MOLECULAR AND CELLULAR BIOLOGY, July 1989, p. 2944–2949
0270-7306/89/072944-06$02.00/0

Vol. 9, No. 7

Defining the Sequence Specificity of DNA-Binding Proteins by Selecting Binding Sites from Random-Sequence Oligonucleotides: Analysis of Yeast GCN4 Protein

ARNOLD R OLIPHANT, CHRISTOPHER J. BRANDL, AND KEVIN STRUHL*

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115

Received 30 January 1989/Accepted 7 April 1989

We describe a new method for accurately defining the sequence recognition properties of DNA-binding proteins by selecting high-affinity binding sites from random-sequence DNA. The yeast transcriptional activator protein GCN4 was coupled to a Sepharose column, and binding sites were isolated by passing short, random-sequence oligonucleotides over the column and eluting them with increasing salt concentrations. Of 43 specifically bound oligonucleotides, 40 contained the symmetric sequence TGA(C/G)TCA, whereas the other 3 contained sequences matching six of these seven bases. The extreme preference for this 7-base-pair sequence suggests that each position directly contacts GCN4. The three nucleotide positions on each side of this core heptanucleotide also showed sequence preferences, indicating their effect on GCN4 binding. Interestingly, deviations in the core and a stronger sequence preference in the flanking region were found on one side of the central C · G base pair. Although GCN4 binds as a dimer, this asymmetry supports a model in which interactions on each side of the binding site are not equivalent. The random selection method should prove generally useful for defining the specificities of other DNA-binding proteins and for identifying putative target sequences from genomic DNA.

A frequent goal in molecular biology is to define the nucleotide sequences required for binding by a specific transcriptional regulatory protein. One approach is to collect wild-type binding sites which occur in biologically related contexts. Sequences responsible for a particular function may then be recognized by their occurrence in most of the identified sites. However, the collection and identification of these sequences can be difficult, and the results will be biased if some of the sequences encode functions other than the one of interest. Another approach is to analyze mutated versions of a given binding site to determine which positions are functionally important. However, this method requires a large number of mutant sites and is biased by the starting wild-type sequence; related sequences which may bind the protein will not be studied.

Random-sequence oligonucleotides have proven to be a useful tool for identifying and defining the sequence requirements of other genetic elements (4, 11, 14, 16, 17). The isolation of functional elements from random-sequence DNA, termed random selection, can be done quickly and with many advantages not applicable to the study of wild-type elements. First, the number of elements generated is sufficiently large to define the element precisely. Second, each sequence is very likely to contain a highly functional example of the element of interest. Third, confounding (nonrelated) elements are unlikely to be present in the surrounding DNA. Fourth, elements are localized to short, easily sequenced segments of DNA and hence are more likely to be seen over random noise.

In this paper, we describe a modification of the random-selection technique (16, 17) that makes it possible to determine the sequence requirements for a DNA-binding protein. The method involves affinity chromatography for the in vitro selection of binding sites from synthetic random-sequence oligonucleotides. In vitro selection with a purified DNA-binding protein ensures that the selected sequences provide a function defined by a single protein-DNA interaction. By this method, the sequence recognition properties of the yeast GCN4 transcriptional activator protein have been accurately defined.

GCN4 is a 281-amino-acid protein that binds to the promoter regions of many yeast amino-acid-biosynthetic genes and activates their transcription during conditions of amino acid starvation (1, 6). GCN4 binds as a dimer, and the 60 C-terminal amino acids are sufficient for dimerization and for specific DNA binding (7, 8). At its extreme C terminus, GCN4 contains a leucine zipper motif (four leucine residues spaced seven residues apart) that is found in CEBP and several oncoproteins and has been proposed to be involved in dimerization (13). Moreover, the GCN4 DNA-binding domain shows about 45% sequence identity to the *jun* oncoprotein (26), and GCN4 and *jun* bind to similar DNA sites (24). The 25-residue region adjacent to the leucine zipper is highly conserved between GCN4 and *jun* and hence is likely to be involved in specific protein-DNA contacts.

The GCN4 recognition sequence has been investigated by detailed mutagenesis of the binding site in the wild-type *his3* promoter (5). Of the sequences analyzed, optimal binding was observed to DNA with the 9-base-pair (bp) symmetric sequence ATGA(C/G)TCAT, a sequence that differs by a single base pair from the wild-type *HIS3* site. This optimal binding site closely resembles putative binding sites from 15 GCN4-regulated genes. The symmetric nature of the recognition sequence and the fact that GCN4 binds as a dimer strongly suggest that the protein-DNA complex consists of two protein monomers interacting with adjacent half-sites. However, from the odd number of residues in the element and from the observation that mutation of the central C · G base pair in the *his3* site to G · C strongly reduced binding,

* Corresponding author.

we suggested that the GCN4 recognition sequence may have an inherent asymmetry (5, 8).

Current knowledge of the GCN4-binding site, although considerable, is insufficient to allow the prediction of protein binding simply by DNA sequence analysis. The relative importance of the positions in the 9-bp sequence and the contributions of flanking sequences is unclear, and the asymmetric nature of the binding site remains to be clarified. In the work described in this paper, we used a novel variation of the random selection method involving in vitro selection of binding sites to obtain a more accurate definition of the GCN4 recognition sequence. The results can be used for predicting GCN4-binding sites and are also discussed in terms of models for the protein-DNA interaction and of the general applicability of the method.

MATERIALS AND METHODS

Construction of DNA molecules. The DNA fragment encoding GCN4 was a derivative of pSP64-Sc4342 (7) that was adapted for expression in *Escherichia coli* by replacing sequences upstream of the initiator ATG codon with the oligonucleotide GAATTCCTCGACTCTAGAAAGGAGG TACGATC. This fragment was cloned into the *Eco*RI site of pRK7, a derivative of pKC30 (18) in which an *Eco*RI site has replaced the *Hpa*I site. The resulting plasmid, pRK7-GCN4, was introduced into *E. coli* C600(pJL23), which contains the temperature-sensitive λ repressor *c*I857, such that GCN4 synthesis can be induced by a temperature shift.

Purification of GCN4. GCN4 was purified from *E. coli* cells by a modification of an unpublished procedure developed by C. R. Wobbe in this laboratory. *E. coli* C600(pJL23)(pRK7-GCN4) cells (18 liters) were grown and induced for expression as described previously (19). All subsequent steps were carried out at 0 to 4°C. Cells were harvested, washed with 2 liters of 10% (wt/vol) sucrose–20 mM Tris hydrochloride (pH 8.0)–0.5 M NaCl–25 mM EDTA, suspended in 360 ml of 10% sucrose–20 mM Tris hydrochloride (pH 8.0)–0.5 M NaCl–1 mM EDTA, and frozen in liquid nitrogen. The cells were thawed, treated with lysozyme (final concentration, 0.2 mg/ml) for 105 min, and then subjected to two freeze-thaw cycles in liquid nitrogen. Cell debris were removed by centrifugation (32,500 × *g* for 30 min), and the supernatant was brought to 0.05% polyethyleneimine (Polymin P) and 1 mM phenylmethylsulfonyl fluoride. The extract was clarified by centrifugation at 15,000 × *g* for 20 min and then adjusted to 50 mM NaCl by 10-fold dilution with buffer A (20 mM Tris hydrochloride [pH 7.5], 1.0 mM EDTA, 10% glycerol, 0.5 mM dithiothreitol, 0.1 mM phenylmethylsulfonyl fluoride, 2.5 μg of antipain per ml, 0.2 μg of aprotinin per ml, 0.4 μg of pepstatin per ml, 0.1 mM benzamadine, 0.5 μg of leupeptin per ml).

A 3.6-liter volume of extract with a protein concentration of 0.6 mg/ml was applied to a 200-ml column of Whatman DE52 DEAE-cellulose equilibrated with buffer A containing 50 mM NaCl. The column was washed consecutively with 3 column volumes of buffer A containing 50, 100, and 300 mM NaCl. Protein from the 300 mM NaCl step (about 1.3 g) was dialyzed against two changes of buffer H (same as buffer A except that buffer H was 20 mM *N*-2-hydroxyethylpiperazine-*N'*-2-ethanesulfonic acid (HEPES)–NaOH [pH 7.5]) containing 50 mM NaCl and then chromatographed on a 200-ml column of Whatman P11 phosphocellulose which was equilibrated with buffer H containing 50 mM NaCl. The column was washed consecutively with 3 column volumes of buffer H containing 50, 300, and 1,000 mM NaCl. Protein from the 1,000 mM NaCl step (30 mg) was dialyzed against buffer H containing 50 mM NaCl.

A DNA affinity column was made by coupling double-stranded DNA containing oligomers of a GCN4-binding site (GGATGACTCATTTTTT) (5) to CNBr-activated Sepharose (Pharmacia Fine Chemicals) essentially as described previously (10). The 1,000 mM NaCl step fraction from phosphocellulose was chromatographed on a 50-ml oligonucleotide affinity column in the presence of 500 μg of poly(dI-dC) (Sigma Chemical Co.) as described previously (10). The column was washed with buffer H containing 0.2 M NaCl, and GCN4 was eluted with a linear gradient (10 column volumes) of NaCl (0.2 to 1.0 M) in buffer H. The peak of GCN4 activity, assayed by its ability to shift the electrophoretic mobility of a DNA fragment containing the *his3* promoter region (2, 3), eluted from the column at about 0.5 M NaCl. Fractions containing GCN4 were pooled, dialyzed against buffer H (without protease inhibitors) containing 50 mM NaCl, and concentrated on a 1.0-ml phosphocellulose column. GCN4 was eluted with buffer H containing 20% glycerol and 1.0 M NaCl in the absence of protease inhibitors. From 18 liters of starting culture, 600 μg of GCN4 was obtained. We estimate GCN4 to be 95% pure by analysis of the fraction after separation in a 10% polyacrylamide–sodium dodecyl sulfate gel.

Construction of the GCN4 affinity column. GCN4 (50 μg) was coupled to CNBr-activated Sepharose essentially as described previously (23). The final concentration of GCN4 on the column was 1 mg/ml of resin, with the coupling efficiency being greater than 95%.

Selection of oligonucleotides containing GCN4-binding sites. An oligonucleotide containing 23 nucleotides of random sequence DNA flanked by *Bam*HI and *Pst*I recognition sites was synthesized chemically (see Fig. 2). ^{32}P-labeled double-stranded oligonucleotide dimers (10 μg) were prepared by mutually primed synthesis at a specific activity of 10^5 cpm/μg (15). The oligonucleotide dimers (10^6 cpm) were suspended in 100 μl of buffer C (1 mM dithiothreitol, 10 μg of gelatin per ml, 50 mM Tris [pH 8]) containing 100 mM NaCl, 5 μg of tRNA, and 5 μg of poly(dI-dC) DNA as carrier and chromatographed on a 50-μl GCN4-Sepharose column. Approximately 7×10^5 cpm remained on the column after it had been washed with 3 column volumes of buffer C at 100 mM NaCl, and 22,000 cpm remained after it had been washed with 400 mM NaCl. Oligonucleotides with potential GCN4-binding sites were eluted in buffer C at 1 M NaCl, yielding 15,000 cpm, or 2% of the material bound at 100 mM NaCl. This 1 M fraction was concentrated by ethanol precipitation and digested with *Bam*HI and *Pst*I to generate oligonucleotide monomers suitable for cloning (see Fig. 2D). This initial selection with oligonucleotide dimers greatly reduced the number of restriction sites and hence the amount of restriction enzyme required for complete digestion. These oligonucleotide monomers were then chromatographed three more times as described above and concentrated by ethanol precipitation.

Cloning of oligonucleotides containing GCN4-binding sites. Owing to the small amounts of DNA selected on the GCN4-Sepharose column, a reverse blue-white screen was used to identify clones containing an oligonucleotide insert. The vector was a derivative of pTZ19 that contained a 1,500-bp insertion in the polylinker and was thus a nonfunctional *lacZ* gene; cells containing this plasmid generated white colonies on 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal) plates. Since the oligonucleotide contained 23 bases of random-sequence DNA flanked by *Bam*HI and *Pst*I sites,

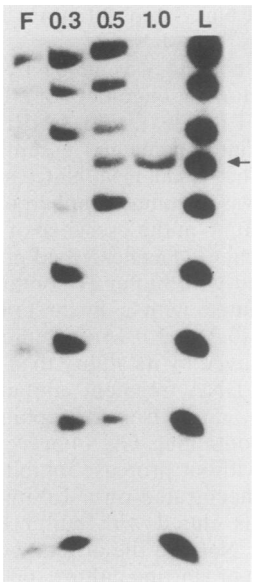

FIG. 1. Selective binding of *his3* DNA to GCN4-Sepharose. Plasmid DNA containing a known GCN4-binding sequence was restricted with *Msp*I and *Bam*HI, labeled with ^{32}P at the 5′ ends, and applied to a 50-μl affinity column in buffer A containing 0.1 M NaCl. The column was eluted in succession with buffer A containing 0.1, 0.3, 0.5, and 1.0 M NaCl, and the DNA eluted after each step was separated by polyacrylamide gel electrophoresis and visualized by autoradiography. The DNA-loaded (lane L) and flowthrough (lane F) fractions are also shown. The arrow indicates the fragment containing the known GCN4-binding site.

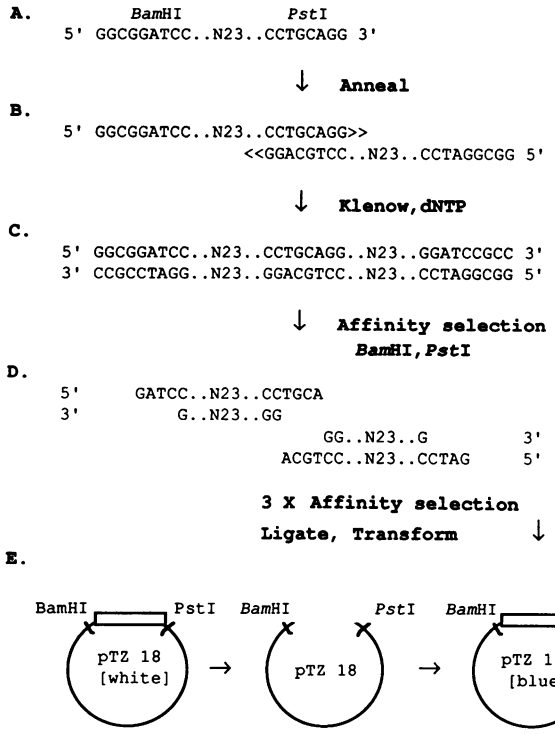

FIG. 2. Cloning and selection of binding sites. (A to D) Generation of double-stranded DNA from single-stranded oligonucleotides by using the technique of mutually primed synthesis (15). Oligonucleotides containing the protein-binding site were purified by affinity chromatography on the GCN4-Sepharose once between steps C and D and three times between steps D and E. These oligonucleotides replace an insert in the pTZ19 vector to regenerate the *lacZ* reading frame and create blue bacterial colonies in step E.

successful cloning would regenerate the correct translational coding frame for functional β-galactosidase and produce blue colonies. This reverse blue-white screen produced a very low background. No blue colonies were seen in ligations without insert DNA, and all blue colonies characterized from ligations with oligonucleotides contained the expected insertion. Although the white-blue screen might prevent the cloning of certain oligonucleotides, it is very unlikely to introduce a systematic bias into the collection of binding sites. Single-stranded DNAs from 43 blue colonies were generated by superinfection with bacteriophage M13-KO7 and sequenced by the chain termination method (21).

RESULTS

Selection of GCN4-binding sites. We have used a GCN4-Sepharose column to select GCN4-binding sites from random-sequence oligonucleotides. GCN4 was synthesized in *E. coli*, purified to homogeneity by affinity chromatography, and coupled to CNBr-activated Sepharose. To demonstrate the functional integrity of GCN4 on the Sepharose column and the feasibility of selecting binding sites from random DNA, restriction fragments of a plasmid containing a GCN4 binding site were loaded on the column and eluted with increasing NaCl concentrations. The fragment containing the GCN4-binding site was selectively retained on the column at 500 mM NaCl (Fig. 1).

Random-sequence oligonucleotides, prepared by mutually primed synthesis (15) (Fig. 2), were chromatographed on the GCN4-Sepharose column in the presence of carrier DNA as described in Materials and Methods. Nonspecifically bound DNA was eluted from the column by using 400 mM NaCl washes, and the specifically bound oligonucleotides were eluted with 1 M NaCl and then chromatographed through three additional cycles (Fig. 2). Approximately 2% of the input DNA was present in the initial 1 M NaCl fraction, indicating a purification of about 50-fold. In the three successive cycles of affinity chromatography, we observed eight-, three-, and twofold purification, with a final yield of 0.01% of the input DNA. These affinity-selected oligonucleotides were cloned by using a white-to-blue *lacZ* screen to identify the correct colonies and then sequenced.

Sequence analysis of GCN4-binding sites. The sequences of the 43 independently derived, affinity-selected oligonucleotides are shown in Fig. 3. Strikingly, 40 of these oligonucleotides contain a common 7-bp dyad-symmetric sequence, TGA(C/G)TCA, and the remaining 3 contain sequences differing at only one of the seven positions. This 7-bp core sequence was used to align all 43 sequences. By using the alignment in which the C residue in the central position has been chosen arbitrarily, the frequencies of each of the bases at 27 positions around the GCN4-binding element are shown in Fig. 4. The central base pair is defined as position 0, such that symmetrically disposed base pairs have the same absolute value (positive for positions to the right and negative for positions to the left).

Analysis of the region flanking the core indicates that three more positions on each side (±4, 5, and 6) have nucleotides that occur at frequencies significantly different from those expected by chance, although the preferences at these positions are much less stringent than those at the core (Fig. 4). A log of the odds (LOD) score of 3.0 in a matrix position

	*Bam*HI		*Pst*I
	---GGATCC...		...CCTGCAGG---
1,	----------G	TGAGTCA	CCATCCCGGTTGGC---
2,	--------GACAGGTC	TGAGTCA	TATGCACG---------
3,	----------GAGCAA	TGAGTCA	TCGTGTCGCG-------
4,	----------GA	TGAGTCA	CAGGCAACGGGCAC---
5,	----------A	TGACTCA	TTGAGCGACTTACCG--
7,	CTGTATTACGACACAA	TGAGTCA	----------
8,	--ATAACGAGTGGGGA	TGACTCA	TT----------
9,	-ATTCTCGCCTTTCTG	TGAGTCA	T----------
11,	----------AAAAAA	TGAGTCA	TCCGAGCT---------
12,	----------AT	TGACTCA	TACGCAGCTAGACT---
13,	-------GAGTGTAGA	TGACTCA	TGGACTG----------
14,	----------A	TGAGTCA	TCGCTAGTCCATGGG--
15,	---GTGTCCTTCGGGA	TGAGTCA	TGC----------
16,	-----GTCACGGGGCC	TGACTCA	TAGAA----------
18,	----------G	TGAGTCA	CGGAAATTGTTGG----
19,	-----GCATTATGGAG	TGACTCA	TCCTT----------
20,	------GAAGCATTTG	TGAGTCA	TCGCT----------
21,	---TCGGTAGTCGGTA	TGAGTCA	TT----------
22,	----------CGG	TGACTCA	CGTAGAGGTAACC----
25,	----------A	TTAGTCA	TCAGAGGGTGGCGAC--
26,	--AAATTTACATGCGA	TGAGTCA	TA----------
27,	--AGCCGCGGTGAGAA	TGAGTCA	TA----------
28,	--TGGCCCCGGGTCTC	TGACTCA	GC----------
29,	----AACGAGACGCGC	TGAGTCA	TCTT----------
30,	--GCCAGTTGATACTC	TGTGTCA	CGG----------
32,	----GTAGCTGAGGAG	TGAGTCA	CGCCTG----------
33,	----------GGGG	TGAGTCA	TAAAGATAAATCT----
34,	-CCCGCGTAGGCTCGA	TGAGTCA	A----------
35,	--TCAGAGTCAGCTTA	TGAGTCA	GG----------
36,	----------CGCTGG	TGACTCA	TCGTGTTCTG-------
37,	--GCAGGCGCCACGCG	TGACTCA	T----------
38,	-------GTCAGTTTG	TGAGTCA	CTCTACC----------
39,	----GTATGACGTGGA	TGTGTCA	GAGG----------
40,	--------TGGCCCTA	TGACTCA	TAAGGCAC---------
41,	--GCCAATGACTTCTC	TGAGTCA	T----------
42,	--CCGTCATCGCGGGT	TGAGTCA	CT----------
43,	----------GA	TGAGTCA	GGACCGGGGTTGGA---
44,	----------TACGTG	TGACTCA	TTCACCACCC-------
45,	----------GCCTG	TGACTCA	TCCTGCGCGTA------
46,	----------GAAATA	TGAGTCA	CGGGACGTCT-------
47,	-AGTGTGTAAAGGGGG	TGAGTCA	T----------
48,	----------AG	TGACTCA	CCGACACGTACGT----
49,	-------TGCGTCGGG	TGAGTCA	GATTGTG----------

FIG. 3. GCN4-binding sites. The 43 affinity-selected oligonucleotides containing GCN4-binding sites are shown as sequenced between the *Bam*HI and *Pst*I restriction endonuclease sites.

means that the probability of the observed preferences occurring by chance is 1 in 1,000 and hence represents a significant contribution to the element. The involvement of positions −4, +4, +5, and +6 is well documented, with the lowest LOD score being 2.8. Positions −5 and −6 are also probably significant, because the nucleotide preferences are symmetrically related to the corresponding positions +5 and +6. At positions farther from the center (±7 and beyond), the frequencies of nucleotide occurrence are essentially random, with the possible exception of positions +10 and −10.

Although the most preferred nucleotides are the same on each side of the element, the frequencies of the preferred bases are affected significantly according to their side. In the central seven positions, the three deviations from TGAC TCA (Fig. 3, derivatives 25, 30, and 39) all occur to the right of the central C. However, beyond these seven positions, this sidedness is reversed (Fig. 4). The nucleotide preferences of the three outer positions to the right of the center (the sum of the LOD scores for +4, +5, and +6 is 13.1) are stronger than those in the equivalent positions to the left (the sum of the LOD scores for −4, −5, and −6 is 5.9).

DISCUSSION

The GCN4 recognition sequence. In the work described in this paper, we used an in vitro selection method for isolating GCN4-binding sites from random-sequence oligonucleotides. By comparing these binding sites, we can determine the nucleotide sequence requirements for GCN4 binding. A significant deviation from a random distribution of the four nucleotides indicates that a position has an effect on GCN4 binding, and the extent of that deviation is an indication of the magnitude of the effect. Although we have not individually assayed each selected oligonucleotide for a GCN4-binding site, the stringency of the selection and the striking similarity between the sequences indicate that this is the case.

In general, the results obtained by random selection of binding sites agree with those of previous experiments which examined the functionality of mutants of the GCN4 binding site in the wild-type *his3* promoter (5). However, several features of the GCN4 recognition sequence have been refined by this approach. First, it is now clear that the central 7 bp (positions −3 to +3) are the most important for GCN4 binding; previously, positions +4 and −4 were proposed to be nearly as important. Second, sequence preferences at positions ±4, ±5, and ±6 have been established, indicating that the GCN4 recognition site extends for 13 bp. Third, the results confirm the previous suggestion that the GCN4 site is inherently asymmetric (5, 8). In this regard, it is noteworthy that the binding site in the wild-type *his3* promoter has its

	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12	+13
G	2	5	2	9	6	10	7	9	11	15	0	43	0	0	0	0	0	5	2	13	4	11	4	11	6	3	4
A	4	2	7	2	3	5	5	4	6	19	0	0	43	0	2	1	43	1	12	4	12	6	7	4	5	8	6
T	2	1	3	2	8	6	4	5	11	2	43	0	0	0	41	0	0	26	8	3	8	7	10	5	6	7	6
C	6	7	6	7	6	4	11	12	9	6	0	0	0	43	0	42	0	11	16	14	10	9	10	8	10	7	7
SUM	14	15	18	20	23	25	27	30	37	42	43	43	43	43	43	43	43	43	38	34	34	33	31	28	27	25	23
LOD	.66	1.4	.86	1.7	.52	.67	.88	1.2	.42	4.3	26	26	26	26	22	24	26	7.4	2.9	2.8	.98	.38	.75	.92	.44	.59	.19
														C_G	T_C		A_T										

FIG. 4. Nucleotide use matrix. Twenty-seven positions of the aligned oligonucleotides from Fig. 3 are shown, with the number of occurrences for the four nucleotides in each position. The strands are aligned such that the central position (defined as 0) contains a C. SUM indicates the total number of nucleotides in each position, and LOD indicates the $\log_{10}$ of the probability that the nucleotides would have the observed pattern if distributed randomly (17). A LOD of 3.0 would occur in 1 of 1,000 positions if the nucleotides were distributed randomly. The probability of observing 43 of the same nucleotides in one position is 10^{-26}. The most common nucleotides for each position are shown below the matrix.

single deviation to the right of the central base pair and that mutations at +4 have stronger effects on DNA binding than do symmetrical mutations at −4 (5).

The binding sites generated in our work by selection in vitro allow one to differentiate between sequences that are necessary for protein binding and those that are associated with the in vivo binding sites for other reasons. The GCN4-binding site in the *his3* promoter contains eight consecutive dT · dA residues just downstream from the core, and other wild-type GCN4-binding sites show a preference for downstream dT · dA base pairs (5). Moreover, deletions into the *his3* dT · dA region can reduce the level of transcriptional activation by GCN4 (5). The results here indicate that the dT · dA residues are unlikely to be important for the inherent ability of GCN4 to bind to DNA and hence that they may play some other role in vivo.

The specificity observed in the selected GCN4-binding sites was greater than that seen in natural yeast promoters whose transcription is regulated by GCN4 (5). This indicates that the in vitro selections for GCN4 binding were more stringent than required for binding and transcriptional activation in vivo. By modifying the elution conditions, selections could be made to generate binding sites with different levels of functions. Selections for the highest levels of functions are most informative in defining the best nucleotides at each position and the number of positions in the binding site. Selections for lower levels of functions are more useful for defining the degree and type of degeneracy allowed. To better define the degeneracy allowed in the central 7 bp of the GCN4-binding site, it would be necessary to select binding sequences under conditions of lower stringency.

Nature of the interaction between GCN4 and its target site. The extreme sequence preference in the 7-bp core suggests that all these positions are involved in direct contact with GCN4. The more modest effects at the adjacent positions may also reflect direct contacts to GCN4, but presumably these would contribute less to the overall affinity. Alternatively, these outside positions might affect GCN4 binding by a more indirect mechanism, such as observed in the interaction between the bacteriophage 434 repressor and the center of its operator (12).

The importance of the central C · G base pair and the asymmetry of the GCN4 recognition sequence strongly support the model that GCN4 dimers bind to nonequivalent half-sites (5, 8). It seems likely that asymmetrical contacts made with the central C · G base pair cause the GCN4 dimer to be shifted from the center of the site. In the 7-bp core, GCN4 probably interacts more avidly with the left half-site (positions −1, −2, and −3) than with the right half-site (positions +1, +2, and +3), because deviations generally occur to the right of the central base. Support for the hypothesis that the left GCN4 monomer interacts with the central C · G base pair comes from the fact that GCN4 can bind the sequence TGACGTCA but not TGAGCTCA (J. W. Sellers and K. Struhl, manuscript in preparation). In contrast to the relative importance of the left side of the core, flanking positions in the right half-site (positions +4, +5, and +6) contribute more to GCN4 binding than equivalent positions in the left half-site (positions −4, −5, and −6) do, perhaps to compensate for the relative weakness of the right side of the core.

Comments on the random-selection method. An important consideration in applying the random-selection method is the specificity of the protein-binding site, i.e., the frequency of the site in random-sequence DNA. A mathematical analysis

A.	4	5	6	7	8	9	specificity
B.	256	1024	4096	16384	65536	262144	frequency
C.	20	19	18	17	16	15	pos./oligo
D.	13	54	227	964	4096	17476	oligos/site
E.	780	190	40	10	2.4	0.6	ng remaining
F.	3.7	5.8	7.8	9.9	12	14	2x selection
G.	1.1	1.7	2.4	3.0	3.6	4.2	10x selection
H.	0.6	0.9	1.2	1.5	1.8	2.1	100x selection

FIG. 5. Numerical analysis of binding-site selection. (A) Number of sequence-specific bases in a potential binding site. (B) Average length of random-sequence DNA containing one binding site. (C) Number of potential positions in an oligonucleotide 23 bases long. (D) Number of oligonucleotides required to contain one binding site (i.e., line B divided by line C). (E) Number of nanograms of oligonucleotide remaining after selection from 10 μg. (F to H) Number of selections required so that 50% of the oligonucleotides contain the desired site if each selection is 2-, 10-, or 100-fold efficient (F, G, and H, respectively).

indicates that the method described here should be generally applicable for defining the nucleotide sequence recognition properties for many DNA-binding proteins (Fig. 5). The 10 μg of double-stranded oligonucleotides contains 2.3×10^{14} molecules before selection. Although the mass of DNA remaining after repeated selections is small, the number of molecules remaining is adequate to define the site. For example, selection for an element with a specificity of 9 bp could yield a maximum of 0.6 ng of DNA, but this represents 1.3×10^{10} different molecules. With a 50- to 100-fold discrimination between binding sites and random-sequence oligonucleotides per chromatographic cycle (such as is observed for the GCN4 column), three or four cycles should provide sufficient purification that essentially all of the oligonucleotides contain a binding site.

However, elements whose sites occur very rarely in random-sequence DNA may be more difficult to characterize. The two problems that are encountered with these elements involve the ability to purify the rare sequences that are bound and the ability to clone the remaining small quantities of DNA. However, since repeated selections for very specific elements should still yield a significant number of molecules, the DNA could be amplified between selections by using the polymerase chain reaction (20). An oligonucleotide hybridizing to the 5′ end of the oligonucleotide dimers created by the mutually primed synthesis reaction (Fig. 2B) could be used as a primer for the polymerase chain reaction. The final selections would be done without amplification after dimers were reduced to monomer oligonucleotides. Repeated selections and amplifications involving the use of the polymerase chain reaction would allow sites with a very high specificity to be accurately defined. Alternatively, a less stringent selection could be performed to isolate oligonucleotides containing nonoptimal binding sequences. Although the selected binding sites will have a lower affinity for the protein, statistical analysis should still reveal the commonly occurring sequences, even if the specificity is distributed over many positions.

If the restriction endonuclease sequences at the edge of the random-sequence DNA contain a portion of the genetic element, the selected elements will be found predominantly in a fixed position with respect to one end of the oligonucleotide. With the GCN4-binding sites shown in Fig. 3, there is no such association with either end of the random-sequence region. If such a problem arose, the experiment could be repeated with other restriction endonuclease sites to clone the random-sequence DNA.

These results of using a protein affinity column show that the selections were suitable for purifying the rare oligonu-

cleotides which contain a binding site. Other techniques for isolating protein-DNA complexes may work equally well. The protein-DNA complexes might be isolated from free oligonucleotides by filter binding, by immunoprecipitation with antibodies to the DNA-binding protein (9), or by gel isolation of the DNA-protein complexes (2, 3). The protein preparation used for complex formation does not have to be pure, but it must not contain other DNA-binding proteins or DNA-modifying enzymes, which would affect the clonability of the oligonucleotides.

The success of this approach in rapidly and accurately defining the specificity of the GCN4-binding site indicates that the method can be used as a part of the standard classification of most DNA-binding proteins and will contribute to defining the structure and function of these protein-DNA interactions. Knowledge of the optimal binding site for DNA-binding proteins will also facilitate their rapid cloning by probing expression libraries with oligonucleotides containing optimal binding sites (22, 25). With the rapid accumulation of sequence data, it will be important to be able to predict functions associated with particular sequences. The nucleotide-use matrices generated by random selection can define a genetic element such that its presence in other DNAs can be accurately predicted. In addition, the method described here should be useful for selecting binding sites from genomic DNA that might represent in vivo targets of the protein.

ACKNOWLEDGMENTS

We thank Rick Wobbe for his unpublished procedure for purifying GCN4 from *E. coli* cells and Victoria Singer for suggesting the potential use of the polymerase chain reaction.

This work was supported by a postdoctoral fellowship to C.J.B. from the Medical Research Council of Canada and by Public Health Service grant GM30186 to K.S. from the National Institutes of Health.

LITERATURE CITED

1. **Arndt, K., and G. R. Fink.** 1986. GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5′ TGACTC 3′ sequences. Proc. Natl. Acad. Sci. USA **83**:8516–8520.
2. **Fried, M., and D. Crothers.** 1981. Equilibrium and kinetics of *lac* repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic Acids Res. **9**:6505–6525.
3. **Garner, M., and A. Revzin.** 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *E. coli* lactose operon regulatory system. Nucleic Acids Res. **9**:3047–3060.
4. **Gronostajski, R. M.** 1987. Site-specific DNA binding of nuclear factor I: effect of the spacer region. Nucleic Acids Res. **15**: 5545–5559.
5. **Hill, D. E., I. A. Hope, J. P. Macke, and K. Struhl.** 1986. Saturation mutagenesis of the yeast *HIS3* regulatory site: requirements for transcriptional induction and for binding by GCN4 activator protein. Science **234**:451–457.
6. **Hope, I. A., and K. Struhl.** 1985. GCN4 protein, synthesized in vitro, binds to *HIS3* regulatory sequences: implications for the general control of amino acid biosynthetic genes in yeast. Cell **43**:177–188.
7. **Hope, I. A., and K. Struhl.** 1986. Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. Cell **46**:885–894.
8. **Hope, I. A., and K. Struhl.** 1987. GCN4, a eukaryotic transcriptional activator protein, binds as a dimer to target DNA. EMBO J. **6**:2781–2784.
9. **Johnson, A. D., and I. Herskowitz.** 1985. A repressor (*MATα2* product) and its operator control expression of a set of cell type specific genes in yeast. Cell **42**:237–247.
10. **Kadonaga, J. T., and R. Tjian.** 1986. Affinity purification of sequence-specific DNA binding proteins. Proc. Natl. Acad. Sci. USA **83**:5889–5893.
11. **Kaiser, C. A., D. Preuss, P. Grisafi, and D. Botstein.** 1987. Many random sequences functionally replace the secretion signal sequence of yeast invertase. Science **235**:312–317.
12. **Koudelka, G. B., P. A. B. Harbury, S. Harrison, and M. Ptashne.** 1988. DNA twisting and the affinity of bacteriophage 434 operator for bacteriophage 434 repressor. Proc. Natl. Acad. Sci. USA **85**:4633–4637.
13. **Landschulz, W. H., P. F. Johnson, and S. L. McKnight.** 1988. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science **240**:1759–1764.
14. **Ma, J., and M. Ptashne.** 1987. A new class of yeast transcriptional activators. Cell **51**:113–119.
15. **Oliphant, A. R, A. L. Nussbaum, and K. Struhl.** 1986. Cloning of random-sequence oligodeoxynucleotides. Gene **44**:177–183.
16. **Oliphant, A. R, and K. Struhl.** 1987. The use of random-sequence oligonucleotides for determining consensus sequences. Methods Enzymol. **155**:568–582.
17. **Oliphant, A. R, and K. Struhl.** 1988. Defining the consensus sequence of *E. coli* promoter elements by random selection. Nucleic Acids Res. **16**:7673–7683.
18. **Rao, R. N.** 1984. Construction and properties of plasmid pKC30, a pBR322 derivative containing the p_L-*N* region of phage lambda. Gene **31**:247–250.
19. **Rosenberg, M., Y. S. Ho, and A. R. Shatzman.** 1983. The use of pKC30 and its derivatives for controlled expression of genes. Methods Enzymol. **101**:123–138.
20. **Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich.** 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239**:487–491.
21. **Sanger, F., A. R. Coulson, B. G. Barell, A. J. Smith, and B. A. Roe.** 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. **143**:161–178.
22. **Singh, H., J. H. LeBowitz, A. S. Baldwin, and P. A. Sharp.** 1988. Molecular cloning of an enhancer binding protein: isolation by screening of an expression library with a recognition site DNA. Cell **52**:415–423.
23. **Sopta, M., R. W. Carthew, and J. Greenblatt.** 1985. Isolation of three proteins that bind to mammalian RNA polymerase II. J. Biol. Chem. **260**:10353–10360.
24. **Struhl, K.** 1987. The DNA-binding domains of the jun oncoprotein and the yeast GCN4 transcriptional activator are functionally homologous. Cell **50**:841–846.
25. **Vinson, C. R., K. L. LaMarco, P. F. Johnson, W. H. Landschulz, and S. L. McKnight.** 1988. In situ detection of sequence-specific DNA binding activity specified by a recombinant bacteriophage. Genes Dev. **2**:801–806.
26. **Vogt, P. K., T. J. Bos, and R. F. Doolittle.** 1987. Homology between the DNA-binding domain of the *GCN4* regulatory protein of yeast and the carboxy-terminal region of a protein coded for by the *onc* gene *jun*. Proc. Natl. Acad. Sci. USA **84**:3316–3319.