Synthetic promoter elements obtained by nucleotide sequence variation and selection for activity

Gerald M. Edelman*†‡, Robyn Meech*, Geoffrey C. Owens†, and Frederick S. Jones*†

*Department of Neurobiology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037; and †The Neurosciences Institute, 10640 John Jay Hopkins Drive, San Diego, CA 92121

Contributed by Gerald M. Edelman, December 23, 1999

Eukaryotic transcriptional regulation in different cells involves large numbers and arrangements of cis and trans elements. To survey the number of cis regulatory elements that are active in different contexts, we have devised a high-throughput selection procedure permitting synthesis of active cis motifs that enhance the activity of a minimal promoter. This synthetic promoter construction method (SPCM) was used to identify >100 DNA sequences that showed increased promoter activity in the neuroblastoma cell line Neuro2A. After determining DNA sequences of selected synthetic promoters, database searches for known elements revealed a predominance of eight motifs: AP2, CEBP, GRE, Ebox, ETS, CREB, AP1, and SP1/MAZ. The most active of the selected synthetic promoters contain composites of a number of these motifs. Assays of DNA binding and promoter activity of three exemplary motifs (ETS, CREB, and SP1/MAZ) were used to prove the effectiveness of SPCM in uncovering active sequences. Up to 10% of 133 selected active sequences had no match in currently available databases, raising the possibility that new motifs and transcriptional regulatory proteins to which they bind may be revealed by SPCM. The method may find uses in constructing databases of active cis motifs, in diagnostics, and in gene therapy.

A central problem in molecular cell biology is to understand how combinations of promoter elements and the proteins to which they bind regulate gene transcription in particular cellular contexts. Although many of the principles of gene regulation originally outlined in studies of prokaryotes (1) also apply to eukaryotes, eukaryotic transcriptional regulation is considerably more complex (2, 3). This complexity arises from several characteristic features: the large number of different DNA regulatory motifs and regulatory proteins to which they bind; the number of different protein components that make up the basic transcription machinery; the contribution of enhancers and silencers that may be located at considerable distances from the core promoter; and the need for chromatin remodeling at specific times and places. One of the first steps in unraveling this complexity is to develop a procedure for surveying the number of cis-regulatory elements that are active in various arrangements and in different cellular contexts.

We describe here a new synthetic promoter construction method (SPCM) based on sequence variation and selection of 18 mers of DNA to reveal cis elements that function to modulate a minimal promoter comprised of a TATA box and an initiator sequence. The method (Fig. 1) involves generation of a retroviral library of synthetic promoters containing random 18 mer sequences (Ran18), packaging of the proviral library, infection of eukaryotic cells, selection first for antibiotic resistance and then for green fluorescent protein (GFP) expression using fluorescence-activated cell sorting (FACS), and recovery of selected Ran18 sequences for analysis of activity and DNA sequencing.

Using SPCM, we identified >100 DNA sequences from the Ran18-promoter library that gave from 4- to 50-fold activation of the minimal promoter in the neuroblastoma cell line Neuro2A. Comparison of these sequences with the TransFac database version 3.5 by using a software package identified eight predominant DNA motifs: AP2, CEBP, GRE, Ebox, AP1, ETS, CRE, and SP1/MAZ. One-half of the active Ran18 elements contained one or more of these motifs. Composites consisting of pairs, triples, or quadruples of these motifs were among the most active in promoter assays. Between 5 and 10% of the active DNA sequences were novel, i.e., were not represented in known transcription factor databases. The SPCM provides a means for discovery of new promoter elements, for analysis of combinations of known and novel elements, and for uncovering new transcriptional regulatory proteins. It also has potential applications in diagnosis and gene therapy, contexts in which cellular responses to synthetic promoters may be usefully controlled.

Materials and Methods

The SPCM was designed to optimize the identification of active synthetic promoters. A library of random 18-bp DNAs (designated Ran18) was inserted 30 bp upstream of a minimal promoter containing TATA box and initiator elements. A retroviral delivery system was used to integrate these promoter constructs into the genome of target Neuro2A cells. A bicistronic enhanced green fluorescent protein (EGFP)/puromycin *N*-acetyltransferase gene cassette was constructed for a double selection procedure (Fig. 1). Synthetic promoters integrated into the cellular genome were identified by their high level of EGFP expression. A second round of selection was performed to minimize false positives that might arise from integration near endogenous enhancers. PCR was used to amplify functional Ran18 sequences from the genomic DNA of selected cells. Active Ran18 elements were inserted into a luciferase reporter plasmid, and after transient transfection, their activities were examined in Neuro2A cells. The sequences of active individual Ran18 elements were then determined and regulatory motifs were identified by the RIGHT software package. This package allows simultaneous comparison of a database of active Ran18 elements to existing databases such as TransFac.

Construction of a Ran18-Promoter Library. Ran18 oligonucleotides were constructed by using a PE Biosystems (Foster City, CA) DNA synthesizer. Ran18 elements were flanked by two different sequences (left, ctactcacgcgtgatcca; right, cgagcgaacgcgtgcaatg) containing the *Mlu* restriction site that allowed cloning into the selection vector. Double-stranded Ran18s were generated by primer extension, digested with *Mlu*I, and purified by extraction from an 8% polyacrylamide gel. The library of Ran18 sequences was ligated into a retroviral vector and transformed into XL1-Blue *Escherichia coli* (Stratagene). Plasmid DNA was prepared using Maxi-Prep columns (Qiagen, Valencia, CA).

Abbreviations: SPCM, synthetic promoter construction method; GFP, green fluorescent protein; EGFP, enhanced GFP; LTR, long terminal repeat; SLA, selected luciferase activators; FACS, fluorescence-activated cell sorting; pCMV, cytomegalovirus promoter.

‡To whom reprint requests should be addressed at: Department of Neurobiology, SBR14, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037.

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.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.040569897. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.040569897

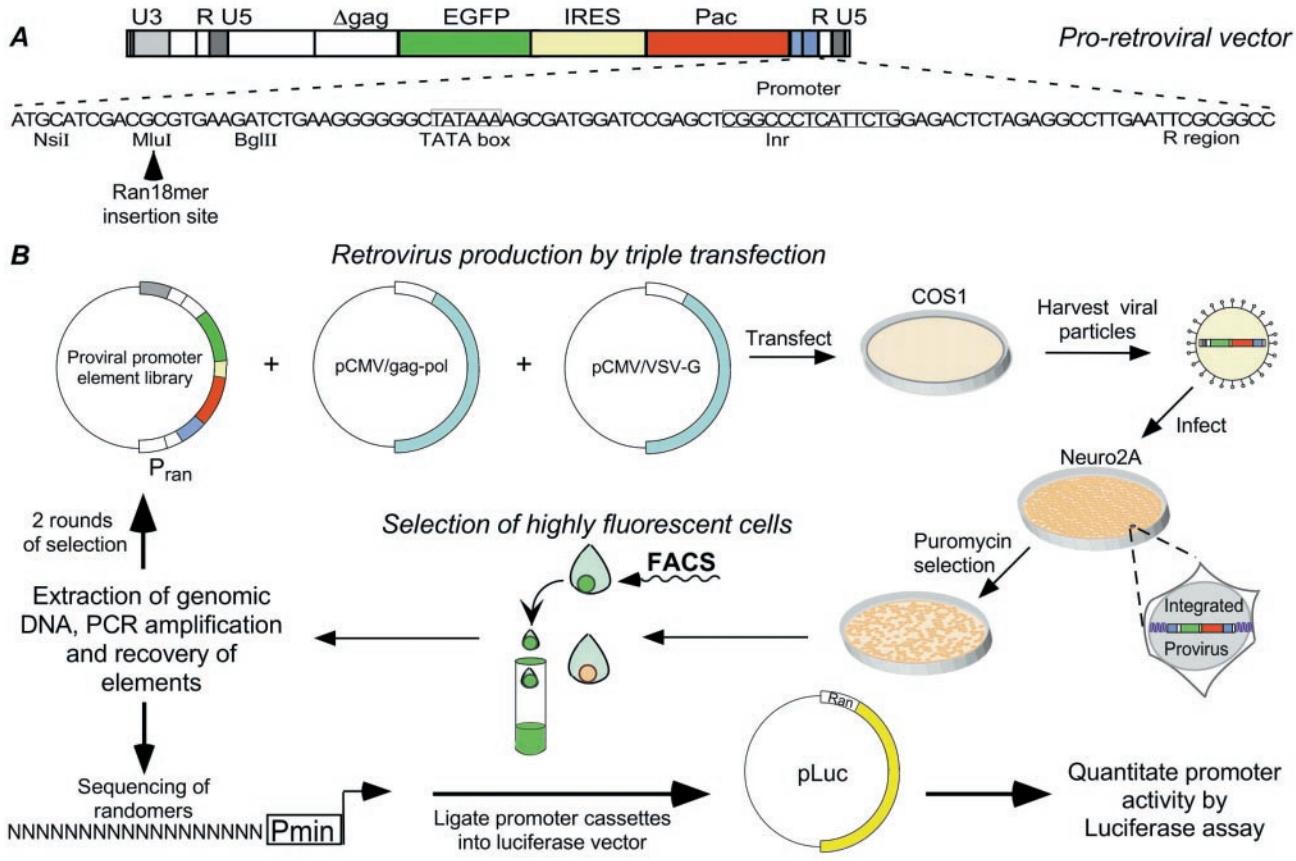

Fig. 1. Strategy for selection of synthetic promoter elements. A library of random 18 mers is constructed in a selection vector (*A*) which is packaged into retroviral particles (*B*) that are harvested and used to infect target cells which are then treated for 3 days with puromycin to kill uninfected or poorly expressing cells. Surviving cells are subjected to FACS, and the most fluorescent cell fractions are collected. Genomic DNA is prepared from these cells and elements are recovered by PCR. Elements are then religated into the retroviral vector for a second round of selection. Finally, the elements are ligated into the pLuc luciferase reporter; the activities of the elements are quantitiated by luciferase assays, and their DNA sequences are determined.

Design of the Retroviral Vector. A retroviral vector called MESV/EGFP/IRES/Pac/pro(ori) was constructed (Fig. 1*A*), based on a variant of the murine embryonic stem cell virus (4). The vector was modified to include a polylinker for insertion of Ran18 sequences, the TATA box from the adenovirus major late promoter (5), and the initiator from the mouse terminal deoxynucleotidyltransferase gene (6). The TATA box cassette replaced the U3 region of the downstream retroviral long terminal repeat (LTR) (Fig. 1*A*). The U3 region of the upstream LTR was replaced by the U3 region of the Rous sarcoma virus to increase viral titer. A cassette containing the gene encoding the enhanced green fluorescent protein (EGFP) (7), an internal ribosome entry site (IRES), and the puromycin *N*-acetyltransferase (pac) gene was constructed and inserted downstream of the LTR. Finally, a simian virus 40 origin of replication was inserted into the proviral plasmid to allow replication in COS1 cells.

Retroviral Packaging. Packaging was achieved by cotransfection of the proviral DNA library into COS1 cells together with two helper plasmids pCMV-GP(sal) and pMD.G. The pCMV-GP(sal) plasmid contains the *gag* and *pol* genes from the Moloney murine leukemia virus under the control of the cytomegalovirus promoter. This plasmid was provided by the University of California at San Diego gene therapy program. The pMD.G plasmid encoded the vesicular stomatitis virus G glycoprotein (8), an envelope protein required for assembly of retroviral particles. Three 100-mm dishes of COS1 cells (8×10^5 cells per dish) were transfected with 4 μg of proviral library DNA, 4 μg of the pCMV/gag-pol plasmid, and 2 μg of the pCMV/VSV-G plasmid using Fugene transfection reagent (Roche Diagnostics). Media were changed 24 hr later, and supernatant containing retroviral particles was collected after an additional 24 hr, filtered, and combined with polybrene to a final concentration of 5 μg/ml. This mixture was used to infect Neuro2A cells in monolayer culture. The ratio of viral particles to cells was optimized to ensure a high probability of single infection/integration events; this ratio generally resulted in infection of 25–40% of the Neuro2A cells.

Selection of Active Promoter Elements. After retroviral infection, each cell incorporated on average a single integrated DNA provirus containing a different Ran18 element upstream of the minimal promoter and the selectable markers, EGFP and Pac. Identification of active Ran18-promoter elements involved two selection steps (Fig. 1*B*). First, 1 mg/ml puromycin was added to the Neuro2A cells 24 hr postinfection for 3 days to kill uninfected and poorly expressing cells. Surviving cells were harvested and subjected to FACS using the FACStar sorter (Becton Dickinson). Control cells were infected with a reporter retrovirus containing either a minimal promoter or a strong promoter (Rous sarcoma virus) to drive expression of the EGFP reporter gene. The EGFP fluorescence in cells driven by the minimal promoter provided a baseline fluorescence threshold above which cells having active Ran18 promoters were selected. The Rous sarcoma virus control provided a measure of infection

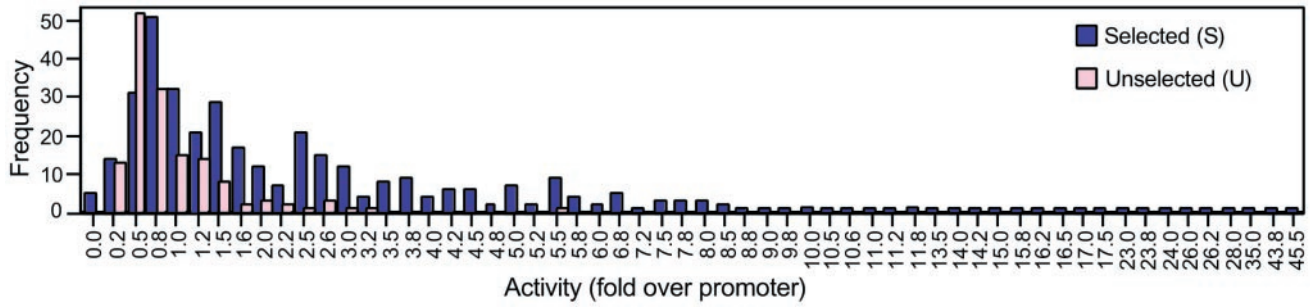

Fig. 2. Distribution by activity of 160 unselected (pink) and 480 selected (blue) Ran18 sequences. Luciferase activities of 120 selected sequences exceeded an activity threshold of 4 times that of the minimal promoter alone.

efficiency. Cells with above-threshold fluorescence were amplified in number by culturing for an additional 3–5 days. Genomic DNA was then extracted from the selected cells using the QiaAmp tissue kit (Qiagen).

Ran18 sequences were recovered by genomic PCR. The pool of amplified Ran18 sequences was then digested with restriction enzymes NsiI and *Bgl*II and recloned into the proviral selection vector to allow a second round of selection. This allowed re-examination of active promoters at different integration sites. After the second round, Ran18 sequences were again amplified by PCR, digested either with *Nsi*I and *Eco*RI (releasing both the 18 mer and minimal promoter) or with *Mlu*I (releasing only the 18 mer). These fragments were cloned into promoterless (pLuc) or promoter-containing luciferase- (pLucPro) reporter vectors, respectively. The pLuc vector was made by inserting *Nsi*I, *Stu*I, and *Eco*RI restriction sites into pGL3basic (Promega). The pLucPro vector is a variant of pLuc containing the minimal promoter from our retroviral vector.

Analysis of Promoter Activity of Selected Ran18 Elements. To quantify the activity of Ran18 elements, the Ran18/pLucPro plasmids were transfected into Neuro2A cells in 24-well tissue culture plates. One hundred nanograms of each reporter was transfected together with CMVβgal to normalize for transfection efficiency, and 48 hr later the cells were harvested and assayed for β-galactosidase and luciferase activity as described (9). The activity of pLucPro was used as a reference standard for measuring the levels of luciferase activity generated by selected Ran18/promoters.

Comparison of Ran18 Sequences with Motifs in the TransFac Database. Ran18 elements were sequenced by using an automated DNA sequencer (model 373, PE Biosystems). Sequences were then searched for candidate transcription factor binding motifs present in the TransFac database (version 3.5) using the RIGHT (Reeke's Interactive Gene Hacking Tool) software package. RIGHT is a motif recognition program based on a regular expression search and is particularly useful for SPCM because it allows a batch format for sequence input and has the capacity to simultaneously analyze large numbers of Ran18-promoter sequences.

Gel Mobility-Shift and Cell Transfection Analyses. Nuclear extracts from Neuro2A cells were examined for binding to ^{32}P-labeled double-stranded DNA probes containing ETS, CREB, and SP1/MAZ motifs derived from highly active Ran18 sequences by using gel mobility-shift analyses as described (9). The contribution of these motifs to the binding and activity of synthetic promoters was examined by mutation of their sequences and assaying for gel shifts and reduction in luciferase activity.

Results

The SPCM (Fig. 1) was used to identify active synthetic promoters in the neuroblastoma cell line Neuro2A from a library of greater than 5×10^7 individual Ran18 sequences. Synthetic promoters driving the highest levels of GFP expression were selected using FACS by collecting the top 1% of fluorescent cells. The first round of FACS-promoter selection yielded 12,000 cells. Promoter elements were recovered by PCR amplification and used to construct a library that was subjected to a second round of FACS selection. Elements recovered by PCR from the top 1% of GFP-expressing cells (75,000 cells) were analyzed by DNA sequencing, and their activities were assessed by luciferase assays.

Activity of Selected Ran18 Sequences. The SPCM generated a population of Ran18 sequences that was enriched for active promoter elements, relative to the original library. To examine the extent of this enrichment, we compared luciferase activities of 480 selected Ran18-promoter cassettes (the set designated S) to a randomly picked sample of 160 promoter cassettes from the original unsorted library (the set designated U in Fig. 2). Assuming that the unselected Ran18 elements had a Gaussian distribution, the mean activity was 0.8 and the SD was 0.4. Using this distribution for the activities of the unselected (U) Ran18 elements and allowing for a confidence interval of $>95\%$, we concluded that 4-fold activity above that of the minimal promoter represented a conservative activity threshold.

Analysis of the distribution of activities of the 480 selected elements (set S, superimposed upon the normal distribution from set U in Fig. 2) revealed that 120 of the selected Ran18 sequences ($\approx$25%) had activity that was 4- to 50-fold greater than that of the minimal promoter. Only one sequence from the U set ($<1\%$ of the total) showed greater than fourfold activity. The SPCM thus provided $\approx$25-fold enrichment of active promoter elements. These selected Ran18 sequences highly active in luciferase assays constituted the SLA set (selected luciferase activators).

DNA Motifs in Ran18 Sequences. The DNA sequences of 106 SLA, 133 S, and 132 U Ran18 elements were determined and compared with known motifs within the TransFac database. Only motifs that had 100% sequence identity with TransFac motifs with a length of 6 bp or greater were scored as matches. Known regulatory motifs were identified in each of the three sets, but the prevalence and linear arrangement of particular motifs differed among the sets.

Eighteen of the most active Ran18 sequences from the SLA set showed 78 matches with known motifs (Fig. 3). A significant number of these matches occurred as composites consisting of two or more motifs that either were overlapping or contiguous. The two most active elements, MS44 and S173 registered six and

Fig. 3. Sequences and identified motifs of 18 synthetic promoters are arranged in decreasing order of luciferase activity. Different colors represent different motifs; composites show multiple colors. Underlining is used to indicate matching sequences found in database searches.

five matches, respectively, with known motifs and contained a composite made up of ETS, AP1, CREB, and GATA motifs. These analyses suggested that composite motif arrangement might contribute significantly to the high level of activity produced by these synthetic promoters.

An analysis of the complete SLA, S, and U sets was performed to compare the number of matches, the distribution of motifs, and the number and type of composite elements. Overall, the SLA and S sets contained approximately twice as many motifs as the U set (Fig. 4*A*). A significant proportion of the motifs identified in all three sets (46% for U, 46.5% for S, and 51% for SLA) was made up of only eight motifs. These represented putative-binding sites for eight different families of transcriptional regulators: AP2, CEBP, GRE, Ebox, ETS, CREB, AP1, and SP1/MAZ. SLA and the S set contained approximately twice as many of these motifs as the U set (Fig. 4*A*). A comparison of the occurrences of each of the eight most frequent motifs among the three sets (Fig. 4*C*) revealed a significant increase in the number of Ebox, ETS, CREB, AP1, and SP1/MAZ motifs in SLA and S sets as compared with the U set. There was no significant increase in the number of AP2, CEBP, and GRE motifs.

As shown in Fig. 4*B*, the total number of composites increased ≈2.8-fold in both the SLA and S sets over the number found in the U set. Composites were further categorized into three types: category A (those containing two or more of the eight most common motifs), category B (those containing one of the eight common motifs and a motif other than one of the eight), and category C (those containing two or more motifs other than the eight most frequent motifs). A comparison of these three categories over the three sets of synthetic promoters (Fig. 4*B*) revealed a dramatic increase in the number of category A composites in the SLA and S sets (3- and 5.7-fold, respectively) over that observed in the U set as well as in category B composites (2.7-fold for SLA and S sets). Category C composites also increased in the S set as compared with the U set (up 2.4-fold) but only increased 1.4-fold in the SLA set. These analyses indicate that composites containing one or more of the eight frequent motifs correlate favorably with highly active synthetic promoters.

Finally, we examined the number of composites containing each of the eight frequent motifs. As shown in Fig. 4*D*, in synthetic promoters of the SLA and S sets as compared with the U set, the number of composites containing GRE, Ebox, AP1, CREB, and SP1/MAZ motifs increased dramatically and those containing ETS increased moderately. However, no increases were observed in the number of composites containing AP2 and CEBP elements (Fig. 4*D*). Taken together with the data presented in Fig. 4*C* showing that only the Ebox, CREB, AP1, and SP1/MAZ increased in numbers in the SLA and S sets, these analyses support the following conclusions: (*i*) increases in both number and presence in composites of the Ebox, AP1, ETS, CREB, and SP1/MAZ were correlated with active synthetic promoters; (*ii*) an increase in the occurrence of GRE elements in composites but not in their abundance was correlated with active synthetic promoters; and (*iii*) there was no correlation between either the number or the presence in composites of AP2 and CEBP elements with activity of synthetic promoters.

A small proportion of active Ran18 sequences from the SLA and S sets (4% and 11%, respectively) showed no matches to known transcriptional regulatory motifs. These sequences are likely to contain novel regulatory elements and are being analyzed in more detail to determine whether they bind to novel proteins.

DNA Binding and Activity of ETS, CREB, and SP1/MAZ Motifs in Synthetic Promoters. To determine whether examples of the eight most frequent motifs identified within the Ran18 sequences actually contributed to DNA binding and promoter activity, gel mobility-shift and promoter assays were performed on native and mutated versions of the ETS, CREB, and MAZ/SP1 motifs in the synthetic promoters MS44 and MS113. The right hand element found in MS44 (designated MS44B) and the Ran18 element in MS113 were examined for binding to Neuro2A

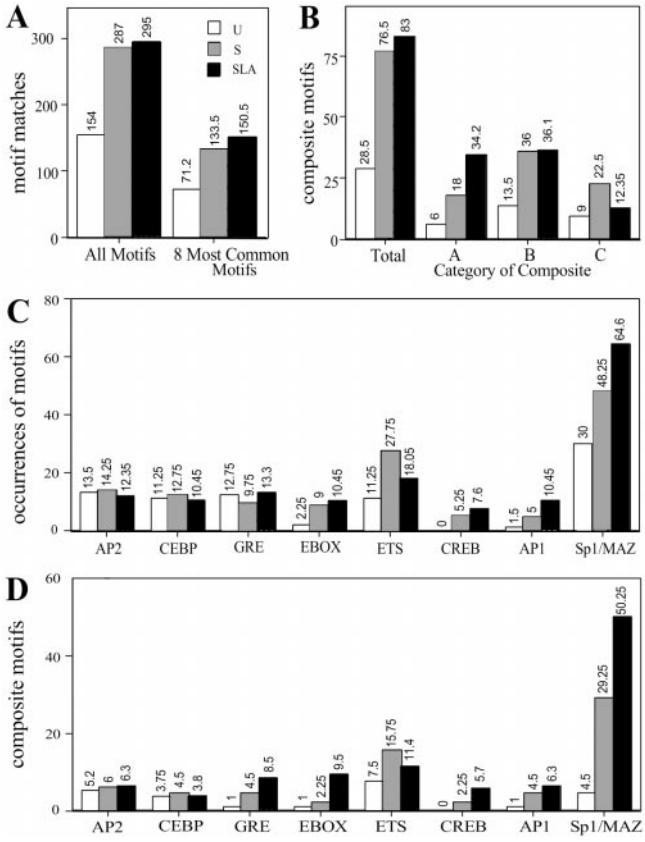

Fig. 4. Histogram of relative frequencies of motifs in unselected (U) and selected (S; SLA) sets. (*A*) Number of motif matches against the TransFac database. (*B*) Number of composite motifs in each set. (*C*) Frequency of eight most common motifs in each set. (*D*) Number of composites formed with the eight most common motifs. All numbers are per 100 Ran18 elements.

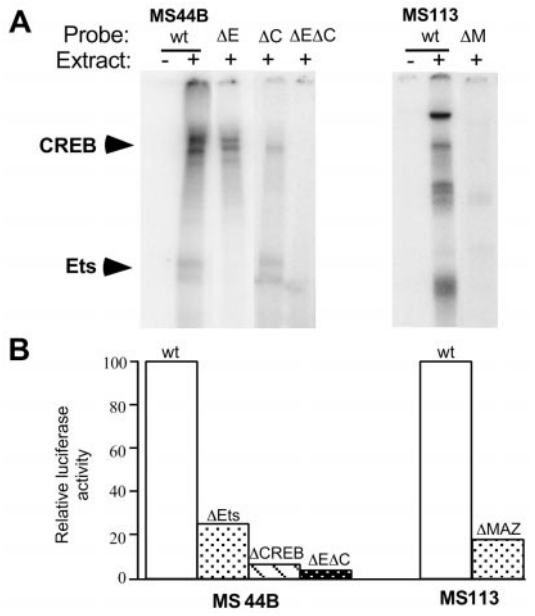

Fig. 5. DNA binding (*A*) and promoter activity (*B*) assays of ETS-, CREB-, and MAZ-containing synthetic promoters MS44B and MS113. Binding reactions were performed with wild-type (wt), ^{32}P-labeled MS44B, and MS113 probes or probes having mutations in CREB (ΔC), ETS (ΔE), ETS and CREB (ΔEΔC), and MAZ (ΔM) elements. (*B*) ETS/CREB and MAZ activities were assayed in native and mutated variants of MS44B and MS113 synthetic promoter sequences. Mutation of each motif within these promoters is indicated (ΔETS, ΔCREB, ΔEΔC, and ΔMAZ).

nuclear extracts (Fig. 5*A*). MS44B contains an ETS/CREB composite and MS113 contains a MAZ/SP1 motif.

Gel mobility-shift experiments using the MS44B probe revealed high and low molecular weight DNA/protein complexes. Formation of high and low molecular weight complexes was eliminated in ^{32}P-labeled variants of the MS44B sequence called ΔC and ΔE, having multiple base pair substitutions in the CREB and ETS motifs, respectively. A probe having both ETS and CREB mutations (ΔEΔC) showed no binding to proteins in nuclear extracts of Neuro2A cells. Experiments that included these and mutated versions of these motifs as cold competitors in binding reactions provided similar results (data not shown). These data indicate that proteins in upper and lower molecular weight complexes most likely represent members of the CREB and ETS families of proteins, respectively. ETS and CREB mutations in MS44B also resulted in substantial reductions of MS44B-promoter activity. Luciferase-reporter variants of MS44B with mutations in the ETS, the CREB, or in both ETS and CREB motifs had only 27%, 5%, and 3%, respectively, of the promoter activity of MS44B.

Similar binding and activity assays were performed to investigate the efficacy of the SP1/MAZ motif in the MS113 promoter. As shown in Fig. 5 *A* and *B*, mutation of the SP1/MAZ motif resulted in a complete elimination of DNA binding of Neuro2A nuclear proteins to the MS113 element. A variant of the MS113 synthetic promoter containing these SP1/MAZ mutations showed only 18% of the promoter activity of MS113 (Fig. 5*B*). Collectively, these experiments indicate that the ETS/CREB composite and SP1/MAZ motifs identified in searches of the TransFac database with RIGHT software are the major contributors to both the binding and activity of the synthetic promoters in which they were found.

Discussion

The SPCM was designed to address several problems confronted in analyzing the complex machinery of eukaryotic gene transcription. A basic problem is to survey the types and frequencies of DNA motifs that contribute to promoter activity. It is therefore important to understand which combinations of cis and trans elements work in concert with a core promoter and the basic transcription machinery in a given cellular context (2, 3). Although the scope of the present study was limited to identifying functional motifs active in the context of one particular cell line, it provides a view into the types, potential combinations, activity, prevalence, and novelty of a relatively large sample of different cis-regulatory motifs.

After GFP selection of 480 sequences, 120 had greater than 4-fold activity over that of the minimal promoter in luciferase assays. We used the RIGHT software package to analyze the occurrence of various motifs in three different sets of synthetic promoters: unselected (the U set), those selected by GFP fluorescence to have promoter activity as integrants in the genome (the S set), and GFP-selected synthetic promoters that, as measured after cellular transfection, gave high levels of activity in an episomal state with the luciferase assay (the SLA set). Approximately twice as many matches with known transcriptional regulatory motifs were found in the SLA and S sets than were found in the U set. Fifty-one percent of the matches were with eight different motifs: AP2, CEBP, GRE, Ebox, ETS, CREB, AP1, and SP1/MAZ. As shown in Fig. 3, the most active sequences were made up of composites of these eight motifs. For example, the most active two sequences had ETS and CRE motifs in an overlapping composite. A BLAST search for occurrence of this composite in natural promoters revealed an exact match with an element in the proximal promoter of a gene encoding a nonstructural protein from the parvovirus B19 (GenBank accession no. AF19028, parvovirus P6 nonstructural protein).

Such composites may also occur in other promoters and this finding prompts further searches to determine whether other highly active composites revealed by SPCM appear in eukaryotic genes.

Although our present survey was not exhaustive, we identified ≈4% to 10% of active DNA sequences that appear to be novel, i.e., do not contain motifs present in current databases. A continued search for native elements of this type is warranted as is a rigorous identification of the actual motifs responsible for their promoter activity. Such sequences may provide the means for discovery of novel regulatory proteins in cellular material. As shown in Fig. 5, searches for proteins binding to motifs found by SPCM may be particularly revealing. The application of procedures that isolate DNA-binding proteins such as Southwestern (10, 11), FROGS (12), and 1-hybrid (13–15) screening procedures along with SPCM should aid in identification of known and novel trans-factors. Clearly, this approach requires the exploration of a variety of differentiated cells to maximize the yield of discovery.

Of the eight prevalent known motifs, several (particularly SP1) have been shown to function within the core promoter (16–18). Others such as ETS and CRE have been shown to be components of enhancers. In its current state of development, the SPCM cannot distinguish whether the activity of a discovered motif is due to direct contributions to a core promoter or to its function as an enhancer. Moreover, the method presently does not distinguish silencer activity. It is obvious, however, that by adapting the method to a negative selection procedure, silencer motifs can be identified.

The present procedure allows separate determinations of the activity of a motif when integrated in the genome or in the episomal state. Of 480 integrated motifs that were selected as active by GFP-sensitive cell sorting, only 120 exceeded the fourfold threshold as plasmids in the luciferase assay. It will be of interest to determine whether certain motifs of the S population can function only in the integrated state. The possibility that some of the activities seen in the integrated state arose because of proximity to unknown enhancers raises the issue of false positive responses. However, application of multiple rounds of selection helps to reduce the frequency of such responses.

Because SPCM in its present form requires cell division for retroviral infection and integration, another technique such as the direct transfection of an SPCM-promoter/reporter construct library might provide an alternative methodology. In unpublished experiments, we have used transfection and selection by antibiotic resistance to Zeocin, constructing permanent cell lines to achieve results similar to those reported here. With this approach, however, integration of promoter constructs was inefficient. Therefore, the retroviral approach presently appears more advantageous. Moreover, the use of retroviruses allows application of SPCM to whole animals. Success in this application may yield a useful lineage analysis of cis motifs in different cell types during particular stages of development.

A number of improvements of the present method can be envisioned. The library used in our study consisted of randomers constructed by oligonucleotide synthesis. It is not currently known whether biasing of that library may have distorted the prevalence of the different types of motifs obtained by the selection procedure. The use of libraries constructed from different lengths of randomers might help guard against such potential biasing. Moreover, obtaining larger cell samples may contribute to improved statistical analysis of the prevalence of particular motifs. Application of such improvements in various cell types and species might shed light on an important evolutionary question: what changes in the prevalence of DNA regulatory motifs have occurred after various speciation events? Obviously, the current databases such as TransFac do not incorporate an exhaustive collection of all DNA-promoter motifs or even a strict criterion for promoter activity and this clearly limits evolutionary comparisons. Consistent application of the current and related SPCM approaches should ultimately enable the creation of databases of truly functional promoters and also include cognate information on various species and developmental states.

Our initial analyses of the synthetic promoters containing eight predominant regulatory motifs and highly active composites made up of these motifs suggest several useful extensions of the SPCM procedure. Besides the selection of random DNA sequences of a particular length, the method can be used to analyze combinations of a single known motif (for example an Octamer) with random sequences. This usage would allow exploration of synergies between various cis elements and the modulation of interactions with corresponding transcription factors. Moreover, deliberately assembling combinations of known elements in various lengths, orders, polarity, and spacings may shed further light on rules governing the effectiveness of DNA motifs in a given cellular context. There is already a hint in the present data that particular combinations (for example ETS/CREB and ETS/SP1) can be particularly effective. The factors that recognize these elements are known to bind cooperatively to promoters that contain both of these sequences (17, 19).

Individual studies of synthetic promoters have already been shown to be useful in a variety of applications in both prokaryotic and eukaryotic systems (20–25). The SPCM approach suggests a wider means of generalizing the synthesis of useful sequences. Of specific interest is the application of a selected set of synthesized promoters in matrix arrays to the detection of differential responses of normal cells and cells from various diseased tissues for diagnostic purposes or drug development. In addition, it is likely that the retroviral approach combined with various synthetic promoters will find uses in gene therapy.

We thank Dr. George Reeke for the use of the RIGHT program, Nicole Son and Judy Yen for technical assistance, and Dr. J. A. Gally for useful criticism. We also thank Drs. D. Copertino and B. Niculescu for help in early explorations of the technique. This work was supported by a grant from the G. Harold and Leila Y. Mathers Charitable Trust.

1. Jacob, F. & Monod, J. (1961) *J. Mol. Biol.* **3,** 318–328.
2. Sauer, F. & Tjian, R. (1997) *Curr. Opin. Genet. Dev.* **7,** 176–181.
3. Roeder, R. G. (1998) *Cold Spring Harbor Symp. Quant. Biol.* **63,** 201–218.
4. Owens, G. C., Orr, E. A., DeMasters, B. K., Muschel, R. J., Berens, M. E. & Kruse, C. A. (1998) *Cancer Res.* **58,** 2020–2028.
5. Langrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. & Ebright, R. H. (1998) *Genes Dev.* **12,** 34–44.
6. Colgan, J. & Manley, J. L. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 1955–1959.
7. Zernicka-Goetz, M., Pines, J., Huneter, S. M., Dixon, J. P. C., Siemering, K. R., Haseloff, J. & Evans, M. J. (1997) *Development (Cambridge, U.K.)* **124,** 1133–1137.
8. Naldini, L., Blomer, U., Gallay, P., Ory, D., Mulligan, R., Gage, F. H., Verma, I. M. & Trono, D. (1996) *Science* **272,** 263–267.
9. Meech, R., Kallunki, P., Edelman, G. M. & Jones, F. S. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2420–2425.
10. Singh, H., LeBowitz, J. H., Baldwin, A. S. & Sharp, P. A. (1988) *Cell* **52,** 415–423.
11. Vinson, C. R., LaMarco, K. L., Johnson, P. F., Landschulz, W. H. & McKnight, S. L. (1988) *Genes Dev.* **2,** 801–806.
12. Mead, P., Zhou, Y., Lustig, K., Huber, T., Kirschner, M. & Zon, L. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11251–11256.
13. Li, K. J. & Herskowitz, I. (1993) *Science* **262,** 1870–1874.
14. Wang, M. W. & Reed, R. R. (1993) *Nature (London)* **364,** 121–126.
15. Dowell, S. J., Romanowski, P. & Diffley, J. F. X. (1994) *Science* **265,** 1243–1246.
16. Parks, C. L. & Skenk, T. (1996) *J. Biol. Chem.* **271,** 4417–4430.
17. Karantzoulis-Fegaras, F., Antoniou, H., Lai, S. L., Kulkarni, G., D'Abreo, C., Wong, G. K., Miller, T. L., Chan, Y., Atkins, J., Wang, Y. & Marsden, D. A. (1999) *J. Biol. Chem.* **274,** 3076–3093.
18. Segal, J. A., Barnett, J. L. & Crawford, D. L. (1999) *J. Mol. Evol.* **49,** 736–749.
19. Sawada, J., Simizu, N., Suzuki, F., Sawa, C., Goto, M., Hasegawa, M., Imai, T., Watanabe, H. & Handa, H. (1999) *J. Biol. Chem.* **274,** 35475–35482.
20. Danner, S. & Soppa, J. (1996) *Mol. Microbiol.* **19,** 1265–1276.
21. Valdivia, R. H. & Falkow, S. (1997) *Science* **277,** 2007–2011.
22. Barker, L. P., Brooks, D. M. & Small, P. L. C. (1998) *Mol. Microbiol.* **29,** 1167–1177.
23. Asoh, S., Lee-Kwon, W., Mouradian, M. & Nirenberg, M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6982–6986.
24. Li, X., Eastman, E., Schwartz, R. & Draghia-Akli, R. (1999) *Nat. Biotechnol.* **17,** 241–245.
25. Wang, S., Wu, H., Jiang, J., Delohery, T. M., Isbell, F. & Goldman, S. A. (1998) *Nat. Biotechnol.* **16,** 196–201