at the 5′ end or the 3′ end or in the center of the oligo), (ii) primary (P) versus secondary (S) G cluster order (i.e., PS/SP), (iii) spacing and classification of base(s) between the primary and secondary G clusters (i.e., $G_3\underline{N}G_2/G_3\underline{NN}G_2$ etc.), (iv) base identity 5′ and 3′ of the primary G cluster (i.e., $\underline{N}G_3\underline{N}$ or $\underline{N}G_4\underline{N}$), and (v) base identity at each position in the octamer. Each feature may contribute to the ability of an octamer to produce HID.

Features That Contribute to an Octamer's Ability To Stimulate Unexpected DNA Synthesis. (i) Position of the Primary G Cluster. It is well documented that the position of the G cluster in the context of the rest of the molecule influences the type of structure formed by G-rich sequences (*19*, *23*, *24*). Therefore, one of the criteria used to analyze the G-rich stimulatory sequences was the positioning of the primary cluster (P) in the context of the octamer and its influence on the outcome of HID production.

The G_3 primary cluster of the type Ib subclass can occupy six different positions within an octamer (Figure 4A). In this subclass, HID-producing octamers have the highest percentage occurrence of the primary cluster at positions 3 and 4 (Figure 4B), i.e., in the center of the oligo. Position 2 is eliminated from consideration, due to its small sample size ($n = 1$). Among the non-HID-producing oligos the primary G cluster occurs with the highest percentage at positions 5 and 6 (Figure 4B), i.e., at the 3′ end of the oligo. Hence, the above results suggest that a central positioning of the G_3 (type Ib) primary G cluster is most favorable for HID production, followed by a 5′ end positioning. The 3′ end is the least favorable position (Figure 4B). This type of primary

Table 2: 168 Non-HID-Producing Octamers Classified According to Type of Primary and Secondary Clusters Present[a]

no.	name/number	sequence	CL	CO	IB	IB-RY
		Type Ib NG_3NG_2N				
1	384.390	GGGAGGAC	3	PS	1	R
2	384.389	GGGAGGAG	3	PS	1	R
3	331PS	GTGGGAGG	3	PS	1	R
4	384.670	CGGAGGGA	3	SP	1	R
5	384.569/270	GAGGAGGG	3	SP	1	R
6	384.456	GGAGGGCA	3	SP	1	R
7	384.455	GGAGGGTC	3	SP	1	R
8	384.462/399	GGAGAGGG	3	SP	3	RRR
9	384.452/421	GGAGTGGG	3	SP	3	RRY
10	384.386	GGGATGGC	3	PS	2	RY
11	384.666	CGGATGGG	3	SP	2	RY
12	384.396	GGGACAGG	3	PS	3	RYR
13	384.471	GGACAGGG	3	SP	3	RYR
14	929	CTGGTGGG	3	SP	1	Y
15	384.441	GGCAGGGA	3	SP	2	YR
16	384.317/176	GGGCAAGG	3	PS	3	YRR
17	384.448	GGCAAGGG	3	SP	3	YRR
18	384.436	GGCATGGG	3	SP	3	YRY
19	384.755	GGGCTGGA	3	PS	2	YY
20	384.405	GGCTGGGA	3	SP	2	YY
21	384.323/171/176	GGGTCAGG	3	PS	3	YYR
22	384.350	GGGTCTGG	3	PS	3	YYY
23	384.398	GGCTTGGG	3	SP	3	YYY
24	384.335	GGTTCGGG	3	SP	3	YYY
		Type Ic G_3NNGNG				
25	384.564	GAGGGCAG	3	SP and PS		
26	384.536	GCAGGGTG	3	SP and PS		
27	384.482	GCTGGGAG	3	SP and PS		
28	384.108/65	GTGGGAGC	3	SP and PS		
29	384.542	GCAGAGGG	3	SP	1	R
30	384.486	GCTGAGGG	3	SP	1	R
31	384.132	GTGAGGGC	3	SP	1	R
32	384.671	TGCGAGGG	3	SP	1	R
33	384.259	CGGGTGAG	3	PS	1	Y
34	384.165	GGGTGTGC	3	PS	1	Y
35	384.140	GAGTGGGC	3	SP	1	Y
36	384.752	GTGCGGGA	3	SP	1	Y
37	384.245	GGGCAGAG	3	PS	2	YR
38	384.258	GGGCTGAG	3	PS	2	YY
39	384.747	GGGTCGTG	3	PS	2	YY
40	384.316	GGGTTGCG	3	PS	2	YY
41	384.570	GAGCTGGG	3	SP	2	YY
42	489	CAGGGAGC	3	PS	1	R
43	960	CCAGGGAG	3	PS	1	R
44	384.254	CCTGGGAG	3	PS	1	R
45	384.107	CTGGGAGC	3	PS	1	R
		Type Id G_3NG				
46	384.391	GGGAGCAC	3	PS	1	R
47	384.387	GGGAGTCC	3	PS	1	R
48	384.706	CCTGAGGG	3	SP	1	R
49	384.131	CTGAGGGC	3	SP	1	R
50	384.567	GAGGGACC	3	SP	1	R
51	384.563	GAGGGCTC	3	SP	1	R
52	280	GAGGGCTC	3	SP	1	R
53	384.082	GAGGGTCC	3	SP	1	R
54	384.330	TCCGAGGG	3	SP	1	R
55	384.547	GCACAGGG	3	SP	4	RYRY
56	384.657	CGGGACTG	3	PS	3	RYY
57	384.392	GGGACTGC	3	PS	3	RYY
58	384.599	GACCAGGG	3	SP	4	RYYR
59	384.394/480	GGGACCTG	3	PS	4	RYYY
60	384.588	GACCTGGG	3	SP	4	RYYY
61	384.584/881	GACTCGGG	3	SP	4	RYYY
62	384.306	CAAGGGCG	3	PS	1	Y
63	384.162/456	CAGGGTGC	3	PS	1	Y
64	384.606/851	CTTGGGCG	3	PS	1	Y
65	384.194/158	GGGTGCTC	3	PS	1	Y
66	384.090	GGGTGTCC	3	PS	1	Y
67	384.307	TCTGGGCG	3	PS	1	Y
68	384.741	CAGCGGGA	3	SP	1	Y
69	384.345	CCAGTGGG	3	SP	1	Y
70	384.702	CCTGTGGG	3	SP	1	Y
		Type Id G_3NG				
71	384.377	GTGGGACC	3	SP	1	Y
72	384.084	GTGGGTCC	3	SP	1	Y
73	6−294M	CAGGGCAG	3	PS	2	YR
74	384.656	CGGGCAGA	3	PS	2	YR
75	384.764/184	GGGCAGAC	3	PS	2	YR
76	384.696	CGCAGGGA	3	SP	2	YR
77	384.624	CTGCAGGG	3	SP	2	YR
78	384.130	GACAGGGC	3	SP	2	YR
79	384.537/819	GCAGGGAC	3	SP	2	YR
80	384.353	GGGCACTG	3	PS	3	YRY
81	384.695	CGCATGGG	3	SP	3	YRY
82	384.226	GGGCACAG	3	PS	4	YRYR
83	384.166	CGGGTTGC	3	PS	2	YY
84	384.203/159/164	GGGTCGTC	3	PS	2	YY
85	384.047	GTTGGGCC	3	SP	2	YY
86	384.483	GCTGGGAC	3	SP	2	YY
87	384.481	GCTGGGTC	3	SP	2	YY
88	384.652	CGGGTCTG	3	PS	3	YYY
89	6−248X	GGGCCTGT	3	PS	3	YYY
90	384.139	GTCTGGGC	3	SP	3	YYY
91	384.243	GGGCTCAG	3	PS	4	YYYR
92	384.234	GGGTCCAG	3	PS	4	YYYR
93	6−223X	GGGCTCTG	3	PS	4	YYYY
94	207	GGGTCCTG	3	PS	4	YYYY
95	384.477	GCTTCGGG	3	SP	4	YYYY
96	384.334	GTCTCGGG	3	SP	4	YYYY
		Type Ie G_3NNNNN				
97	384.397	GGGAACCC	3			
98	384.395	GGGACCAC	3			
99	384.393	GGGACTCC	3			
100	384.385	GGGCAACC	3			
101	384.170	GGGCACTC	3			
102	384.196	GGGCTCTC	3			
103	384.378	GGGTCCAC	3			
104	384.013	GGGTTCCC	3			
105	384.045	CAAGGGCC	3			
106	384.129	CACAGGGC	3			
107	384.329	CACCAGGG	3			
108	384.343	CACCTGGG	3			
109	384.332	CACTCGGG	3			
110	384.739	CAGGGACC	3			
111	384.342	CCACTGGG	3			
112	384.732	CCAGGGAC	3			
113	384.708	CCTCTGGG	3			
114	384.213	CCTGGGTC	3			
115	384.659	CGGGAACC	3			
116	384.658	CGGGACAC	3			
117	384.653	CGGGTCAC	3			
118	384.197	CGGGTCTC	3			
119	384.639	CTCCAGGG	3			
120	384.632/411	CTCCTGGG	3			
121	384.630	CTCTCGGG	3			
122	384.138	CTCTGGGC	3			
123	160	CTGGGACC	3			
124	384.083/809/831	CTGGGTCC	3			
125	384.333	TCCTCGGG	3			
126	384.046	TCTGGGCC	3			
		Type IIa G_4NG_2				
127	384.475	GGAAGGGG	4	SP	2	RR
		Type IIb G_4NGNG				
128	338PS	GTGGGGAG	4	SP and PS	1	YR
		Type IIc $G_4NGNN/G_4NNGN/G_4NNNG$				
129	384.253	CAGGGGAG	4	PS	1	R
130	384.765/189	GGGGAGAC	4	PS	1	R
131	6−50X PS	GGGGAGCT	4	PS	1	R
132	384.337	CAGAGGGG	4	SP	1	R
133	384.560/591	GAGGGGTC	4	SP	1	R
134	765SP	GAGGGGAC	4	SP	1	R
135	681	GGGGAAGC	4	PS	2	RR
136	384.134	GAAGGGGC	4	SP	2	RR
137	384.601	GACAGGGG	4	SP	3	RYR
138	384.582	GACTGGGG	4	SP	3	RYY

Table 2: (Continued)

no.	name/number	sequence	CL	CO	IB	IB-RY	no.	name/number	sequence	CL	CO	IB	IB-RY
		Type IIc G_4NGNN/G_4NNGN/G_4NNNG							Type IId G_4NNNN				
139	384.654	CGGGGTGA	4	PS	1	Y	154	384.655	CGGGGAAC	4			
140	384.766	GGGGTGAC	4	PS	1	Y	155	384.640	CTCAGGGG	4			
141	340PS	CAGGGGTG	4	PS	1	Y	156	384.212	CTGGGGTC	4			
142	384.340	CAGTGGGG	4	SP	1	Y	157	384.767/217	GGGGAACC	4			
143	384.609/433	CTGTGGGG	4	SP	1	Y	158	384.759	GGGGACAC	4			
144	654SP	CGTGGGGA	4	SP	1	Y	159	384.172	GGGGACTC	4			
145	384.224	GGGGCAAG	4	PS	3	YRR	160	384.757/231	GGGGCAAC	4			
146	384.751	GGGGCTTG	4	PS	3	YYY	161	384.763/73	GGGGTCAC	4			
147	384.476	GCTTGGGG	4	SP	3	YYY	162	384.136	TCTGGGGC	4			
		Type IId G_4NNNN							Type III G_5NNN				
148	384.221	GGGGCTTC	4				163	384.758	GGGGGAAC	5			
149	384.133	CAAGGGGC	4				164	384.756	GGGGGTGA	5			
150	384.336	CACAGGGG	4				165	384.753	GTGGGGGA	5			
151	384.738	CAGGGGAC	4				166	384.616	CTGGGGGA	5			
152	384.328	CAGGGGCA	4				167	758REV	AACGGGGG	5			
153	384.211	CAGGGGTC	4				168	756REV	TGAGGGGG	5			

[a] The name or number and sequence of the octamers are indicated. The octamers are sorted by type, primary cluster length (CL), primary (P) versus secondary (S) cluster order (CO), intervening base separating the primary and secondary clusters (IB), and intervening base classification (IB-RY; R = purine; Y = pyrimidine).

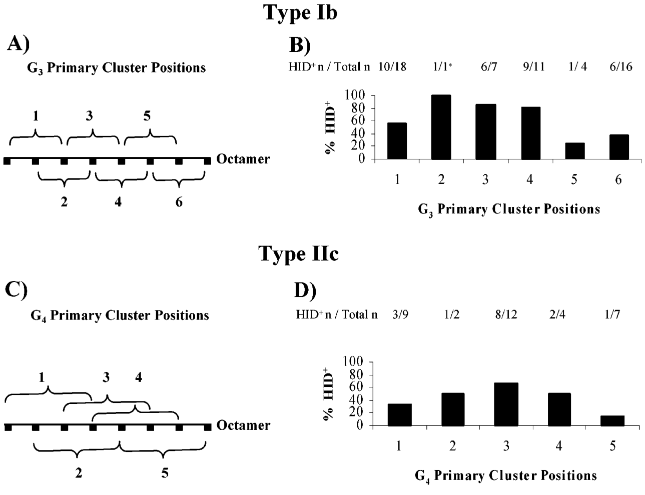


FIGURE 4: Positions of the G_3 and G_4 primary clusters. (A) Possible positions at which the G_3 primary cluster (P) can occur in a type Ib octamer. (B) Percent HID-producing octamers at each position in type Ib. (C) Possible positions at which the G_4 primary cluster (P) can occur in a type IIc octamer. (D) Percent HID-producing octamers at each position in type IIc. The value above each bar is the number of HID-producing octamers over the total number of octamers in that category. The asterisk indicates that there is only one octamer in this category.

cluster preference indicates that the secondary G cluster in this subclass is locked in position toward either the 5′ or 3′ end of the oligo.

Similarly, the G_4 primary G cluster of the type IIc subclass can occur at five different positions within an octamer (Figure 4C). In this subclass, HID-producing oligos have the highest percentage occurrence of the primary cluster at the third position (i.e., the center of an octamer, Figure 4D) and lower percentage occurrence at the first and fifth positions (i.e., the 5′ and 3′ ends, Figure 4D). The occurrence of the G_4 primary G cluster at the center of the octamer, favoring HID production, dictates that, like the type Ib subclass, the secondary G cluster in this subclass is also located at either end of the molecule.

(ii) Primary versus Secondary G Cluster Order. To determine whether the order in which the primary and the secondary G clusters occur within an octamer impacts HID production, cluster order names were assigned to each of the 198 initially selected octamers (183 oligos from the "384" library and 15 nonlibrary oligos). Accordingly, all octamers in which the primary cluster (P) occurs 5′ to the secondary cluster (S) are designated "PS", and all octamers in which the primary cluster occurs 3′ to the secondary cluster are designated "SP". Some octamers contain both types of cluster orders and are designated as "PS and SP" (Tables 1 and 2).

Initial analysis of the HID-producing versus the non-HID-producing octamers in the type Ib and IIc subclasses suggested that PS is the favorable cluster order for HID production, since the percentage of HID-producing oligos with a PS type cluster order is slightly higher. To test this hypothesis, 21 oligos were selected from the type Ib, IIc, and III subclasses, corresponding oligos were synthesized containing the same base composition but reversed cluster order (i.e., converting a PS oligo into a SP oligo or vice versa), and these oligos were assayed for HID production (Table 3).

On the basis of our hypothesis, we predicted that the new PS permutations of oligos would have an increased frequency of HID production. HID analysis of the 21 newly synthesized oligos shows that 70% of the oligos that were originally HID producing and had a SP cluster order stimulate HID production after cluster order reversal to PS (Table 3, no. 1−10). Furthermore, 75% of the oligos that were originally non-HID producing and SP type test positive for HID production after cluster order reversal to PS (Table 3, no. 11−14). One oligo that was originally HID producing and PS type remains HID positive after cluster order reversal to SP (Table 3, no. 15). Of the four originally PS and non-HID-producing oligos tested, cluster order reversal to SP results in two HID positive and two HID negative oligos (Table 3, no. 16−19). Two non-HID oligos containing five contiguous Gs at their 5′ end fail to produce HID extension products after the G cluster is moved to the 3′ end (Table 3, no. 20 and 21).

The new PS permutations of oligos have a higher percentage of HID-producing oligos. Additionally, the overall percentage of HID-producing oligos with a PS-type cluster order is slightly higher (data not shown), suggesting that the

Table 3: 21 Octamers Selected from the Original Library and Newly Designed Octamers with Reversed Cluster Orders[a]

	original sequence							cluster reversed sequence						
no.	name/number	sequence	CL	IB	RY	CO	HID	name/number	sequence	CL	IB	RY	CO	HID
1	384.454/ATsub	GGAGGGTG	3	1	R	SP	+	454PS	GGGAGGTG	3	1	R	PS	+
2	384.569/270	GAGGAGGG	3	1	R	SP	+	SH15/569PS	GAGGGAGG	3	1	R	PS	+
3	384.346	CAGGTGGG	3	1	Y	SP	+	346PS	CAGGGTGG	3	1	Y	PS	+
4	TAsub	GGTGGGAG	3	1	Y	SP	+	TASUB PS	GGGTGGAG	3	1	Y	PS	+
5	384.749/Tsub	GGTGGGTG	3	1	Y	SP	+	749PS	GGGTGGTG	3	1	Y	PS	+
6	384.214	GGTGGGTC	3	1	Y	SP	+	384.216/214PS	GGGTGGTC	3	1	Y	PS	+
7	384.551	GAGTGGGG	4	1	Y	SP	+	551PS	GAGGGGTG	4	1	RY	SP&PS	+
8	384.331	GTGGAGGG	3	1	R	SP	+	331PS	GTGGGAGG	3	1	R	PS	−
9	6−50X	GAGGGGCT	4	1	R	SP	+	6−50X PS	GGGGAGCT	4	1	R	PS	−
10	384.338	GTGAGGGG	4	1	R	SP	+	338PS	GTGGGGAG	4	1	YR	SP&PS	−
11	384.67	CGGAGGGA	3	1	R	SP	−	SH12/670PS	CGGGAGGA	3	1	R	PS	+
12	384.455	GGAGGGTC	3	1	R	SP	−	SH13/455PS	GGGAGGTC	3	1	R	PS	+
13	929	CTGGTGGG	3	1	Y	SP	−	SH14/929PS	CTGGGTGG	3	1	Y	PS	+
14	384.34	CAGTGGGG	4	1	Y	SP	−	340PS	CAGGGGTG	4	1	Y	PS	−
15	384.219	GGGGTGTC	4	1	Y	PS	+	219SP	GTGGGGTC	4	1	Y	SP	+
16	384.389	GGGAGGAG	3	1	R	PS	−	389SP/Asub/849	GGAGGGAG	3	1	R	SP	+
17	384.39	GGGAGGAC	3	1	R	PS	−	390SP/Asub3″C	GGAGGGAC	3	1	R	SP	+
18	384.765/189	GGGGAGAC	4	1	R	PS	−	765SP	GAGGGGAC	4	1	R	SP	−
19	384.654	CGGGGTGA	4	1	Y	PS	−	654SP	CGTGGGGA	4	1	Y	SP	−
20	384.758	GGGGGAAC	5	0			−	758REV	AACGGGGG	5				−
21	384.756	GGGGGTGA	5	0	Y	PS	−	756REV	TGAGGGGG	5	1	Y	SP	−

[a] The octamer name or number, sequence, primary cluster length (CL), intervening base separating the primary and secondary clusters (IB), intervening base classification (IB-RY; R = purine; Y = pyrimidine), primary (P) versus secondary (S) cluster order (CO), and the presence or absence of HID production are indicated.

PS-type cluster order may favor DNA polymerization, especially in the type Ib octamers.

(iii) Spacing and Classification of Non-G Base(s) between the Primary and Secondary G Clusters. The spacing (one, two, or three bases) and classification [purine (R) or pyrimidine (Y)] of the non-G base(s) between the primary and secondary G clusters appear to be the most important factors that influence the octamer's ability to stimulate HID production. These features are also among the primary factors that determine the type of structure a G-rich sequence forms (*20*, *25*−*27*). Assuming a total of five Gs as the number of bases comprising the two clusters (as is observed for both type Ib and type IIc subclasses), the total number of non-G bases in the rest of the octamer is three bases. Hence, one, two, or three bases can occur between the primary and secondary G clusters.

From our analysis, the most striking feature is the frequent occurrence of a single, non-G base between the primary and the secondary G clusters, among the type Ib and IIc HID-producing octamers (Figure 5). The classification of this intervening base [i.e., whether it is a purine (R) or a pyrimidine (Y)] is the most significant factor that distinguishes between HID-producing and non-HID-producing categories, especially in the type Ib subclass. A pyrimidine occurs in 95% of type Ib and in 54% of type IIc HID-producing octamers (Figure 5). These observations indicate that the influence of the intervening base on the outcome of HID production is reduced in the type IIc subclass. This also suggests that the type Ib and IIc oligos may form different polymorphic structures that stimulate DNA synthesis.

(iv) Base Identity Immediately 5′ and/or 3′ of the Primary G Cluster. Base identity of the non-G bases flanking the primary G cluster strongly influences the type of structure formed, as well as its thermal stability (*9*, *19*, *20*). To determine if the base identity 5′and 3′ of the primary G cluster correlates with the ability of an octamer to produce HID, these bases were analyzed in both subclasses (types Ib and IIc). There are 24 possible sequence permutations 5′ and 3′ of the primary G cluster, relative to its location within the octamer. However, due to the criteria used to design the library for octamer-primed sequencing technology (*3*) it

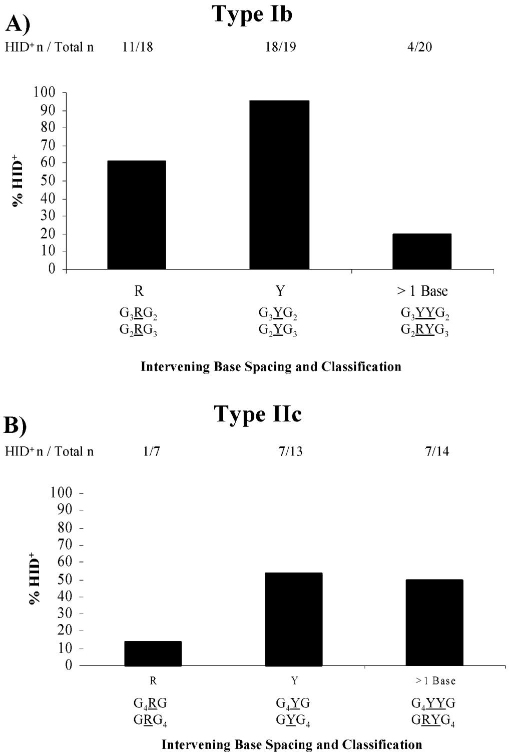


FIGURE 5: HID-producing octamers sorted by intervening base. (A)-Type Ib single intervening base, $G_3\underline{R}G_2/G_2\underline{R}G_3/G_3\underline{Y}G_2/G_3\underline{Y}G_2$; more than one intervening base, $G_3\underline{YY}G_2$. (B) Type IIc single intervening base, $G_4\underline{R}/G$, $G/\underline{R}G_4$, $G_4\underline{Y/G}$, and $G/\underline{Y}G_4$; more than one intervening base, $G_4\underline{YY/G}$. The value above each bar shows the number of HID-producing octamers over the total number of octamers in that category.

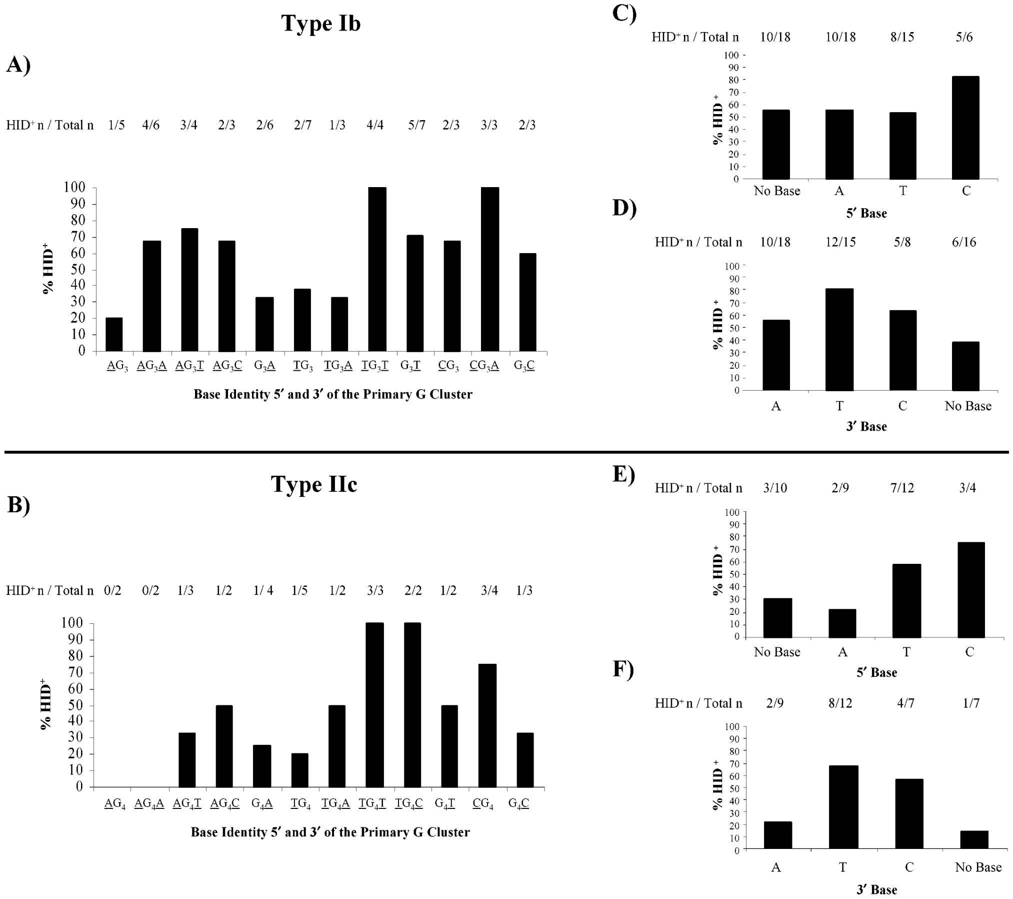


FIGURE 6: HID-producing octamers sorted by base identity 5′ and/or 3′ of the primary G cluster. (A) Type Ib: $\underline{N}G_3\underline{N}$. (B) Type IIc: $\underline{N}G_4\underline{N}$. Percent HID-producing octamers within type Ib sequences sorted by base identity (C) 5′ of the primary G cluster ($\underline{N}G_3$) or (D) 3′ of the primary G cluster ($G_3\underline{N}$). Percent HID-producing octamers within type IIc sequences sorted by base identity (E) 5′ of the primary G cluster ($\underline{N}G_4$) or (F) 3′ of the primary G cluster ($G_4\underline{N}$). The value above each bar shows the number of HID-producing octamers over the total number of octamers in that category.

should be noted that certain sequence permutations are underrepresented.

Sequence analysis reveals the importance of thymine (T) for HID production. The 5′ and 3′ sequences of “$\underline{T}G_3\underline{T}$”, “$\underline{T}G_4\underline{T}$”, “$\underline{A}G_3\underline{T}$”, and “$\underline{T}G_4\underline{C}$” (Figure 6A,B) occur with the highest percentages among the HID-producing octamers. The sequences “$G_3\underline{T}$” and “$G_4\underline{T}$” also occur at a higher frequency among HID-producing octamers (Figure 6A,B). It appears that the presence of a T at the 5′ and 3′ ends, but especially at the 3′ end of the primary G cluster, appears to have a positive influence on the ability of an octamer to stimulate HID production.

Almost all the $\underline{N}G_3$, $\underline{N}G_4$ and $G_3\underline{N}$, $G_4\underline{N}$ sequence permutations appear to be less preferred, consistent with the hypothesis that a central positioning of the primary cluster is more efficient in stimulating HID production (Figure 6A,B).

Thymine 5′ or 3′ of the primary G cluster is not the only base that influences HID production. In type Ib, if a cytosine (C) is either 5′ or 3′ of the primary cluster and the alternate base is either an adenine (A) or absent (due to end placement), HID production is favored (Figure 6A). On the other hand, in type IIc the presence of a C either 5′ or 3′ of the primary G cluster and a T or no base at the alternate 5′ or 3′ position favors HID production (Figure 6B). Analysis of the base 5′ of the primary cluster in both the type Ib and type IIc subclasses reveals that a large percentage of HID-producing octamers contain a C (Figure 6C,E). The positive influence of a T is reinforced by analysis of the 3′ base individually. A large percentage of the oligos with a T positioned immediately 3′ of the G cluster are HID producing (Figure 6D,F).

(v) Base Identity at Each Position in the Octamer. Tabulation of the base identity at each position of the octamer for all of the 198 initially selected octamers (183 oligos from the 384 library and 15 nonlibrary primers) demonstrated an overall lower percentage of Cs at any position among the HID-producing octamers. To test the hypothesis that the presence of a C would depress HID production, 18 oligos with 75% GC and varied position of C within the octamer (octamers with a SH prefix followed by a number) were designed and tested for production of HID (Table 1). Contrary to our expectation, each stimulates DNA polymerization. Thus, the primers may be underrepresenting C due to the requirement of a primary and secondary G cluster and the 75% GC limit imposed in library design (3).

The descending order of occurrence of Gs among the type Ib HID-producing octamers is at the fifth > fourth > third positions, respectively, consistent with the preference of a central position for the primary G cluster. Ts occur in the descending order of first ≡ eighth > seventh > sixth ≡ third positions, respectively, again consistent with the preference for Ts 5′ and 3′ of the primary G cluster. Adenines occur in the order of second > seventh position, with lower percent-

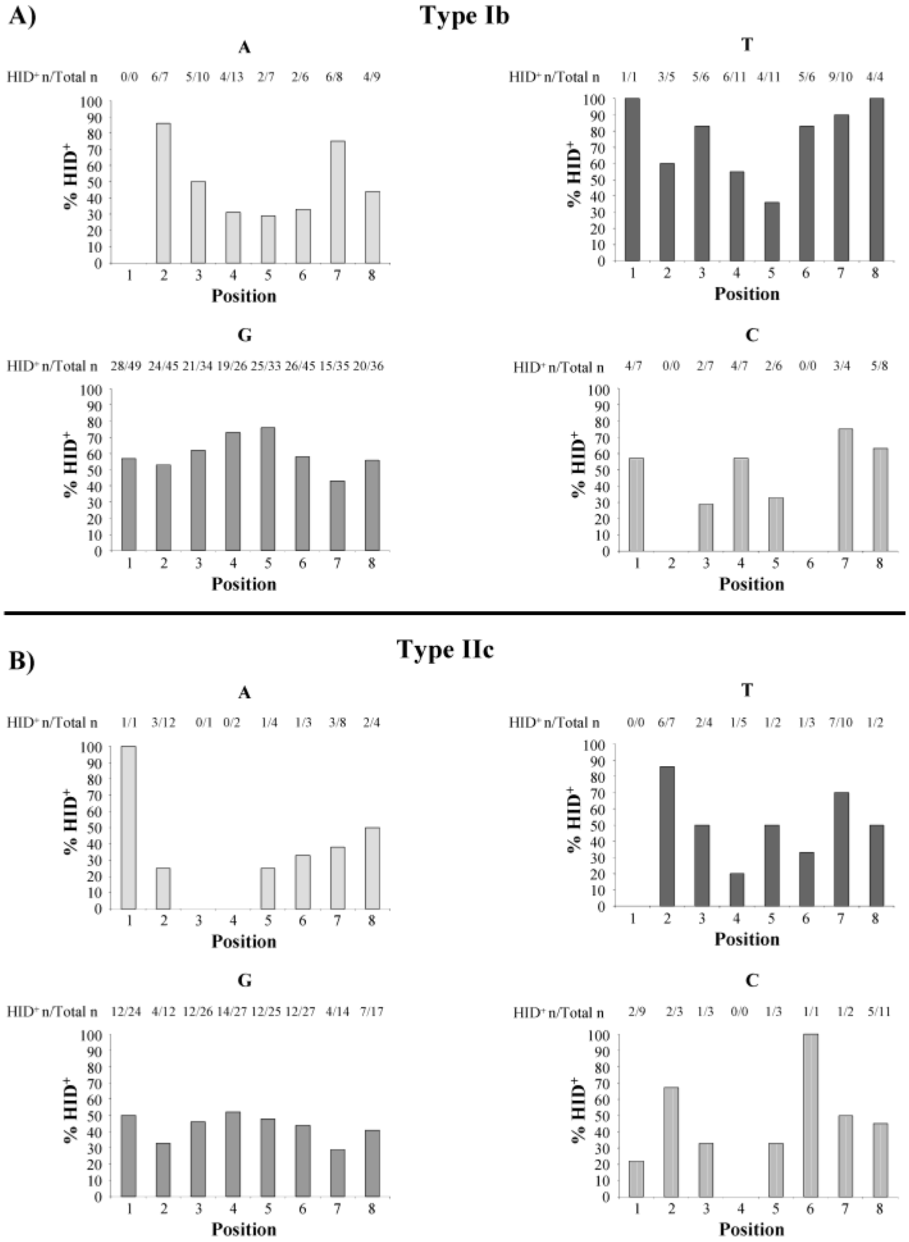


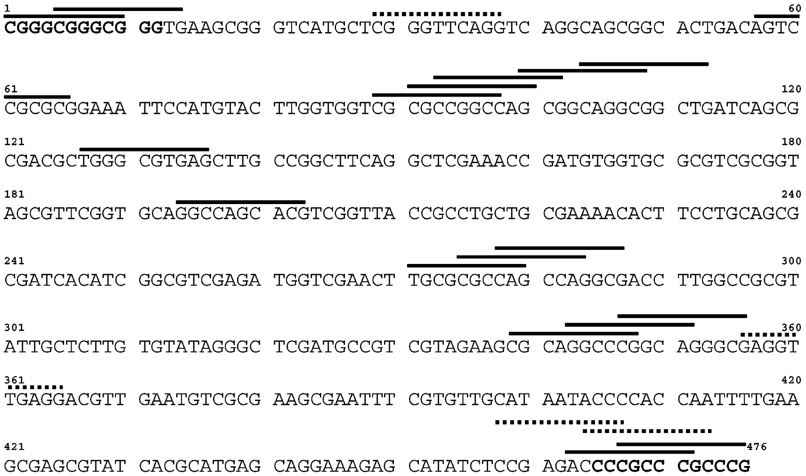
FIGURE 7: Base identity percentages in the HID-producing octamers. (A) Type Ib: base A, T, C, and G. (B) Type IIc: base A, T, C, and G. The base position is indicated below the corresponding bars which indicate the percent HID-producing oligos in that position. The value above each bar shows the number of HID-producing octamers over the total number of octamers in that category.

ages at all other positions. Cs occur in the order of seventh > eighth position with lower percentages at all other positions. The net result is consistent with the overall base preference of G > T > A > C (Figure 7A).

There is no overwhelming preference for Gs at any position within the type IIc HID-producing oligos. Ts occur in the order of second > seventh position, adenines in eighth position, and Cs in second > eighth position, with lower percentages occurring at other positions. The increased presence of Ts at the second and seventh position is consistent with the preference of Ts 5′ and 3′ of the primary G cluster (Figure 7B). These position identity differences between the type Ib and type IIc subclasses may suggest possible structural differences between the stimulatory structures formed by these different sequences.

Cloning and Sequencing of Products Generated by G-Rich Oligos (CGGG)$_3$ and G$_4$T$_2$G$_4$ (Tet1.5). HID production is typically characterized by uninterpretable high-intensity fluorescent sequence data and the presence of the Watson–Crick complement of the primer at the 3′ end of the "sequencing" reaction. This observation led us to propose that the unexpected extension products were amplified via a polymerase chain reaction (PCR).

To gain additional insight into the mechanism responsible for the unexpected polymerization products, two primers, (CGGG)$_3$ and G$_4$T$_2$G$_4$, were independently incubated with *Pfu* DNA polymerase. These primer sequences were chosen because they vary in GC content, form G-quadruplex stabilized structures, and are biologically interesting. (CGGG)$_3$ is similar to the sequence involved in the human triplet disease, and G$_4$T$_2$G$_4$ is the telomeric repeat found in *Tetrahymena* (*11*, *16*, *25*, *28*). *Pfu* DNA polymerase was chosen since this enzyme lacks the 5′ to 3′ exonuclease domain associated with primer/template independent ampli-



```
1                                                                60
CGGGCGGGCG GGTGAAGCGG GTCATGCTCG GGTTCAGGTC AGGCAGCGGC ACTGACAGTC

61                                                              120
CGCGCGGAAA TTCCATGTAC TTGGTGGTCG CGCCGGCCAG CGGCAGGCGG CTGATCAGCG

121                                                             180
CGACGCTGGG CGTGAGCTTG CCGGCTTCAG GCTCGAAACC GATGTGGTGC GCGTCGCGGT

181                                                             240
AGCGTTCGGT GCAGGCCAGC ACGTCGGTTA CCGCCTGCTG CGAAAACACT TCCTGCAGCG

241                                                             300
CGATCACATC GGCGTCGAGA TGGTCGAACT TGCGCGCCAG CCAGGCGACC TTGGCCGCGT

301                                                             360
ATTGCTCTTG TGTATAGGGC TCGATGCCGT CGTAGAAGCG CAGGCCCGGC AGGGCGAGGT

361                                                             420
TGAGGACGTT GAATGTCGCG AAGCGAATTT CGTGTTGCAT AATACCCCAC CAATTTTGAA

421                                                         476
GCGAGCGTAT CACGCATGAG CAGGAAAGAG CATATCTCCG AGACCCCGCC CGCCCG
```

FIGURE 8: Sequence and tiling pattern observed in one unique $(CGGG)_3$ cloned product (GenBank ID AY145510). The insert sequence is bordered by the input primer and its Watson−Crick complement (in bold). The solid lines above the sequence indicate the pattern matches with ≥70% identity to the sequence $(CGGG)_2CG$. The dashed lines above the sequence indicate the patterns present when a nonrelated primer sequence, $G_4T_2G_4$, is used in a pattern search requiring ≥70% identity.

fication (*1*, *29*−*31*). Reaction products synthesized without the addition of conventional templates were cloned and sequenced. Similarity searches were used to classify the sequences as either unique or matched. The number of unique sequences recovered is 15 [11 from $(CGGG)_3$; 4 from $G_4T_2G_4$ polymerization products], and the number of matched sequences recovered is 17 [3 from $(CGGG)_3$; 14 from $G_4T_2G_4$].

Analysis of the unique and matched sequences reveals the presence of the input primer at one end of the clone and, in almost all the cases, its Watson−Crick complement at the other end of the clone. The average GC content of the unique sequences is 61% ± 6% average deviation, while that for the matched sequences is 47% ± 4% average deviation. Additionally, while the matched products resulted from amplification of a contaminating DNA, there are interesting tile-like patterns associated with several of the unique products (Figure 8). The sequence of a clone amplified using the $(CGGG)_3$ primer is shown, as are the regions in the sequence that are >70% matched with the sequence $(CGGG)_2CG$. When the $G_4T_2G_4$ primer sequence is used to search this same sequence, no tiled pattern is evident, suggesting that the extension product is based (to some extent) on the sequence of the input primer. No other obvious patterns are discernible. Thus, while Watson−Crick and non-Watson−Crick base-pairing rules appear to dictate nucleotide incorporation, Watson−Crick interactions are not utilized for template formation. The mechanism responsible for the synthesis of these products remains elusive.

DISCUSSION

To better understand the sequence characteristics that ultimately determine the structural features of a stimulatory sequence, we determined the ability of a large and diverse population of octamers (sharing common features of 75% GC content and presence of three contiguous Gs) to stimulate the synthesis of unexpected DNA polymerization products [HID (*1*)]. Our observations demonstrate that, among all the criteria used to identify the sequence requirements for HID production in an octamer containing minimally three contiguous Gs, the presence of a secondary G cluster and the number and nature of intervening bases separating the G clusters (especially for a type Ib) are the most important factors.

The HID phenomenon involves alternative, non-Watson−Crick interactions between individual primer molecules to stabilize the stimulatory structure(s). This current study demonstrates additional sequence requirements for the unexpected DNA polymerization. By library design the assayed octamers have a 75% GC content, and most of the primers that are able to stimulate DNA synthesis have Gs in five of the eight positions. Two of the three remaining bases are A and/or T, and the third base could be a C or a G. Taking into consideration the sequence information of the stimulatory G-rich sequences, one expects maximally a single A•T or a single G•C base pair in an intramolecular hairpin or intermolecular primer association. Neither of these standard Watson−Crick interactions is expected to be stable at even the lowest temperature used in the cycling regime (40 °C). Additionally, "fill-in" synthesis of an antiparallel duplex (involving alternative base pairings) by a polymerase cannot account for the production of long HID sequences (lengths >170 bases). Therefore, the HID phenomenon involves alternative, non-Watson−Crick interactions that stabilize the stimulatory structure(s).

Our earlier work demonstrated the importance of Hoogsteen base pairing in structure formation (*1*). Hoogsteen interactions can stabilize parallel-stranded duplexes, triplex, and G-quartets. However, parallel duplexes are not a suitable substrate for DNA synthesis and cannot account for the observed product lengths. The formation of a triplex would require the presence of an additional pyrimidine-rich third strand, which is lacking from our assay. Furthermore, we detect high-intensity data at elevated temperatures, far above those at which mismatched sequences will form duplex DNA (*22*). Although DNA can adopt diverse and interesting structures, the evidence is most consistent with a structure involving G−G Hoogsteen base pairing, such as a G-quartet stabilized structure, as the most likely structure(s) to be stable in these conditions and relevant for the polymerase activity detected in our system.

Short G-rich sequences containing a single run of contiguous Gs at their 3′ or 5′ end with a G as the terminal base are known to form G4-DNA in the presence of Na^+, K^+, NH_4^+, or Mg^{2+} ions (*7*, *9*, *17*, *18*, *32*, *33*). In our polymerization assays the only ions added to the reaction are 2 mM $MgCl_2$, but depending on the efficiency of oligo desalting after synthesis, some ammonium ions may remain associated with the oligo (Dr. Anthony Yeung, personal communication). A low concentration of NH_4^+ or Mg^+ ions may be responsible for stabilizing the stimulatory structure under our polymerization assay conditions. G-quartets can form between 4 and 65 °C (*15*, *34*). Additionally, they can form within a wide pH range, in general the optimum being between pH 6.0 and pH 8.0. Once formed, these structures are stable to pH 12.0 (*23*), well beyond the pH 9.0 in our assay.

Generally, the number of Gs (out of the total number of bases in an oligo) that form G-quartets positively correlates with both the rate of G4-DNA formation and the thermal stability of the structure (*23*). This may explain the inability of the four or five contiguous Gs (types IId and III) containing octamers to support DNA polymerization. A contiguous stretch of four or five Gs may form a very stable structure that is rendered essentially inaccessible to the polymerase. Interestingly, large percentages of type IIa−c sequences, containing a G4 primary cluster and a secondary G cluster, produce HID (Figure 3). Our analysis of type IIc HID-producing sequences reveals that a large number contain one or more bases between the primary and secondary clusters (Figure 5B). Perhaps this is because the presence of bases that are not involved in base pairing destabilizes G-quartet structures (*35*). The stability of a given structure depends on the balance between the favorable free energy of quartet formation and the unfavorable contribution of bases not involved in pairing or quartet formation (*23*). The presence and positioning of the non-G base(s) in HID-positive sequences may sufficiently destabilize the structure to allow polymerase access.

The majority of the HID-producing type I sequences also contain a G3 primary cluster and a secondary G cluster. However, our analysis of type Ib HID-producing sequences reveals a distinct preference for a single base between the primary and secondary clusters (Figure 5A), which may reflect the relatively reduced stability of these structures. In our analysis, T is the most frequent single base present between the primary and secondary cluster (79%; $n = 15/19$; type Ib). Thus, the presence of a T or an A (61%; $n = 11/18$; type Ib) between the primary and secondary clusters in the majority of the HID-producing octamers may be an important feature that facilitates the polymerase's access to base information within the template structure.

Substituting an A residue for a G in the telomeric consensus sequence results in a decrease in the structure's stability (*9*). It has also been observed by Guo and co-workers (*35*) that adding a T or increasing the number of Ts 5′ to a stretch of Gs decreases the quadruplex's stability. In most telomeric DNA sequences the base 5′ of the G cluster is limited to T or A (*36*). There is a striking similarity between these characteristics present in telomeric sequences and the observed preference for T residues both 5′ and 3′ of the primary G cluster for promoting HID production in the present study.

Prediction of Structure(s) Capable of Stimulating DNA Polymerization. One of the diagnostic tests for the presence of G-tetrads is to establish involvement of N7 of guanines in Hoogsteen-type base pairing (*10*, *37*). Using base analogues 7-deaza-dG and 2-aminopurine substituted oligos, we established that the structure formed under our sequencing assay conditions is stabilized by interactions consistent with Hoogsteen base pairing (*1*). Since the polymerase utilizes primers containing 7-deaza-dG and 2-aminopurine in extension reactions and similarly substituted DNA is recognized as a template for synthesis, an affect on HID production results from altered association between the input DNA, the "primer". More specifically, data demonstrating differential thermal stabilities (relative to Watson−Crick G•C base pairing) and formation of extended DNA structures up to approximately 10 times longer than the input oligo length suggested that a structure stabilized by alternative base interactions, such as those stabilizing G-quartets, stimulates the unexpected DNA polymerization activity (*1*).

The crystal structure of *Taq* DNA polymerase indicates that the enzyme is in very close contact with the DNA through the minor groove of an A-form DNA duplex located immediately adjacent to the site of nucleotide addition (*38*). The polymerase interacts with the template/primer duplex over approximately an 8 base pair region, and the templating base is flipped into the major groove of the newly synthesized DNA (*38*−*40*). Additionally, it is well established that the polymerase requires a 3′-OH recessed primer strand for a template-directed extension (*29*, *38*).

Taking together all of the above-mentioned information and the analysis of the eight base sequences presented in this study, we suggest that the most favorable structure to stimulate the unexpected DNA polymerization products is an intermolecular structure with shallow minor groove dimensions similar to those found in replicating DNA. Hence, we propose a model in which a parallel G-quadruplex and/or a hairpin fold-back quadruplex structure(s) stimulate(s) DNA polymerization (Figure 9).

There are several additional requirements of the stimulatory structure. First, it is essential that the DNA polymerase gain access to base information within the structure. This structure must be stable enough to form in conditions that promote polymerase activity but not be too stable that it is not able to open along its length. Second, the structure must be flexible so that it can be accommodated in the polymerase active site. For these reasons, we propose that the stimulatory structure is initially a G-quartet stabilized structure but that the quartet is opened to allow the polymerase base access and to provide the flexibility needed for template manipulation.

An elongated, open-ended box can be used to illustrate this model. In closed "box" form it represents the rigid, closed G-quartet stabilized structure. Alternatively, the box can be opened along its length to access the "contents" and increase flexibility. The open structure is more flexible along its length, and information that had been previously inside the closed (box) structure is accessible (Figure 9A). An alternative hairpin structure may also function as a template (Figure 9B).

Biological Significance. Most telomeres contain a consensus sequence of $d(T/A)_{1-4}G_{1-8}$ (*41*, *42*). The 3′-terminus of the G-rich strand extends beyond the corresponding 5′

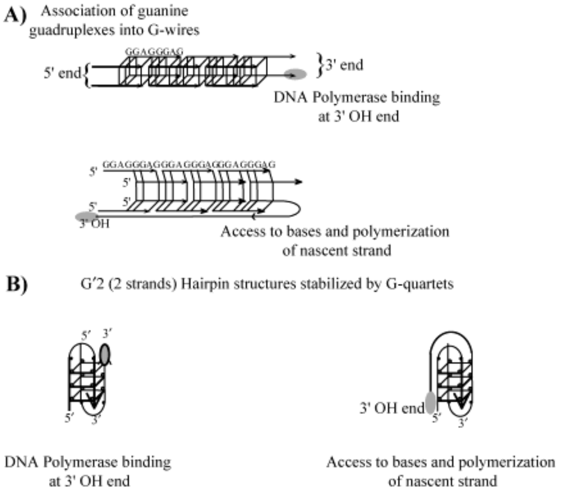

A) Association of guanine guadruplexes into G-wires
GGAGGGAG
5' end
3' end
DNA Polymerase binding at 3' OH end
GGAGGGAGGGAGGGAGGGAGGGAG
5'
5'
5'
5'
3' OH
Access to bases and polymerization of nascent strand
B) G′2 (2 strands) Hairpin structures stabilized by G-quartets
5′
3′
5′
3′
3' OH end
5′
3′
DNA Polymerase binding at 3' OH end
Access to bases and polymerization of nascent strand

FIGURE 9: Proposed model of the mechanism for DNA polymerization. (A) Formation of an intermolecular parallel quadruplex DNA stabilized by G-quartets and further association of these quadruplexes into higher order G-wires. Association of DNA polymerase (shaded oval) to the 3′ end and access to base information within the structure lead to synthesis of the nascent DNA strand. (B) Formation of a hairpin/fold-back quadruplex DNA stabilized by G-quartets. Association of DNA polymerase to the 3′ end and access to base information within the structure lead to synthesis of the nascent DNA strand.

complementary C-rich strand (*43*). The exact length of the G-rich strand varies, and these single-strand, G-rich sequences are effective substrates for telomerase and may form several noncanonical G-quartet stabilized structures (*16*). To date there is no direct evidence for the formation of such structures in vivo. However, several proteins are identified from *Tetrahymena*, *Oxytricha*, *Saccharomyces*, and humans that promote the formation, stabilization, or resolution of quadruplex structures (*32*, *37*, *44*−*47*). Additionally, the discovery of a naturally occurring G4 DNA binding antibody from mice (*48*) and the recent finding of reaction with *Stylonychia* macronuclei of in vitro generated G-quadruplex-specific antibodies (*49*) suggest the existence and importance of such structures within the cell. G-quadruplexes are proposed to be involved in the alignment of chromosomes during meiosis (*12*, *16*), replication (*47*), transcriptional regulation (*50*), recombination (*25*, *51*), and telomere maintenance (*45*, *52*).

Our analyses of the G-rich oligos reveal the preference for thymine as the intervening base and thymine or adenine as the G cluster flanking bases bearing striking similarities to telomeric sequences. Furthermore, telomeric sequences tested in our assay are able to polymerize the nonconventional DNA products (*1*). Hence, we suggest that this unusual polymerization activity may occur in vivo and that noncanonical base pairs formed from G-rich sequences, such as G-quadruplex stabilized structures, may be recognized by a DNA polymerase within the cell. While it is possible that the activity we describe could be an artifact of the in vitro reaction, our observations suggest that there is still much to learn about DNA polymerase action.

ACKNOWLEDGMENT

We thank Anelia Kraltcheva, Brent Mulder, Nathan Agnew, and Steve Anaya for discussions and critical review of the manuscript. Prince John Thommen is gratefully acknowledged for assistance with assembly and electrophoresis of the polymerization reactions.

REFERENCES

1. Ying, J., Bradley, R. K., Jones, L. B., Reddy, M. S., Colbert, D. T., Smalley, R. E., and Hardin, S. H. (1999) *Biochemistry 38*, 16461−16468.
2. Marsh, T. C., and Henderson, E. (1994) *Biochemistry 33*, 10718−10724.
3. Hardin, S. H., Jones, L. B., Homayouni, R., and McCollum, J. C. (1996) *Genome Res. 6*, 545−550.
4. Abad, J. P., and Villasante, A. (1999) *FEBS Lett. 453*, 59−62.
5. Venczel, E. A., and Sen, D. (1993) *Biochemistry 32*, 6220−6228.
6. Wang, Y., and Patel, D. J. (1993) *Structure 1*, 263−282.
7. Sen, D., and Gilbert, W. (1992) *Biochemistry 31*, 65−70.
8. Smith, F. W., and Feigon, J. (1992) *Nature 356*, 164−168.
9. Hardin, C. C., Henderson, E., Watson, T., and Prosser, J. K. (1991) *Biochemistry 30*, 4460−4472.
10. Williamson, J. R., Raghuraman, M. K., and Cech, T. R. (1989) *Cell 59*, 871−880.
11. Henderson, E., Hardin, C. C., Walk, S. K., Tinoco, I., Jr., and Blackburn, E. H. (1987) *Cell 51*, 899−908.
12. Sen, D., and Gilbert, W. (1988) *Nature 334*, 364−366.
13. Howell, R. M., Woodford, K. J., Weitzmann, M. N., and Usdin, K. (1996) *J. Biol. Chem. 271*, 5208−5214.
14. Kettani, A., Kumar, R. A., and Patel, D. J. (1995) *J. Mol. Biol. 254*, 638−656.
15. Jin, R. Z., Breslauer, K. J., Jones, R. A., and Gaffney, B. L. (1990) *Science 250*, 543−546.
16. Sundquist, W. I., and Klug, A. (1989) *Nature 342*, 825−829.
17. Guschlbauer, W., Chantot, J. F., and Thiele, D. (1990) *J. Biomol. Struct. Dyn. 8*, 491−511.
18. Sen, D., and Gilbert, W. (1990) *Nature 344*, 410−414.
19. Hardin, C. C., Watson, T., Corregan, M., and Bailey, C. (1992) *Biochemistry 31*, 833−41.
20. Balagurumoorthy, P., and Brahmachari, S. K. (1994) *J. Biol. Chem. 269*, 21858−21869.
21. Hardin, C. C., Corregan, M. J., Lieberman, D. V., and Brown, B. A., II (1997) *Biochemistry 36*, 15428−15450.
22. Jones, L. B., and Hardin, S. H. (1998) *Nucleic Acids Res. 26*, 2824−2826.
23. Sen, D., and Gilbert, W. (1992) *Methods Enzymol. 211*, 191−199.
24. Henderson, E. R., Moore, M., and Malcolm, B. A. (1990) *Biochemistry 29*, 732−737.
25. Balagurumoorthy, P., Brahmachari, S. K., Mohanty, D., Bansal, M., and Sasisekharan, V. (1992) *Nucleic Acids Res. 20*, 4061−4067.
26. Lu, M., Guo, Q., and Kallenbach, N. R. (1992) *Biochemistry 31*, 2455−2459.
27. Miura, T., and Thomas, G. J., Jr. (1994) *Biochemistry 33*, 7848−7856.
28. Usdin, K. (1998) *Nucleic Acids Res. 26*, 4078−4085.
29. Kornberg, A. and Baker, T. (1992) *DNA Replication*, 2nd ed., W. H. Freeman, New York.
30. Hanaki, K., Odawara, T., Muramatsu, T., Kuchino, Y., Masuda, M., Yamamoto, K., Nozaki, C., Mizuno, K., and Yoshikura, H. (1997) *Biochem. Biophys. Res. Commun. 238*, 113−118.
31. Hanaki, K., Odawara, T., Nakajima, N., Shimizu, Y. K., Nozaki, C., Mizuno, K., Muramatsu, T., Kuchino, Y., and Yoshikura, H. (1998) *Biochem. Biophys. Res. Commun. 244*, 210−219.
32. Schierer, T., and Henderson, E. (1994) *Biochemistry 33*, 2240−2246.
33. Nagesh, N., and Chatterji, D. (1995) *J. Biochem. Biophys. Methods 30*, 1−8.
34. Marotta, S. P., Tamburri, P. A., and Sheardy, R. D. (1996) *Biochemistry 35*, 10484−10492.
35. Guo, Q., Lu, M., and Kallenbach, N. R. (1993) *Biochemistry 32*, 3596−3603.
36. Shida, T., Suda, M., and Sekiguchi, J. (1998) *Nucleosides Nucleotides 17*, 575−584.
37. Giraldo, R., and Rhodes, D. (1994) *EMBO J. 13*, 2411−2420.
38. Li, Y., Korolev, S., and Waksman, G. (1998) *EMBO J. 17*, 7514−7525.
39. Eom, S. H., Wang, J., and Steitz, T. A. (1996) *Nature 382*, 278−281.
40. Korolev, S., Nayal, M., Barnes, W. M., Di Cera, E., and Waksman, G. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 9264−9268.
41. Shampay, J., Szostak, J. W., and Blackburn, E. H. (1984) *Nature 310*, 154−157.

42. Zakian, V. A. (1989) *Annu. Rev. Genet. 23*, 579−604.
43. Henderson, E. R., and Blackburn, E. H. (1989) *Mol. Cell. Biol. 9*, 345−348.
44. Fang, G., and Cech, T. R. (1993) *Cell 74*, 875−885.
45. Liu, Z., Frantz, J. D., Gilbert, W., and Tye, B. K. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 3157−3161.
46. Lu, Q., Schierer, T., Kang, S. G., and Henderson, E. (1998) *Nucleic Acids Res. 26*, 1613−1620.
47. Sun, X. G., Cao, E. H., He, Y. J., and Qin, J. F. (1999) *J. Biomol. Struct. Dyn. 16*, 863−872.
48. Brown, B. A., II, Li, Y., Brown, J. C., Hardin, C. C., Roberts, J. F., Pelsue, S. C., and Shultz, L. D. (1998) *Biochemistry 37*, 16325−16337.
49. Schaffitzel, C., Berger, I., Postberg, J., Hanes, J., Lipps, H. J., and Pluckthun, A. (2001) *Proc. Natl. Acad. Sci. U.S.A. 98*, 8572−8577.
50. Simonsson, T., Pecinka, P., and Kubista, M. (1998) *Nucleic Acids Res. 26*, 1167−1172.
51. Muniyappa, K., Anuradha, S., and Byers, B. (2000) *Mol. Cell. Biol. 20*, 1361−1369.
52. Sun, D., Thompson, B., Cathers, B. E., Salazar, M., Kerwin, S. M., Trent, J. O., Jenkins, T. C., Neidle, S., and Hurley, L. H. (1997) *J. Med. Chem. 40*, 2113−2116.

BI020380W