Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 2811–2816, April 1996
Biochemistry

Sequence specificity of triplex DNA formation: Analysis by a combinatorial approach, restriction endonuclease protection selection and amplification

PAUL HARDENBOL AND MICHAEL W. VAN DYKE*

Department of Tumor Biology, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030

Communicated by Peter B. Dervan, California Institute of Technology, Pasadena, CA, November 20, 1995

ABSTRACT **We have devised a combinatorial method, restriction endonuclease protection selection and amplification (REPSA), to identify consensus ligand binding sequences in DNA. In this technique, cleavage by a type IIS restriction endonuclease (an enzyme that cleaves DNA at a site distal from its recognition sequence) is prevented by a bound ligand while unbound DNA is cleaved. Since the selection step of REPSA is performed in solution under mild conditions, this approach is amenable to the investigation of ligand–DNA complexes that are either insufficiently stable or not readily separable by other methods. Here we report the use of REPSA to identify the consensus duplex DNA sequence recognized by a G/T-rich oligodeoxyribonucleotide under conditions favoring purine-motif triple-helix formation. Analysis of 47 sequences indicated that recognition between 13 bases on the oligonucleotide 3′ end and the duplex DNA was sufficient for triplex formation and indicated the possible existence of a new base triplet, G·AT. This information should help identify appropriate target sequences for purine-motif triplex formation and demonstrates the power of REPSA for investigating ligand–DNA interactions.**

Strategies for identifying consensus nucleic acid binding sequences involving multiple rounds of ligand–nucleic acid complex selection from a pool of random oligonucleotides followed by amplification of the selected sequences have proven to be a powerful means of identifying specific ligand binding sites. These combinatorial approaches, also termed CASTing (1), *in vitro* genetics (2), or directed molecular evolution (3), have been used to determine sequence requirements of several types of ligand–nucleic acid interactions, including protein–DNA (4–7), protein–RNA (8, 9), RNA–small molecule (10), DNA–small molecule (11), and RNA–DNA triplexes (12).

A flow diagram depicting an archetype combinatorial approach is shown in Fig. 1*A*. A population of nucleic acids containing a region of random sequence is incubated with a ligand under conditions that allow formation of specific ligand complexes. The subset of sequences able to form complexes is then isolated from uncomplexed nucleic acids by a selection method. By using an amplification method such as PCR, this subset of nucleic acids is amplified to an amount suitable for further manipulation. This series of steps, including complex formation, selection, and amplification, constitutes one round of a combinatorial method. Because the ligand-binding sequences usually represent only a very small fraction of the total pool of sequences, and most selection methods provide only a limited degree of enrichment, several rounds are often necessary before the population of sequences capable of ligand-specific interactions constitutes a majority.

A limitation to these techniques arises from the reliance on a physical separation of ligand-bound from unbound nucleic acids. Included are methods that exploit the different physical properties of the ligand–nucleic acid complex, e.g., reduced electrophoretic mobility and an electrophoretic mobility shift assay (6, 7) or increased hydrophobicity and filter binding (4, 8). Alternatively, affinity methods, including immunoprecipitation (5, 9) and matrix-immobilized ligands (10–12), can be used to enrich for ligand-binding sequences. Such methods are not suitable for selecting ligand-binding sequences if the physical properties of the specific complex are not sufficiently different from the uncomplexed nucleic acids or if affinity methods are not available.

As an alternative combinatorial approach for investigating ligand–DNA interactions, we have developed a selection process, restriction endonuclease protection selection and amplification (REPSA), that relies on the inhibition of an enzymatic activity to select for ligand-bound DNAs (Fig. 1*B*). Protection from endonuclease cleavage is a well-proven method for assaying DNA–ligand binding (13–15). In REPSA, a restriction endonuclease that cleaves at a site distal from its recognition sequence is used to selectively cleave unbound DNA, while DNA complexed with ligand is protected by occlusion of the cleavage site. Selection, therefore, can be performed in solution under mild conditions. Because the only requirement for the ligand is that it be able to block cleavage, no prior knowledge of the characteristics of the ligand is needed.

We have used REPSA to investigate the spectrum of duplex DNAs capable of specifically interacting with a single-stranded oligodeoxyribonucleotide through purine-motif triple-helix formation. Purine-motif triplexes form when a purine-rich oligonucleotide binds in an antiparallel orientation to a run of purine acceptors in the major groove of duplex DNA. Base triplets in this motif include G·GC, A·AT, and T·AT (16, 17), where the convention is the nucleotide in the third strand hydrogen bonds to (·) the purine acceptor in the duplex DNA. Through REPSA we were able to determine the consensus duplex sequence recognized by a purine-motif triplex-forming oligonucleotide, the major and minor base triplets involved, the average length of duplex DNA recognized, and unexpected consensus protein binding sites that arose from this method.

MATERIALS AND METHODS

Oligonucleotides. Phosphodiester oligodeoxyribonucleotides were prepared on a Millipore Cyclone DNA synthesizer. The nucleotide sequences (5′ → 3′) of oligonucleotides used in this study are as follows: ODN1, TGGGTGGGGTGGGGTGGGT; RPR, CTAGGAATTCGTGCAGTCTAGAG; RPL, CTCCAAGCTTGTGCAGCTGCAGG; 65R19, CTAGGAATTCGTGCAGTCTAGAG-N_{19}-CCTGCAGCTGCACAAGCTTGGAG; MS5, TGTTGTGTGGAATTGTG; MS6, CAAGGCGATTAAGTTGG. For the oligonucleotide

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

Abbreviations: REPA, restriction endonuclease protection assay; REPSA, restriction endonuclease protection selection and amplification.

*To whom reprint requests should be addressed.

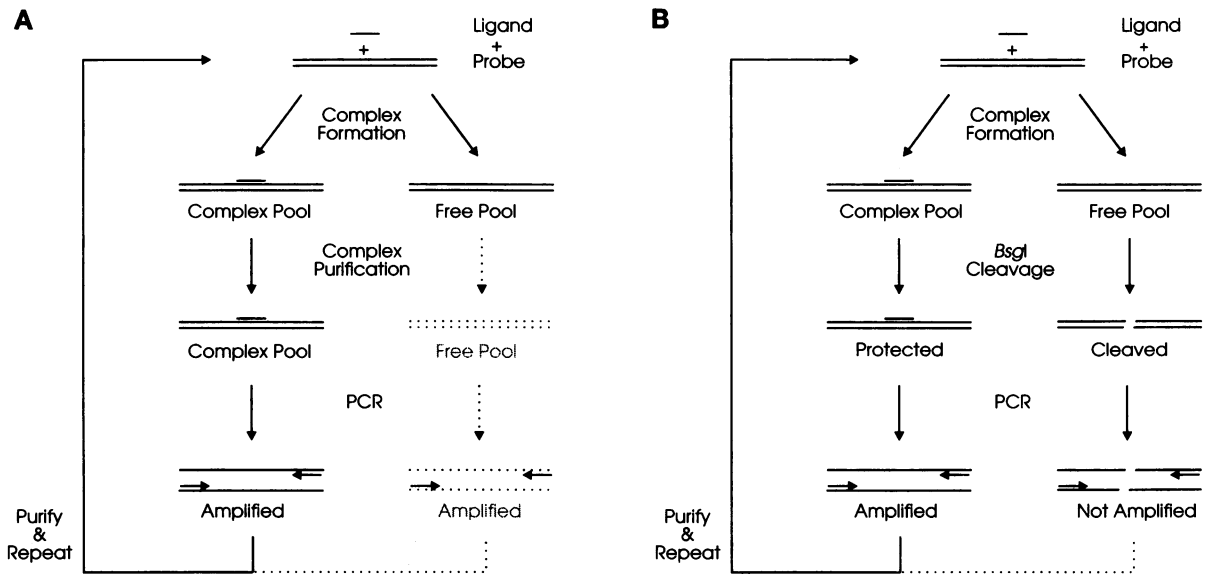

FIG. 1. Flow diagrams of combinatorial approaches for selecting duplex DNAs containing ligand-specific binding sites. (*A*) Cyclic amplification and selection of targets, CASTing. (*B*) REPSA.

65R19, sites containing mixed bases (N) were synthesized by using an equimolar mixture of each phosphoramidite.

REPSA. The double-stranded selection template ST1 was synthesized by four rounds of PCR using the oligonucleotide 65R19 and the RPR and RPL amplimers. To allow triplex formation, 10 ng of ST1 was incubated in 10 μl with 5 μM ODN1/12 mM $MgCl_2$/40 mM Hepes·NaOH, pH 8.4/0.02% Nonidet P-40 for 2 h at either 30°C or 37°C. After triplex formation, 10 μl of 4 units of *Bsg* I (New England Biolabs)/80 μM *S*-adenosylmethionine/2× *Bsg* I reaction buffer (40 mM Tris acetate/100 mM potassium acetate/20 mM magnesium acetate/2 mM dithiothreitol, buffered to pH 7.9) was added to each sample, and the incubation was continued for an additional 30 min. To amplify the *Bsg* I cleavage-resistant duplex DNA subpopulation, 200 ng of RPR/200 ng of RPL/5 units of *Taq* DNA polymerase/0.25 mM dATP/0.25 mM dCTP/0.25 mM dGTP/0.25 mM dTTP/10 mM Tris·HCl, pH 9.0/50 mM KCl/1 mM $MgCl_2$/2 μCi of [α-^{32}P]dATP (1 Ci = 37 GBq) was added to each sample, to a final total volume of 100 μl. The amplification profile used for PCR was 94°C for 1 min followed by 50°C for 3 min. Duplicate reactions were amplified for six and nine cycles. After PCR amplification, 2 μl of each reaction mixture was analyzed by PAGE and autoradiography to determine relative levels of amplification. The balance of each mixture was phenol-extracted, and the aqueous phase was concentrated on a Millipore Ultrafree-MC 5000 cellulose spin filter by centrifugation for 30 min at 15,000 × *g*. Filters were washed 10 min with 100 μl of Tris/EDTA and centrifuged for 30 min, and the remaining DNA was resuspended in 20 μl of Tris/EDTA. These steps—triplex formation, *Bsg* I cleavage, PCR amplification, and filter purification—were repeated until a population of cleavage-resistant DNAs was detected by comparing the levels of PCR-amplified DNA from triplex-selected reactions and control reactions.

Sequence Determination. The emergent triplex forming duplex DNAs were digested with *Eco*RI and *Hin*dIII and cloned into similarly cut plasmid pUC19 by standard protocols (18). Individual colonies were used to inoculate 5-ml overnight cultures in Luria broth medium containing ampicillin at 0.2 mg/ml. To screen these colonies for the presence of plasmids with triplex-forming inserts, 5 μl of these bacterial suspensions was added to PCR mixtures containing the MS5 and MS6 primers, and 20 cycles of PCR were done as described above. The resulting 187-bp DNA fragments containing the inserts were subjected to *Bsg* I cleavage with or without added ODN1 under triplex forming conditions; the resulting DNA fragments were analyzed by nondenaturing PAGE and autoradiography. Miniplasmid preparations were made from the positive-scoring clones, and their inserts were sequenced by Sanger enzymatic sequencing (18).

Statistical Analysis. The starting nucleotide distribution in the random sequence was determined by sequencing eight subclones (a total of 149 nt) containing the unselected ST1 starting material. A nucleotide distribution of 36% T, 24% A, 22% G, and 18% C was found. The significance of experimentally determined consensus sequences was determined by a χ^2 comparison of distributions in consensus sequences to the starting distribution, with P values <0.05 considered significant.

Affinity Determination. Representative sequences were radiolabeled by amplification in a PCR containing the RPR and RPL primers and 10 μCi of [α-^{32}P]dATP. Each probe (0.06 pmol) was incubated with increasing concentrations of ODN1 for 3 h at 30°C under triplex-forming conditions and then subjected to *Bsg* I cleavage as described above. Reaction products were analyzed by PAGE and autoradiography. Relative levels of triplex-protected (uncleaved) vs. unprotected (cleaved) probe were determined by densitometry of the appropriate bands. The concentration of ODN1 that confers 50% protection of a probe was considered equivalent to the affinity of that sequence for ODN1.

RESULTS

Design of the REPSA Selection Template. The selection template ST1 (Fig. 2) was designed to provide a pool of all possible 19-bp sequences and a method for selecting, amplifying, and subcloning the triplex-forming subset of this population. The center of the template contains a 19-bp randomized cassette with 4^{19} or 275 billion possible sequence combinations. On either side of this cassette are 23-bp flanks with multiple functions. Each contains a recognition sequence for *Bsg* I, a type IIS restriction endonuclease that cleaves DNA at a site 16 bases 3′ of its GTGCAG recognition sequence (19), positioned so that their cleavage sites are centered in the random cassette. Redundant sites were chosen to increase enzyme cleavage efficiency, thereby maximizing potential selection efficiency. We had previously found that *Bsg* I cleavage

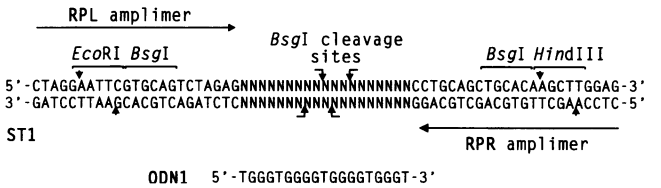

FIG. 2. Design of the selection template, ST1, used for the identification of duplex DNA sequences capable of triplex formation with the G/T-rich oligodeoxyribonucleotide ODN1. Locations of restriction endonuclease binding (brackets) and cleavage (arrows) sites are indicated. Long horizontal arrows correspond to the sequences of the PCR amplimers RPL and RPR. N, random nucleotides.

was effectively prevented when its cleavage site was involved in a triple helix (data not shown). Thus, selective cleavage of triplex-unprotected ST1 by *Bsg* I is the basis of selection in REPSA. Intact (triplex protected) selection templates can then be enriched by PCR amplification using the RPR and RPL primers that anneal to the flanks of ST1. Note that *Bsg* I recognition sites were included within the primer annealing sequences to prevent the emergence of *Bsg* I binding site mutants. These flanks also contained an *Eco*RI and a *Hin*dIII restriction site for the purpose of subcloning the template into the plasmid vector pUC19.

Selection of a Triplex-Forming Population. A flow diagram of the REPSA protocol is shown in Fig. 1*B*. To begin the selection for triplex forming sequences, we incubated 10 ng (1.4×10^{11} molecules) of the selection template ST1 with ODN1 under conditions that facilitated triplex formation. After 2 h at either 30°C or 37°C, these mixtures were subjected to cleavage by *Bsg* I. Selections at different temperatures were performed to test the effects of temperature on the specificity of purine-motif triplex formation, a correlation previously observed for pyrimidine-motif triplex formation (20). Duplex DNA sequences were selectively cleaved while DNA sequences involved in triplexes were protected from cleavage. After challenge with *Bsg* I, the remaining intact DNAs were amplified by PCR using the RPR and RPL primers. We found that amplification should be limited to nine or less cycles to prevent the formation of "bubbles," template strands annealed with mismatched cassettes that arose when the PCR consumed a significant proportion of the available primers. In each round of enrichment, three sets of PCRs were performed for each selection temperature. In addition to amplification of the triplex selection described above, control amplifications where either ODN1 or both ODN1 and *Bsg* I were omitted from the REPSA protocol were performed to give the maximum (ODN1$^-$, *Bsg* I$^-$) and minimum (ODN1$^-$, *Bsg* I$^+$) possible levels of amplified DNA. By comparing relative levels of DNA amplified in the triplex selection (ODN1$^+$, *Bsg* I$^+$) to these controls, the progress of enrichment could be monitored.

In the first 10 rounds of enrichment, amplified DNA levels in the triplex selection lanes were similar to the levels observed in the ODN1$^-$, *Bsg* I$^+$ control lanes, indicating that the percentage of templates capable of forming triplexes was below the detection limit of our cleavage protection assay (round one shown, Fig. 3*A*). In round 11, an increase in amplified DNA in the triplex selection lanes was observed relative to the minimum amplification control lanes (1.9-fold for the selection at 30°C; 1.7-fold for 37°C as determined by densitometry, Fig. 3*B*), suggesting the emergence of triplex competent sequences. To examine these emergent sequences, each of the control and triplex-selected pools from the 11th round of selection was subcloned into pUC19 and transformed into bacteria.

Selected Triplex Sequences. Radiolabeled DNA probes containing ST1 inserts were generated by PCR amplification directly from individual bacterial clones. The ability of these DNA fragments to form triplexes with ODN1 was determined

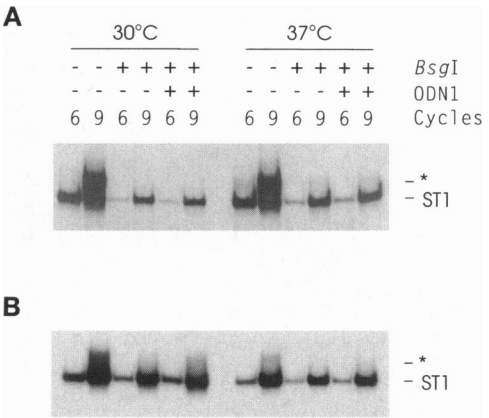

FIG. 3. Identification of an emergent *Bsg* I-cleavage-resistant population by semiquantitative PCR and nondenaturing PAGE. Shown are the autoradiograms of PCR products obtained by using *Bsg* I cleavage-selected ST1 DNA and the RPL and RPR amplimers. Cycles, number of PCR temperature cycles used. The locations of the bands corresponding to properly annealed, duplex ST1 DNA (ST1) and incompletely annealed or "bubble"-containing ST1 DNA (*) are indicated. (*A*) DNA products after one round of REPSA. (*B*) DNA products after 11 rounds of REPSA.

by a restriction endonuclease protection assay (REPA) (13–15). Here ODN1-dependent protection from *Bsg* I cleavage indicated the ability of a particular sequence to form a triplex with ODN1. The results of a series of representative assays are shown in Fig. 4, with clone G36 demonstrating a pattern consistent with triplex formation. DNA fragments from 55 colonies from each of the 30°C and 37°C selections and 25 colonies from the minimum amplification control (ODN$^-$, *Bsg* I$^+$) were assayed for their ability to form a triplex with ODN1: 26 and 21 fragments from the 30°C and 37°C pools, respectively, demonstrated REPA patterns consistent with triplex formation, whereas none of the control fragments did so. Two other patterns, including DNA fragments with reduced electrophoretic mobility (clone E5) and a mixture of cleavage products (clone E7), were found to occur independently of added ODN1. These contained sequences selected as a result of using *Bsg* I in our REPSA experiments (see below).

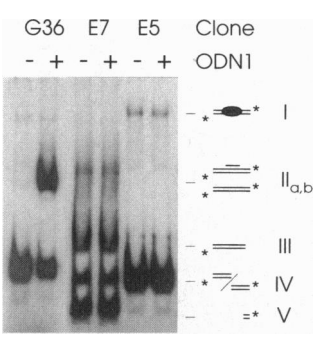

FIG. 4. Analysis of REPSA-selected clones by *Bsg* I cleavage protection and nondenaturing PAGE. Shown is an autoradiogram of the reaction products after incubation with *Bsg* I, and ODN1 as indicated, for three representative ST1 clones. Relative electrophoretic mobilities and schematic representations of the resulting DNA species are indicated at right of figure. Asterisks indicate sites of radiolabeling for these DNA fragments. Diagrams: I, ST1 probe with reduced electrophoretic mobility, presumably resulting from a specific protein–DNA complex; IIa, ST1 probe in a triplex with ODN1; IIb, ST1 fragment alone; III–V, *Bsg* I cleavage products of ST1; IV, cleavage in the center of ST1 resulting from *Bsg* I bound to sites flanking the random cassette; III and V, large and small ST1 fragments, respectively, resulting from *Bsg* I bound to a site within the random cassette and subsequent cleavage in the flanking sequences.

Table 1. Alignment of triplex-forming sequences

Clone		Reference sequences																			
		Sequence (5′ → 3′)																			
Purine strand		**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**A**	
		30°C selection																			
E11		C	A	G	G	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	A	C	C	AGAA
E12	ATTT	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	T	T	C	C	C	T	G	
E17	T	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	T	T	C	A	**G**	C	
E20		T	T	T	C	C	C	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	A	**A**	
E22	TTA	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	T	A	G	A	C	C	T	
E23	CAG	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	T	**G**	C	C	T	
E24	CT	T	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	T	C	C	
F1		A	G	G	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	C	C	C	T	C	T	A	
F2		G	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	T	A	G	TA
F7	TC	**A**	**G**	**G**	**G**	C	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	C	A	T	C	C	
F9	T	C	A	**G**	**G**	G	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	T	C	C	C	
F12		C	A	**G**	**G**	G	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	C	T	A	A	C	
F13	GTG	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	T	A	G	C	C	C	T	
F16	CAT	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	C	T	C	A	C	C	T	
F19		G	**G**	A	T	**A**	A	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**A**	TA
F20	T	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	C	C	C	T	C	C	
F21	TCTGGT	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	C	C	T	**G**	C	A	G	
F22	T	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	C	T	C	T	A	C	
F24	TACT	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	C	C	C	T	G	
F27		**A**	A	**G**	**G**	G	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	T	C	A	G	
F28	TAT	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	T	**G**	C	C	T	
F29	ACA	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	A	A	T	T	C	T	C	
F32	G	T	T	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	G	A	T	T	C	
F35	AG	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	T	A	T	A	A	C	T	
F36	AG	C	T	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	G	T	C	C	T	
F38		A	G	A	G	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	G	TATG
		37°C selection																			
E26	T	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	G	C	C	**G**	C	
E27	T	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	T	T	A	C	C	
E32	GA	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	T	T	**G**	T	C	C	
E37		G	G	C	A	T	A	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**A**	CA
G1		C	A	G	G	G	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	G	T	T	A	**A**	AGA
G3		**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	A	C	**A**	T	T	T	C	
G8		A	G	G	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	C	C	T	C	C	T	C	
G10		A	G	G	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	A	C	C	C	C	C	T	
G11	GCAG	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	G	**G**	**G**	**G**	A	G	C	T	C	T	
G12	ACAA	**A**	**G**	**G**	**G**	C	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	T	C	T	C	T	
G13		**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	G	A	**G**	**G**	G	ATG
G20		T	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	T	A	**G**	A	C	
G21		A	G	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	T	C	T	A	G	TC
G22		T	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	C	T	A	A	T	
G23		**A**	**G**	A	**G**	**A**	C	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**A**	GGTT
G26	TT	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	A	T	G	**G**	A	C	C	
G28	A	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	T	C	C	A	T	C	
G29	TT	**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	C	T	C	C	C	C	
G33		G	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	G	C	A	C	T	
G34		A	G	G	**G**	C	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	G	C	C	A	**A**	AAA
G36		**A**	T	C	**G**	**A**	A	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**A**	
A, no.		26	3	1	1	37	3	0	0	0	46	0	0	4	5	12	7	9	9	7	
C, no.		3	0	2	1	5	2	0	0	0	0	0	0	3	7	9	12	5	4	3	
G, no.		0	28	36	40	4	42	47	47	47	1	47	47	34	28	11	13	8	8	5	
T, no.		5	4	1	1	1	0	0	0	0	0	0	0	5	6	13	9	10	3	1	
Total no.		34	35	40	43	47	47	47	47	47	47	47	47	46	46	45	41	36	24	16	
Consensus		**A**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	-	-	-	-	-	

Bases shown in boldface type are identical to the presumed consensus target sequence (purine strand) for purine-motif triplexes containing ODN1. Bases present in the defined flanks bordering the random cassette are underlined. Bases that were part of the random cassette but did not align with the consensus sequence are shown extending on either side of the aligned central region. The consensus sequence was determined by comparison of selected base distribution (excluding bases originally from flanking sequences, underlined) to the starting distribution of bases using a χ^2 analysis. Bases with a significantly higher than chance representation ($P < 0.05$) are listed as consensus.

Triplex-Forming Sequences. The sequences of the 47 triplex-positive fragments were determined by dideoxynucleotide sequencing (Table 1). All possessed substantial homology to the sequence $AG_3AG_4AG_4AG_3A$, the binding site based on

the known rules of purine-motif triplex formation (16, 17, 21). These sequences were aligned to give the longest continuous stretch of bases homologous to this preferred sequence (Table 1, purine-rich strand shown). No difference in consensus sequence, average length of homology, or base distribution was observed between sequences obtained through the selection at 30°C from those selected at 37°C; thus both sets of sequences were considered as one population. To derive a consensus binding sequence, the number of each nucleotide at each of the 19 positions was tabulated and compared to the starting distribution of nucleotides by a χ^2 analysis. Nucleotides in positions 1–14 were found to be identical to the sequence 5′-$AG_3AG_4AG_4$-3′; however, no consensus emerged for positions 15–19. Given the antiparallel binding orientation of purine-motif triplexes, this data suggests that complementarity between nucleotides on the 3′ end of the third strand and the 5′ end of the purine-rich strand of the duplex facilitate triplex formation to a greater extent than potential interactions at the other end of the triple helix.

To investigate the length of complementarity required for purine-motif triplex formation, the length of homology of each sequence to the presumed consensus sequence $AG_3AG_4AG_4AG_3A$ was determined, with an average of 13 bases of homology to the consensus found. To verify this number, affinities for sequences with 12, 13, 14, and 15 bases of homology (clones F35, F22, E27, and F2, respectively) were determined. The equilibrium dissociation constants for sequences with 13, 14, and 15 bases of homology were 5 to 8 $\times$ 10^{-8} M, while the constant for the sequence with 12 bases of homology was substantially higher (1 $\times$ 10^{-6} M), indicating a drop in affinity from 13 to 12 bases of homology. No difference was observed for sequences selected at either 30°C or 37°C, suggesting that a higher range of temperatures and perhaps a heat-stable endonuclease might be required to detect a relationship between temperature and minimum triplex length.

Clearly, the G·GC and T·AT base triplets were the consensus base triplets. However, other base triplets were found within consensus triplex-forming sequences. Comparing the frequency of these base triplets to the starting distribution of bases allowed us to determine whether a selective pressure favored any of these nonconsensus base triplets. For our analysis, mismatches were defined as single nonconsensus bases flanked by consensus bases. Of the 11 T·NN mismatches present, two T·TA, four T·GC, and five T·CG mismatches were found, suggesting that no selective pressure favored triplets involving a third-strand T. However, of the six G·NN mismatches present, one G·CG, no G·TA, and five G·AT mismatches were found. Similar results were also found with an independent REPSA experiment where the starting nucleotide concentrations were T, A > G, C (data not shown). Thus these data suggest that the G·AT mismatch emerged due to a selective pressure, e.g., by stabilizing the triplex through the formation of a weak G·A hydrogen bond or through stacking interactions, allowing neighboring bases to better engage in triplex formation.

Other Consensus Sequences. In addition to the REPA pattern indicating triplex formation, two other patterns were observed. Of the 110 colonies assayed, 46 had REPA patterns exhibiting multiple cleavage products (Fig. 4, clone E7). Nine representatives from this group were sequenced and found to contain a third *Bsg* I recognition site in the 3′ end of the random cassette region (Table 2). Here the consensus sequence was 5′-AGTGCAGT-3′, the defined 6-bp *Bsg* I recognition sequence with a 5′ flanking A and a 3′ flanking T. The emergence of this consensus sequence, and its resulting characteristic REPA pattern, can be explained by a *Bsg* I enzyme preferentially occupying this third site. Cleavage by the two *Bsg* I enzymes bound to the ST1 flanks could be prevented by a *Bsg* I protein bound to a kinetically favored site within the randomized cassette. Though the template was cleaved by the third *Bsg* I, the resulting large ST1 fragment would still be capable of annealing to the RPL amplimer under our PCR conditions, thus the observed *Bsg* I-binding sequences were selected.

Table 2. Alignment of *Bsg* I binding sequences

Clone	Reference sequences: Sequence (5′ → 3′)
(*Bsg* I)	**G T G C A G**
E3	T T T C T A A C C G T A **G T G C A G** T
E7	T C C T A C G A G T T A **G T G C A G** T
F26	T G T A A A A A A A A A **G T G C A G** T
I7	T A T T G G C T T A C A **G T G C A G** A
I8	G G A T A C T C G T T A **G T G C A G** T
I10	A T A G G C A A A T T A **G T G C A G** T
I11	A T A T T A G T G A T A **G T G C A G** T
I13	T T T T C G C C T G T A **G T G C A G** T
I15	T A T T T C T T A T T A **G T G C A G** A
A, no.	2 2 3 1 3 3 3 3 3 3 1 9 0 0 0 0 9 0 2
C, no.	0 1 1 1 1 4 2 3 1 0 1 0 0 0 0 9 0 0 0
G, no.	1 2 0 1 2 2 2 0 3 2 0 0 9 0 9 0 0 9 0
T, no.	6 4 5 6 3 0 2 3 2 4 7 0 0 9 0 0 0 0 7
Consensus	- - - - - - - - - - - - - **A G T G C A G T**

Bases shown in boldface type are identical to the previously described *Bsg* I recognition sequence, shown. The consensus sequence was determined by comparison of selected base distribution to the starting distribution of bases by using a χ^2 analysis. Bases with a significantly higher than chance representation ($P < 0.05$) are listed as consensus.

In three of the 110 colonies screened, a REPA pattern characterized by a low-mobility ODN1-independent *Bsg* I-cleavage-resistant band was observed (Fig. 4, clone E5). These clones were sequenced and found to contain a 5-bp inverted repeat (Table 3). Possible explanations for the selection of this consensus sequence include the formation of a cruciform structure resistant to *Bsg* I cleavage or binding of a previously unidentified protein present in the *Bsg* I protein fraction. Either could account for the shift in mobility seen in the REPA assay; however, a DNA-binding protein contaminant is more likely the cause for two reasons: this mobility shift was seen only upon the addition of *Bsg* I and not after any other REPSA steps (data not shown), and cruciform structures are thought to form in response to helical stress, a condition not present in our short DNA fragments (22). The consensus binding sequence for this putative *Bacillus sphaericus* protein does not match the consensus sequence of any known prokaryotic

Table 3. Alignment of *Bacillus sphaericus* protein–DNA binding sequences

Clone	Reference sequences: Sequence (5′ → 3′)
Inverted repeat	**T G G G A** **T C C C A**
E5	**T G G G A** C T T T T A T T G **T C C C A**
G7	**T G G G A** T A A T C T C G G **T C C C A**
I3	**T G G G A** T A G G A A T T G **T C C C A**
A, no.	0 0 0 0 3 0 2 1 0 1 2 0 0 0 0 0 0 0 3
C, no.	0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 3 3 3 0
G, no.	0 3 3 3 0 0 0 1 1 0 0 0 0 1 3 0 0 0 0 0
T, no.	3 0 0 0 0 2 1 1 2 1 1 2 2 0 3 0 0 0 0
Consensus	**T G G G A** - - - - - - - - - **G T C C C A**

Bases shown in boldface type are identical to the 5-bp inverted repeat sequence shown above.

DNA-binding protein (23), and though it is reminiscent of a split restriction-modification recognition sequence, no endonuclease activities other than *Bsg* I or any *Bsg* I methylase activity were detected in this fraction (M. J. McMahon, New England Biolabs, personal communication).

DISCUSSION

We have devised a technique, termed REPSA, that allows the isolation of a population of duplex DNAs capable of sequence-specific interactions with a particular ligand. Through the present work, we demonstrate the utility and flexibility of this method by examining triplex binding specificity, as well as serendipitously determining binding sequences for ligands present in a protein mixture. Advantages of REPSA stem from the use of restriction endonuclease cleavage protection to achieve enrichment of desired sequences, a method that does not rely on the physical separation of complexed from uncomplexed DNA. Ligand–DNA complexes that may not be sufficiently stable to allow their partitioning or lack a readily available means for their physical isolation become amenable to investigation using REPSA. The former may be demonstrated with our identification of triplex-forming sequences; though other techniques have been used in their determination (24, 25), REPSA allowed identification of weak binding sites with triplex binding constants as high as 10^{-6} M. Similarly, as demonstrated by our identification of two consensus sequences for ligands present in a commercial preparation of the restriction endonuclease *Bsg* I, this technique may be used to identify the preferred binding sites of ligands about which very little is known. Generally, if a ligand–DNA interaction can be investigated by a nuclease cleavage assay (e.g., DNase I footprinting), it should be possible to determine its consensus binding sequence by REPSA.

While the identification of optimal oligonucleotide sequences capable of recognizing a prescribed duplex target is of greater practical concern (e.g., in the development of triplex-based gene therapeutics), our data on the preferred targets for triplex formation by the G/T-rich oligonucleotide ODN1 should aid in defining the parameters governing purine-motif triple-helical interactions. The consensus triplex binding sequence that emerged corresponded to only the 5′ end of the binding site, which complements our prior observations that modifications to the 3′ end of ODN1 decreased triplex formation to a greater degree than changes in the 5′ end (26). Because the 3′ end of ODN1 binds the 5′ end of the binding site, the two studies indicate that this asymmetry in binding is not a property of the third-strand or duplex alone, but rather an intrinsic property of purine-motif triplex formation. Indications of a G·AT base triplet within a purine triple helix, while not observed in earlier published reports (21), are consistent with recent work demonstrating that the G·AT base triplet has the lowest free energy of formation of any of the other G·NN mismatches (27). An average of 13-base triplets was required for the selection of target sequences, which is similar to our prior determination that 12-base deletion mutants of ODN1 were kinetically better able to form triplexes than ones that were 14, 15, or 19 nt long (26). This finding was independently confirmed after determining the binding affinities of several representative sequences. Those having 13, 14, and 15 bases of homology to the consensus had similar apparent affinities; however, one with 12 homologous bases had a 50-fold lower affinity. Thus these data suggest that at least 12 bases, whether defined by the binding site or the third strand, are sufficient for defining a purine-motif triple helical interaction.

The number of rounds necessary to select a desired sequence depends on the efficiency of the enzyme cleavage and the representation of the sequence in the starting pool. For the REPSA selection reported here, an enrichment level of roughly 10-fold was observed, as estimated by the levels of amplified DNA in the triplex selection reaction mixture compared to a control amplification. While this enrichment is lower than obtained by other methods, the burden of performing additional rounds with REPSA is compensated by the relative ease of performing each round. Also, 34 other type IIS restriction endonucleases are presently commercially available (19). These could be used singly or in combination to optimize selection efficiency under a variety of conditions and to prevent the emergence of artifactual sequences (e.g., restriction endonuclease recognition sites). In the final analysis, REPSA was capable of determining consensus sequences of 8, 10, and 14 bp for various protein- or nucleic acid-based ligands, thus, demonstrating the utility of this method in investigating complex binding sites.

We thank William E. Jack and Michèle Sawadogo for critical reading of the manuscript. This work was supported by grants from the American Cancer Society (NP-839) and the Welch Foundation (G-1199). P.H. was supported by a National Research Award predoctoral traineeship from the National Cancer Institute (CA60440).

1. Szostack, J. W. (1992) *Trends Biochem. Sci.* **17,** 89–93.
2. Wright, W. E. & Funk, W. D. (1993) *Trends Biochem. Sci.* **18,** 77–80.
3. Kenan, D. J., Tsai, D. E. & Keene, J. D. (1994) *Trends Biochem. Sci.* **19,** 57–64.
4. Thieson, H.-J. & Bath, C. (1990) *Nucleic Acids Res.* **18,** 3203–3209.
5. Polloc, R. & Treisman, R. (1990) *Nucleic Acids Res.* **18,** 6197–6204.
6. Mavrothalassitis, G., Beal, G. & Papas, T. S. (1990) *DNA Cell Biol.* **9,** 783–788.
7. Blackwell, T. K. & Weintraub, H. (1990) *Science* **250,** 1104–1110.
8. Tuerk, C. & Gold, L. (1990) *Science* **249,** 505–510.
9. Tsai, D. E., Harper, D. S. & Keene, J. D. (1991) *Nucleic Acids Res.* **19,** 4931–4936.
10. Ellington, A. D. & Szostack, J. W. (1990) *Nature (London)* **346,** 818–822.
11. Ellington, A. D. & Szostack, J. W. (1992) *Nature (London)* **355,** 850–852.
12. Pei, D., Ulrich, H. D. & Schultz, P. G. (1991) *Science* **245,** 1408–1411.
13. Hanvey, J. C., Shimizu, M. & Wells, R. D. (1989) *Nucleic Acids Res.* **18,** 157–161.
14. Maher, L. J., Wold, B. & Dervan, P. B. (1989) *Science* **245,** 725–730.
15. François, J., Saison-Behmoaras, T., Thuong, N. T. & Hélène, C. (1989) *Biochemistry* **28,** 9617–9619.
16. Beal, P. A. & Dervan, P. B. (1991) *Science* **251,** 1360–1363.
17. Durland, R. H., Kessler, D. J., Gunnell, S., Duvic, M., Pettit, B. M. & Hogan, M. E. (1991) *Biochemistry* **30,** 9246–9255.
18. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd Ed.
19. Szybalski, W., Kim, S. C., Hasan, N. & Podnajska, A. J. (1991) *Gene* **100,** 13–26.
20. Singleton, S. F. & Dervan, P. B. (1992) *J. Am. Chem. Soc.* **114,** 6957–6965.
21. Beal, P. A. & Dervan, P. B. (1992) *Nucleic Acids Res.* **20,** 2773–2776.
22. Mizuuchi, K., Mizuuchi, M. & Gellert, M. (1982) *J. Mol. Biol.* **156,** 229–243.
23. Bickle, T. A. & Kruger, D. H. (1993) *Microbiol. Rev.* **57,** 434–450.
24. Ito, T., Smith, C. L. & Cantor, C. R. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 495–498.
25. Nishikawa, N., Oishi, M. & Kiyama, R. (1995) *J. Biol. Chem.* **270,** 9258–9264.
26. Cheng, A.-J. & Van Dyke, M. W. (1994) *Nucleic Acids Res.* **22,** 4742–4747.
27. Greenberg, W. A. & Dervan, P. B. (1995) *J. Am. Chem. Soc.* **117,** 5016–5022.