Identification of preferred hTBP DNA binding sites by the combinatorial method REPSA

Paul Hardenbol+, Jo C. Wang and Michael W. Van Dyke*

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

Received April 1, 1997; Revised and Accepted June 26, 1997

ABSTRACT

Here we describe the application of a novel combinatorial method, restriction endonuclease protection selection and amplification (REPSA), to identification of a consensus DNA binding site for the TATA binding subunit (hTBP) of the human general transcription factor TFIID. Unlike most combinatorial methods, REPSA is based on inhibition of an enzymatic template inactivation process by specific ligand–DNA complexes. The mild conditions of this method allow examination of proteins with atypical binding characteristics (e.g. limited discrimination between specific and non-specific binding sites), such as those found with hTBP. Analysis of 57 emergent sequences identified 47 sequences containing consensus 6 bp TATA elements as previously defined. However, further examination of these sequences indicated that a larger consensus, 5′-TATAAATA-3′, could be supported by the data. Studies of the binding affinities and transcriptional activities of these four consensus TATA sequences demonstrated that hTBP binding affinity correlated directly with transcriptional activity *in vitro* and that the TATAAATA sequence was the best among the TATA sequences investigated.

INTRODUCTION

Combinatorial approaches have emerged as the preferred method for determining the consensus binding sequences for DNA binding ligands (1–3). As exemplified by CASTing (cyclical amplification and selection of targets), these methods are characterized by selection of ligand binding sequences from a pool of random sequences, amplification of selected sequences and repetition of this process to enrich for ligand-bound sequences. Using these techniques, consensus DNA binding sites for RNA (4), small molecules (5) and DNA (6) ligands have been determined. These methods have been most frequently used, however, to identify consensus sequences for DNA binding proteins, owing in part to their large number and their biological importance in processes such as transcription and replication. All combinatorial protocols employ a selection step to separate ligand-bound DNA from unbound DNA. Selection methods used in identifying protein binding sequences have included electrophoretic mobility shift assays (7,8), filter binding (9), immunoprecipitation (10,11) and matrix immobilized proteins (12). Each of these selection methods relies on a physical separation of protein-bound DNA from unbound DNA as the means of isolating desired sequences.

We have developed a combinatorial approach based on an enzymatic selection process, restriction endonuclease protection selection amplification (REPSA), which was used to determine the consensus DNA binding sequence of a triplex-forming oligodeoxyribonucleotide (6). This method relies on a type IIS restriction endonuclease (IISRE), a class of nucleases that cleave DNA without regard to sequence at a specific distance from its recognition sequence, to selectively cleave unbound DNA while triplex-bound DNA is protected from cleavage. Because this selection occurs in solution under mild conditions, DNA–triplex interactions with binding constants as weak as 10^{-6} M were found. In addition to the desired triplex consensus, serendipitous consensus sequences also emerged for DNA binding proteins present in the endonuclease fraction used in the selection. This observation suggested that REPSA could be used to determine the consensus binding site of DNA binding proteins under physiological conditions.

As a test of REPSA capabilities, we applied it to determine the consensus DNA binding sequence of the human TATA binding protein hTBP. TBP, as part of the holoTFIID complex, plays a critical role in transcription of class II genes through its sequestration at the gene promoter and its nucleation of preinitiation complex assembly (reviewed in 13). In many cases this involves direct recognition of a TATA sequence by TBP. Crystal structures of the human C-terminal/core TBP complexed with either TATAAAAG or TATATATA sequences have recently been described (14,15). As found for TBPs from *Saccharomyces cerevisiae* and *Arabidopsis thaliana*, hTBP possesses a positively charged, saddle-shaped convex surface that forms minor groove contacts with an 8 bp stretch of DNA, resulting in a pronounced local unwinding of the DNA and a 100° bend. These characteristics make TBP unique among eukaryotic DNA binding proteins.

TBP has been shown to bind several TATA boxes that together do not conform to a simple consensus sequence. A consensus of TATA@A@ (where @ signifies A or T) has been determined from a comparative sequence analysis of over 500 class II gene promoters (16). However, TBP has also been found to bind to several native sequences that vary greatly from the consensus (17–19). In addition to its promiscuous DNA binding pattern,

*To whom correspondence should be addressed. Tel: +713 792 8954; Fax: +713 794 4784; Email: mishko@odin.mdacc.tmc.edu

+Present address: Department of Biochemistry, Stanford University, Stanford, CA 94305-5307, USA

TBP also exhibits slow DNA binding kinetics and requires higher temperatures to bind DNA relative to other DNA binding proteins (13). Owing to these characteristics, hTBP was considered a challenging test of the ability of REPSA to determine the consensus sequence of a DNA binding protein.

MATERIALS AND METHODS

Oligonucleotides

Phosphodiester oligodeoxyribonucleotides were prepared on a Millipore Cyclone DNA synthesizer. The nucleotide sequences of oligonucleotides used in this study were (5′→3′): 63AL, CTAGGAATTCGTGCAGAGGTGA; 63AR, GTCCAAGCTT-CTGGAGGGATG; 63R14, CTAGGAATTCGTGCAGAGGT-GA$(N)_{14}$TTACCATCCCTCCAGAAGCTTGGAC; MS5, TGT-TGTGTGGAATTGTG; MS6, CAAGGCGATTAAGTTGG. For oligonucleotide 63R14 the sites containing mixed bases (N) were synthesized using an equimolar mixture of each phosphoramidite. The distribution of nucleotides incorporated into the random cassette was 27% A, 18% C, 16% G and 39% T, as determined by sequencing of eight individual clones derived from the starting material.

Preparation of hTBP

Recombinant hTBP was expressed in *Escherichia coli* and purified essentially as previously described (20). The concentration of active hTBP present in the final Mono S fraction was estimated to be 5 µM, as determined by a restriction endonuclease protection assay in the presence of TATA-containing oligonucleotides in vast excess of the expected TBP dissociation constant (17,21).

REPSA

The double-stranded selection template ST2 was synthesized by four rounds of PCR using oligonucleotide 63R14 as template and 63AL and 63AR as amplimers. To effect TBP binding, 2 ng ST2 were incubated with 1.2 pmol hTBP in a 10 µl volume containing 40 mM HEPES–NaOH, pH 8.4, 6 mM $MgCl_2$, 50 mM KCl, 10% glycerol, 0.05% Nonidet P-40, 1 mM dithiothreitol (binding buffer) and 1 µg poly(dG·dC) for 30 min at 30°C. Following TBP binding, either 3 U *Bpm*I, 2 U *Bsg*I or 2 U *Fok*I (New England Biolabs) in a 3 µl volume containing the appropriate reaction buffer was added and the incubation continued for an additional 30 min. To amplify the cleavage-resistant duplex DNA subpopulation, 200 ng each of 63AL and 63AR, 5 U *Taq* DNA polymerase, 0.25 mM dATP, dCTP, dGTP and dTTP, 10 mM Tris–HCl, pH 9.0, 50 mM KCl, 1 mM $MgCl_2$ and 2 µCi [α-^{32}P]dATP (1 Ci = 37 GBq) were added to each sample, to a final 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. Following PCR amplification, 2 µl of each reaction mixture were analyzed by PAGE and autoradiography to determine relative levels of amplification. The balance of each mixture was phenol extracted and the aqueous phase concentrated on a Millipore Ultrafree-MC 5000 cellulose spin filter by centrifugation for 30 min at 15 000 *g*. Filters were washed for 10 min with 100 µl 10 mM Tris–HCl, pH 7.4, 1 mM EDTA and centrifuged for 30 min. The retained template DNA was resuspended in 20 µl Tris–EDTA. These steps, TBP binding, enzyme cleavage, PCR amplification and filter purification, were repeated for a total of 10 times.

Sequence determination and statistical analysis

The finally selected ST2 templates were digested with *Eco*RI and *Hin*dIII and cloned into similarly cut plasmid pUC19. Individual colonies were used to inoculate 5 ml overnight cultures in Luria broth medium containing 0.2 mg/ml ampicillin. Mini-plasmid preparations were made from the clones and their inserts sequenced by Sanger enzymatic sequencing.

The significance of differences in experimentally determined consensus sequences was determined by a χ^2 comparison of distributions in consensus sequences to the nucleotide distribution present in the total population of sequences isolated after the final REPSA round. $P < 0.05$ was considered significant. In the TBP selection the final nucleotide distribution was 32.3% A, 16.8% C, 13.8% G and 37.1% T from a total of 797 nt sequenced.

Binding affinity determination

The binding affinity of hTBP to probes containing different consensus TATA sequences was determined by a IISRE cleavage protection assay (6,21). Radiolabeled DNA fragments were generated by PCR amplification of clones L23, L36, L16 and K19, containing TATA sequences TATAAATA, TATAAA, TAAATA and TATATA respectively, using 5′-end-labeled primer MS5 and unlabeled primer MS6. To effect DNA binding by TBP, 1 fmol gel-purified, labeled probe DNA, 1 µg poly(dG·dC) non-specific carrier DNA and various concentrations (0–20 nM, as indicated in the figure legends) of unlabeled competitor DNA containing the sequence TATAAATA were incubated with 0.8 pmol hTBP in 10 µl binding buffer for 30 min at 30°C. Following TBP binding, 0.3 U *Bpm*I were added and the incubation continued for an additional 5 min at 37°C. Reaction products were analyzed by non-denaturing PAGE and quantitated using a Storm 840 phosphoimager (Molecular Dynamics). The apparent binding affinity was considered to be the concentration of DNA required to give 50% maximal cleavage protection by bound TBP. For example, if 61% probe cleavage was observed in the absence of TBP and 40% cleavage was observed when TBP and only the labeled probe DNA were present, then the concentration of unlabeled competitor DNA allowing 51% probe cleavage would correspond to the apparent binding constant.

Transcription

Transcription templates were constructed that contained the TATA box sequences TATAAATA, TATAAA, TAAATA and TATATA cloned upstream of a G-less cassette (22). In each case the identical flanking sequences were present and the initial T of the TATA element was located 31 bp upstream of the same initiation site. Given the different lengths of G-less cassette used in these constructs, the expected transcript lengths were 347 (TATAAATA) and 377 nt (TATAAA, TAAATA and TATATA). A similar template containing the adenovirus-2 major late core promoter (from –40 to +10), which yields a 388 nt transcript, was also used for comparison purposes.

In vitro transcription assays were performed essentially as previously described (22,23). Briefly, 25 µl reaction mixtures containing binding buffer and 0.6 mM ATP, 0.6 mM CTP, 0.1 mM 3′-*O*-methyl GTP, 0.025 mM UTP, 13 µCi [α-^{32}P]UTP at

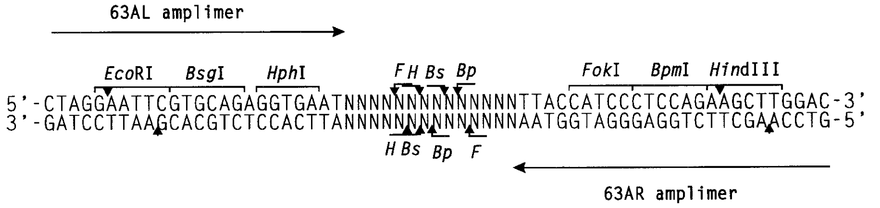
63AL amplimer
EcoRI
BsgI
HphI
F H Bs Bp
FokI
BpmI
HindIII
5'-CTAGGAATTCGTGCAGAGGTGAATNNNNNNNNNNNNNNTTACCATCCCTCCAGAAGCTTGGAC-3'
3'-GATCCTTAAGCACGTCTCCACTTANNNNNNNNNNNNNNAATGGTAGGGAGGTCTTCGAACCTG-5'
H Bs Bp F
63AR amplimer

Figure 1. Schematic of the selection template ST2 used for the identification of duplex DNA sequences that preferentially bind hTBP. Locations of restriction endonuclease binding (brackets) and cleavage (arrows) sites are indicated. Long horizontal arrows correspond to the sequences of the PCR amplimers 63AL and 63AR. N, random nucleotides.

700 Ci/mmol, 5 U RNase T1 (Calbiochem), 8 U rRNasin (Promega), 0.1 pmol hTFIIB, 0.025 pmol hTBP, 0.2 µl RNA polymerase II (10 U, phosphocellulose fraction), 2 µl Bio-Gel A1.5 fraction containing at least 2.4 U each TFIIE, TFIIF and TFIIH (22) and 0.2 pmol total supercoiled plasmid template DNA were incubated for 30 min at 30°C and then processed by standard protocols. RNA products were resolved by denaturing PAGE and visualized by autoradiography. Quantitation was by direct β counting using a Betascope 600 (Betascan).

RESULTS

REPSA selection template design

A REPSA selection template should be designed to contain a central region of random sequence nucleotides of sufficient length to provide an adequate binding site for the protein investigated and defined flanks suitable for PCR amplification that contain type IIS and conventional restriction endonuclease recognition sites, for selection and subcloning purposes respectively. In REPSA, selection arises as a result of protection from type IIS restriction endonuclease (IISRE) cleavage within the cassette by a bound ligand for a subpopulation of the sequences present. The 63 bp selection template used here, ST2, contained a 14 bp region of random sequence and defined flanks having nested IISRE binding sites (either *Bsg*I and *Hph*I or *Bpm*I and *Fok*I) and terminal *Eco*RI or *Hin*dIII cleavage sites. ST2 is shown schematically in Figure 1. Locations of the IISRE cleavage sites within the randomized cassette are indicated. The incorporation of multiple, different IISRE recognition sites in this selection template allowed substitution of different IISREs in different selection rounds. This substitution was done in order to minimize selection of sequences recognized by proteins present in any one enzyme preparation, as was found during selection of purine motif triplex-forming sequences with the IISRE *Bsg*I (6).

Selection of DNAs binding TBP

A flow diagram for the REPSA selection of hTBP binding sequences is shown in Figure 2. In the first round of selection, 48 fmol double-stranded ST2 was incubated with a 25-fold molar excess of hTBP under conditions favorable for TBP binding. The amount of selection template chosen provided a good over-representation of all possible 14 bp sequences (i.e. 2.9×10^{10} template molecules >> $4^{14} = 2.7 \times 10^{8}$ different sequences possible), while the amount of hTBP used provided a final concentration that was greater than its expected dissociation constant (10^{-9} M; 17) but did not non-specifically interfere with IISRE cleavage. Two units of *Fok*I, sufficient for cleaving >95% of the unbound template under these conditions, were then added and the reaction mixture incubated for an additional 30 min. After endonuclease challenge, the mixture was heated for 5 min at 90°C to denature both hTBP and the selection enzyme. PCR buffers, primers and *Taq* polymerase were then added and intact DNA amplified for nine cycles. Afterwards the amplified selection templates were purified from unused primers, IISRE cleavage products, proteins, nucleotides, buffers and salts by spin filtration and then subjected to additional rounds of REPSA as described above. In subsequent rounds the IISRE selection enzymes used were *Bsg*I (round 2), *Fok*I (rounds 3–6), *Bpm*I (rounds 7–9) and *Fok*I (round 10). Cleavage efficiency of *Bsg*I and *Hph*I was found to be lower than *Fok*I and *Bpm*I under optimal hTBP binding conditions, thus the latter two enzymes were predominately used during the course of this selection.

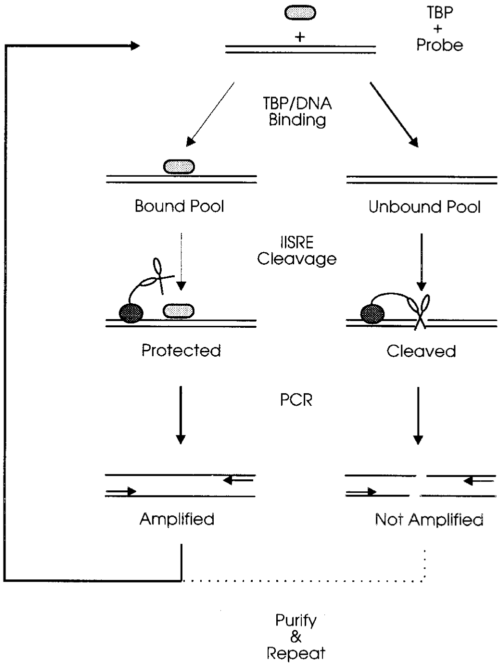

Figure 2. Flow diagram of the combinatorial method restriction endonuclease protection selection and amplification, REPSA. Light gray oval, the human TATA box binding protein hTBP; dark gray oval with attached scissors, DNA binding and cleaving moieties of a type IIS restriction endonuclease (IISRE) respectively.

TBP-selected sequences

The insert sequences of 75 clones were determined by dideoxy sequencing. Eighteen clones contained templates in which the random cassette had been deleted. Forty five clones with intact templates contained sequences with substantial homologies to one of three 6 bp TATA box sequences as defined previously by saturation point mutagenesis: TATAAA, TAAATA and TATATA (18,24,25). These data are shown in Table 1. Notably, 11 sequences appeared in both the TATAAA and TAAATA groups. Twelve template sequences did not fit one of these three patterns. To determine the consensus sequences of each of the three groups of selected sequences, the relative frequency of each nucleotide at each position was determined. In the case of the sequence TATATA, its palindromic nature complicated this analysis and no larger consensus was derived. Remarkably, the consensus sequences for both the TATAAA and the TAAATA groups both contained the 8 bp sequence TATAAATA, suggesting that the two sequences TATAAA and TAAATA were part of a larger consensus sequence. For this reason, the sequences of these groups were combined and the relative frequency of each nucleotide at each position determined (Table 2). From a χ^2 analysis, the frequency at which the extended TATA sequence appeared was found to be extremely significant ($P < 0.01$), given that it appeared in four of the 45 selected sequences (K18, L23, P44 and P52).

Table 1. REPSA/TBP sequence[a]

Clone	Sequence
(A)	**TATAAA**
K16	**GAAT**ATAAAGC
K18[b]	**TAA**TATAAATA
K20	CGCTATAAAAG
K23	**GAAT**ATAAGTT
K30[b]	**AA**ATATAATCA
K33	TGATATAAAAG
K36	**T**AATATAAAAG
L10	**GAAT**ATAAAGT
L14[b]	AGCTATAAATT
L23[b]	**AT**GTATAAATA
L36	CCTTATAAACG
L39	**A**GATATAAACA
P16[b]	**GAAT**ATAAGGT
P20[b]	TAATATAACGC
P32[b]	TAGTATAA**ATT**
P35	**TAA**TATAAAGA
P41[b]	CCATATAA**ATT**
P43	GCTTATAAGT**T**
P44[b]	CGGTATAAATA
P47	**AA**GTATAAGGA
P52[b]	TAGTATAAATA
P56[b]	CCCTATAATTT
	CGA**TATAAATA**
(B)	**TAAATA**
K18[b]	**A**TATAAATACA
K26	GGATAA**ATTCA**
K30*	**TGGTAA**ATATA
K34	**A**GATAAATAGG
L4	ATTTAAATCCA
L7	ACTTAAATTAC
L14[b]	CTATAA**ATTCA**
L15	GCTTAAA**TTAC**
L16	CCTTAAATACA
L20	**TGGTAA**ATATA
L22	TCGTAAATA**TT**
L23[b]	GTATAAATACG
L35	TAATAAAT**ATT**
L40	**TGGTAA**ATAAG
P16[b]	**TGGTAA**ATACA
P20[b]	TATTAA**ATTCA**
P27	**A**GATAAATAGA
P30	TTTTAAATCA**A**
P32[b]	GTATAA**ATTCA**
P41[b]	ATATAA**ATTCA**
P44[b]	GTATAAATA**TT**
P48	**A**GTTAAATTCC
P52[b]	GTATAAATACG
P56[b]	**AA**ATAAATTAT
P58	**TGGTAA**ATTAC
	T**TATAAATA**CA
(C)	**TATATA**
K19	GGCTATATAC**T**
K21	AATTATATAGG
K25	CATTATATACC
K27	**GAAT**ATATACC
K32	**GAAT**ATATACC
L18	CTATATATACA
L19	AAATATAT**ATT**
L24	**GAAT**ATATAAG
L25	**TAA**TATATATT
	TATATA
(D)	Other
K22	TACACGTTAATATA
K29	TATATGGTAGAAC
K31	ATATGACACCACTT
L17	AAACGGGCTGC
L34	CTTAAAAGGATATT
L37	GATAAAAATTGACG
P19	TTTGCTTAAAACC
P23	GGCATATACCTTAG
P38	CAAAATATCATATT
P40	ACACTCGGGGTATCT
P46	TACAGGGCGTAAAG
P54	GATGCCGTCAATAT

[a] Sequences are shown in 5′→3′ orientation. Alignment sequences are shown in bold above; consensus shown below. Underlining indicates sequences present in the defined flanks of ST2.

[b] Sequences present in both TATAAA and TAAATA alignments.

Table 2. TBP consensus sequence

	Position							
	1	2	3	4	5	6	7	8
A	1	25	0	32	32	27	3	14
C	4	0	0	0	0	0	3	3
G	4	1	0	0	0	5	5	5
T	18	6	32	0	0	0	20	7
Totals	27	32	32	32	32	32	31	29
Consensus	**T**	**A**	**T**	**A**	**A**	**A**	**T**	**A**

Consensus sequence derived from the TATAAA and TAAATA alignments in Table 1. The consensus sequence was determined by comparison of selected base distribution 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.

Table 3. TBP binding affinity and promoter strength for different TATA sequences

Sequence	TBP affinity (%)	Transcription (%)
TATAAATA	100 ± 2.4	100
TATAAA	90 ± 0.5	72 ± 1.5
TAAATA	48 ± 14	20 ± 0.5
TATATA	52 ± 6.4	54 ± 1.5
TATAAAAG (MLP)	n.d.	199 ± 0.7

TBP affinity determined by a restriction endonuclease protection assay using a TATAAATA-containing competitor. Transcription efficiency determined by an *in vitro* transcription assay reconstituted with hTBP. Values were derived from either three (TBP affinity) or two (transcription) independent experiments.
n.d., not determined.

Functional characteristics of different TATA sequences

The TATA consensus sequences selected by REPSA were investigated for their hTBP binding affinity and for their ability to serve as a minimal promoter in an *in vitro* transcription assay. Relative binding affinities were determined by a IISRE cleavage protection assay (6,21). Radiolabled 187 bp DNA fragments were generated containing the TATA sequences TATAAATA, TATAAA, TAAATA and TATATA, these representing the extended consensus and the three groups of TATA sequences as previously defined. These radiolabeled probes were incubated with 80 nM hTBP and increasing concentrations of an identical competitor DNA fragment containing the sequence TATAAATA. After a 30 min incubation to allow hTBP binding, the IISRE *Bpm*I was added to cleave the probe unprotected by hTBP. An example of this analysis for the TATAAATA probe is shown in Figure 3B. Note that under these conditions the maximal amount of *Bpm*I cleavage observed (lane C, no hTBP, no competitor DNA) was only 61%. The minimal amount of cleavage, when hTBP and no competitor DNA were present, was 40%. These reactions provided endpoints for the cleavage protection assay. We arbitrarily chose a protection of 50% as our measure of relative binding affinity. For the TATAAATA sequence, 50% protection occurred when 10 nM competitor was present. Experiments were performed with the other TATA sequences using the TATAAATA fragment as competitor. Their values as a percentage of the affinity found for the TATAAATA sequence are presented in Table 3. We found that the extended consensus TATA and the shorter TATAAA sequence demonstrated similarly strong binding to hTBP, while the dissociation constants of TATATA and TAAATA were 52 and 48% of the value for TATAAATA respectively.

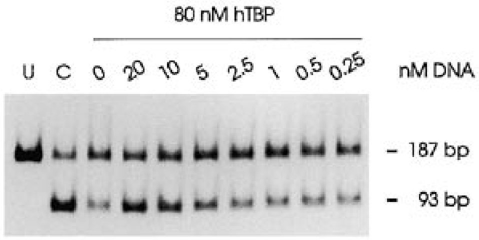

Figure 3. Determination of hTBP binding affinity to the sequence TATAAATA using a restriction endonuclease protection assay. Shown is an autoradiogram of the reaction products after partial cleavage with *Bpm*I and resolution by non-denaturing PAGE. U, uncut probe control; C, cleaved probe control.

The ability of these TATA sequences to serve as a core promoter was determined by an *in vitro* transcription assay. The sequences TATAAATA, TATAAA, TAAATA and TATATA were cloned upstream of G-less reporter cassettes such that the initial T was located 31 bp upstream of an initiating A. These templates are schematically represented in Figure 4A. *In vitro* transcription assay mixtures were reconstituted with purified general transcription factors, including hTBP and equimolar concentrations of two templates, one containing the TATAAATA sequence, which served as an internal control. The results of a typical transcription experiment are shown in Figure 4B. Relative transcription efficiencies were determined by comparing the relative intensity of the RNA products from the test templates with that of the TATAAATA control. These data are presented in Table 3. In each case the relative transcription efficiencies of the TATA sequences were similar to their relative hTBP binding affinities, suggesting a correspondence between hTBP binding and *in vitro* transcription. Note that the MLP promoter exhibited twice the transcription rate as found for the control TATAAATA-containing template. Previously we had shown that the purified protein fractions used in these *in vitro* transcription assays do not support initiator-dependent transcription activation (23). Thus our data suggest that other characteristics intrinsic to the core MLP (e.g. G-rich sequences flanking the TATA element) may play a role in determining the overall transcription efficiency of this promoter.

DISCUSSION

As an example of its use in combinatorially selecting consensus protein binding sites on duplex DNA, REPSA successfully identified preferred hTBP binding sequences. Of the sequences found, 45 could be classified into the three TATA box sequences TATAAA, TAAATA and TATATA as described by Struhl *et al.* Examination of the consensus sequence obtained for each of these groups revealed a larger consensus TATA sequence, TATAAATA. That this sequence was a consensus was supported by a statistically significant number of appearances in the emergent population (four appearances, $P < 0.01$). The only other significant 8 bp sequence identified corresponded to the adenovirus-2 major late TATA box sequence, TATAAAAG (three appearances, $P < 0.01$), well recognized as an efficient TATA box. Both TATAAATA and TATAAAAG have been identified in an optimized weight matrix derived from a comparative sequence analysis of 502 unrelated class II gene promoters (16). However, it is interesting to note that the equally expected, related 8 bp

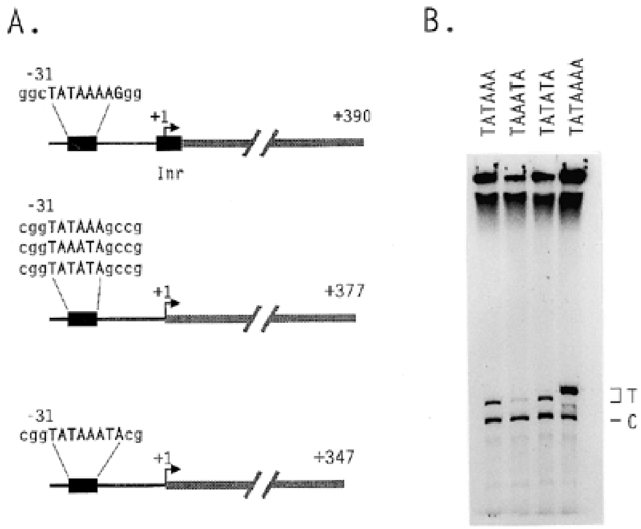

Figure 4. Determination of core promoter strength of different TATA sequences using an *in vitro* transcription assay. (**A**) Schematic representations of the transcription templates. (Top) The adenovirus-2 MLP core promoter (from –40 to +11) containing the sequence TATAAAAG. (Middle) Test templates containing the 6 bp sequences TATAAA, TAAATA and TATATA. (Bottom) Control template containing the 8 bp sequence TATAAATA. Core promoter elements are indicated by black boxes. Inr, initiator element. The G-less cassette is indicated by a broken gray bar with its length indicated at right. (**B**) *In vitro* transcription of templates containing different TATA sequences. TATA sequences present in the test templates are indicated above each lane. T, transcripts from the test templates; C, transcripts from the TATAAATA control template.

sequences TATAAAAA and TATAAATG were not selected for by REPSA. A combinatorial search of duplex sequences recognized by *Acanthamoeba* TBP also identified TATAAATA and TATAAAAG in only a small subset of the selected sequences, with TATATAAG and TATATATA predominating instead (26). Differences between the combinatorially derived consensus sequences might reflect intrinsic specificity differences between *Acanthamoeba* and human TBPs. Alternatively, it could reflect differences resulting from the selection methods employed, i.e. differences in TBP–DNA complex stability during non-denaturing PAGE or during IISRE cleavage respectively.

Is the REPSA-selected sequence TATAAATA truly a better TATA box? *In vitro* assays to determine hTBP binding affinities and transcriptional efficiencies found that the extended consensus sequence exhibited both higher affinity and promoted more transcription than the other consensus sequences found. In general, binding affinity and promoter strength correlated directly among the TATA sequences investigated. Such a correlation between binding affinity and transcription could be an artifact of the *in vitro* transcription assay, which can be made highly dependent on TFIID activity (22). Nonetheless, our data suggest that under some circumstances TATA box binding by TBP can be the rate limiting step in transcription initiation by RNA polymerase II.

Our ability to select functional TBP binding sequences demonstrates the value of REPSA for determining protein binding sites. The enzymatic selection employed gives REPSA greater flexibility than many other combinatorial methods. REPSA selection conditions are compatible with many physiological conditions, thus facilitating selection of biologically relevant protein binding sites. Likewise, the mild selection conditions allow identification of consensus sequences of proteins with unusual, as is the case with TBP, or weak binding characteristics. Though a well-characterized protein was used in this proof-of-concept study, this method should also be suitable for identifying protein binding sites for poorly characterized proteins, e.g. those that have not been purified to homogeneity or for which no antibodies are available. Ultimately, it should be possible to use REPSA with a crude mixture of proteins, for example a yeast whole cell extract, to identify a set of preferred binding sequences for the DNA binding proteins present therein. Identifying these sequences within the control regions of various genes could provide insights into transcriptional regulatory pathways present in an organism. Given current advances in sequencing whole genomes, combinatorial methods such as REPSA may well become important in the next generation of studies, protein annotation and regulatory program identification, thus bridging the gulf between raw sequence data and actual biological processes.

ACKNOWLEDGEMENTS

We thank Michèle Sawadogo for critical reading of this manuscript. This research was supported by grants from the Welch Foundation (G-1199) and the American Cancer Society (RPG-97-028-01-DHP) and by a Physicians' Referral Service research award. P.H. was the recipient of a National Research Service Award pre-doctoral traineeship from the National Cancer Institute (T32 CA60440).

REFERENCES

1 Szostack,J.W. (1992) *Trends Biochem. Sci.*, **17**, 89–93.
2 Wright,W.E. and Funk,W.D. (1993) *Trends Biochem. Sci.*, **18**, 77–80.
3 Kenan,D.J., Tsai,D.E. and Keene,J.D. (1994) *Trends Biochem. Sci.*, **19**, 57–64.
4 Ellington,A.D. and Szostack,J.W. (1990) *Nature*, **346**, 818–822.
5 Pei,D., Ulrich,H.D. and Schultz,P.G. (1991) *Science*, **245**,1408–1411.
6 Hardenbol,P. and Van Dyke,M.W. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 2811–2816.
7 Mavrothalassitis,G., Beal,G. and Papas,T.S. (1990) *DNA Cell Biol.*, **9**, 783–788.
8 Blackwell,T.K. and Weintraub,H. (1990) *Science*, **250**, 1104–1110.
9 Thieson,H.-J. and Bath,C. (1990) *Nucleic Acids Res.*, **18**, 3203–3209.
10 Polloc,R. and Treisman,R. (1990) *Nucleic Acids Res.*, **18**, 6197–6204.
11 Tsai,D.E., Harper,D.S. and Keene,J.D. (1991) *Nucleic Acids Res.*, **19**, 4931–4936.
12 Czernik,P.J., Shin,D.S. and Hurlburt,B.K. (1994) *J. Biol. Chem.*, **269**, 27869–27875.
13 Burley,S.K. and Roeder,R.G. (1996) *Annu. Rev. Biochem.*, **65**, 769–799.
14 Nikolov,D.B., Chen,H., Halay,E.D., Hoffman,A., Roeder,R.G. and Burley,S.K. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 4862–4867.
15 Juo,Z.S., Chiu,T.K., Leiberman,P.M., Baikalov,I., Berk,A.J. and Dickerson,R.E. (1996) *J. Mol. Biol.*, **261**, 239–254.
16 Bucher,P. (1990) *J. Mol. Biol.*, **212**, 563–578.
17 Hahn,S., Buratowski,S., Sharp,P.A. and Guarente,L. (1989) *Proc. Natl. Acad. Sci. USA*, **86** 5718–5722.
18 Singer,V.L., Wobbe,R. and Struhl,K. (1990) *Genes Dev.*, **4**, 636–645.
19 Wiley,S.R., Kraus,R.J. and Mertz,J.E. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 5814–5818.
20 Innis,J.W., Moore,D.J., Kash,S.F., Ramamurthy,V., Sawadogo,M. and Kellems,R.E. (1991) *J. Biol. Chem.*, **266**, 21765–21772.
21 Ward,B. (1996) *Nucleic Acids Res.*, **24**, 2435–2440.
22 Sawadogo,M. and Roeder,R.G. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4394–4398.
23 Wang,J.C. and Van Dyke,M.W. (1993) *Biochim. Biophys. Acta*, **1216**, 73–80.
24 Chen,W. and Struhl,K. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2691–2695.
25 Wobbe,C.R. and Struhl,K. (1990) *Mol. Cell. Biol.*, **10**, 3859–3867.
26 Wong,J.M. and Bateman,E. (1994) *Nucleic Acids Res.*, **22**, 1890–1896.