Quantifying DNA–protein interactions by double-stranded DNA arrays

Martha L. Bulyk[1], Erik Gentalen[2], David J. Lockhart[2], and George M. Church[1]*

*[1]Harvard University Graduate Biophysics Program and Harvard Medical School Department of Genetics, Boston, MA 02115. [2]Affymetrix, Santa Clara, CA 95051. *Corresponding author (e-mail: church@salt2.med.harvard.edu).*

Received 2 February 1999; accepted 1 April 1999

We have created double-stranded oligonucleotide arrays to perform highly parallel investigations of DNA–protein interactions. Arrays of single-stranded DNA oligonucleotides, synthesized by a combination of photolithography and solid-state chemistry, have been used for a variety of applications, including large-scale mRNA expression monitoring, genotyping, and sequence-variation analysis. We converted a single-stranded to a double-stranded array by synthesizing a constant sequence at every position on an array and then annealing and enzymatically extending a complementary primer. The efficiency of second-strand synthesis was demonstrated by incorporation of fluorescently labeled dNTPs (2′-deoxyribonucleoside 5′-triphosphates) and by terminal transferase addition of a fluorescently labeled ddNTP. The accuracy of second-strand synthesis was demonstrated by digestion of the arrayed double-stranded DNA (dsDNA) on the array with sequence-specific restriction enzymes. We showed *dam* methylation of dsDNA arrays by digestion with *Dpn*I, which cleaves when its recognition site is methylated. This digestion demonstrated that the dsDNA arrays can be further biochemically modified and that the DNA is accessible for interaction with DNA-binding proteins. This dsDNA array approach could be extended to explore the spectrum of sequence-specific protein binding sites in genomes.

Keywords: dsDNA arrays, restriction enzymes, DNA–protein interactions

Sequence-specific DNA binding by proteins controls transcription[1], recombination[2], restriction[3], and replication[4]. Sequence requirements are usually determined by assays that measure the effects of mutations on binding of DNA and amino acid residues implicated in these interactions. These assays, which include nitrocellulose binding assays[5], gel shift analysis[6], Southwestern blotting[7,8], or reporter constructs in yeast[9], are usually considered too laborious for the analysis of many DNA variants. Therefore, we have developed a highly parallel method for studying the sequence specificity of DNA–protein interactions.

We have taken advantage of oligonucleotide arrays, or DNA arrays, that have previously been used for mRNA expression analysis[10–12], polymorphism analysis[13–16], deletion strain analysis[17], and for identifying clones from genetic selections[18]. However, the arrays used for these purposes contain single-stranded DNA (ssDNA) oligonucleotides, and most sequence-specific regulatory DNA-binding proteins bind double-stranded DNA (dsDNA). Therefore, we present a method for enzymatically converting ssDNA arrays into arrays of duplex DNA. Sequence-specific digestion at the cognate restriction sites has been demonstrated using restriction-enzyme digestion of dsDNA arrays. In addition, we show that the dsDNA can be altered biochemically. Arrays of biochemically modified DNA may be useful for applications that seek to determine the effects of modifications, such as methylation, on sequence-specific binding. The results presented here suggest that these dsDNA arrays will be well suited for the analysis of DNA–protein interactions, particularly for the discovery of the sequences recognized by transcription factors and the quantitative assessment of those important interactions.

Results and discussion

Second-strand synthesis. ssDNA arrays were made on an Affymetrix (Santa Clara, CA) DNA array synthesizer. A constant sequence was synthesized before any variable sequences were introduced, and these strands were used as templates for enzymatic second-strand synthesis. A primer complementary to the constant sequence was used in primer extension reactions, producing all the second strands on the array in a single enzymatic reaction.

For our experiments, there are a number of advantages to creating dsDNA via primer extension instead of by chemically synthesizing single-stranded, self-complementary oligonucleotides[19]. First, 5′-(4,4′-dimethoxytrityl) (DMT) synthesis occurs with higher efficiency than that achieved with light-directed, 5′-(α-methyl-2-nitropiperonyl)oxycarbonyl (MeNPOC)[20,21] synthesis. Therefore, longer strands of dsDNA can be made because only half as many nucleotides need to be produced by light-directed synthesis when the complementary strand is created via primer extension. Second, the exact complement of each template strand, including any degenerate nucleotides synthesized into the first strand, will be made because the Klenow fragment of DNA polymerase I is a highly processive polymerase with an error rate of approximately 10^{-5}. Third, this mode of second-strand synthesis ensures a low mismatch rate as creation of dsDNA does not rely upon annealing a complex mix of exogenous complementary sequences.

In order to verify initially that the primer was annealing to all sequences, a fluorescein-labeled primer was hybridized to the array, and signal intensity was seen over the entire chip (data not shown). Subsequently, unlabeled primers were used in all primer-extension reactions. To confirm enzymatic extension of the primer, we included fluorescein-labeled dATP in a reaction along with unlabeled 2′-deoxyribonucleoside 5′-triphosphates (dNTPs) (Figs. 1 and 2A). As expected, there tended to be higher signal intensity in features with a greater proportion of adenine in the second strand (Fig. 3B). Of the features with identical subsites, those with longer spacers had higher signal intensities, as expected, because longer spacers allowed a greater number of fluorescein-labeled dATPs to be incorporated.

The duplex DNA also can be end-labeled after synthesis (Fig. 2B) instead of being labeled by incorporation of fluorescein-tagged dNTPs. In this scheme, only unlabeled dNTPs were used in the

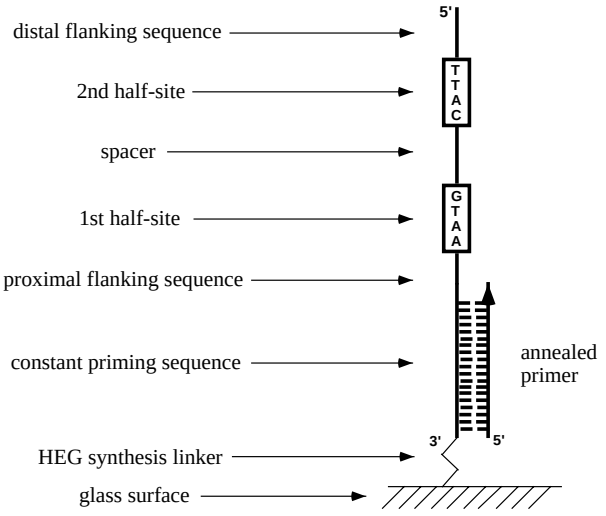

Figure 1. Generalized array strand (not to scale). DNA is attached to the glass surface by either 1 or 2 hexaethylene glycol (HEG) synthesis linkers. The sequences of the two priming sequences that were used are:

1s: 3′-AGCTGTGCGAGGTTGT-5′
2s: 3′-CCTGGCTAACTGAACT-5′

primer-extension reactions. The 3′-ends of the newly synthesized strands were then end-labeled by addition of fluorescein-labeled ddNTP with terminal transferase (Fig. 4A). Only the 3′-ends of the second strands were available for addition in these terminal transferase reactions, because the 3′-ends of the first strands were covalently attached to hexaethylene glycol (HEG) linkers. The observed variation in signal intensity from row to row was due to either different synthesis efficiencies or different efficiencies of terminal transferase addition for different sequences.

Restriction enzyme digestion. To determine that the duplex DNA was both physically accessible and of proper structure for interaction with a protein, we digested dsDNA arrays with a restriction enzyme. This also confirmed that the second strands were synthesized correctly. A restriction enzyme with a 4 bp recognition site was chosen because the two subsites on the arrays were each either 3 or 4 bp long, although the design of the array can be changed according to the particular type of restriction enzyme being studied. The fluorescein-labeled dNTP included in the primer-extension reaction was chosen to be distal to the cleavage site (relative to the glass surface), so that after digestion the fluorescent label that had been incorporated into the second strand would be released (Fig. 3A). For end-labeled dsDNA arrays, the signal was distal to the cleavage site irrespective of the restriction site.

Strand density and the distance of the strands from the array surface were varied to measure the effects of accessibility of the DNA strands for primer-extension reactions and enzymatic digestions. The distance from the surface was varied using either one or two HEG linkers. The two HEG linkers were expected to make the duplex DNA more flexible and more accessible by reducing steric hindrance from the glass surface and neighboring molecules. An array with variable densities and number of linkers was extended in the presence of fluorescein-labeled dATP, then digested with *Rsa*I (Fig. 3B). As *Rsa*I digestion leaves blunt ends between the T and the A of its recognition site (5′-GTAC-3′), incorporated label is lost with the portion of the strand that is released.

Signal intensity loss was evaluated by calculating a *z* score for each feature. This statistic measures the amount of signal intensity loss beyond that due to photobleaching or other effects that might cause general signal intensity loss over the whole array. The average *z* score in the 30 features containing the *Rsa*I recognition site was 7 ($p = 3 \times 10^{-13}$) A graph of the *z* scores for these features (Fig. 5) indicates that arrays made with two HEG linkers and a strand density of about 0.6 pmol/cm^2 achieved the most highly significant signal intensity loss ($z = 9$, $p = 7 \times 10^{-20}$). No other subsites showed a significant loss in signal intensity, indicating that the digestion was sequence-specific. The 4-mer subsite with the second highest *z* score was 5′-GGTA-3′ ($z = 3$, $p = 1 \times 10^{-3}$). As the spacer (0–14 nucleotides) consisted of degenerate sequence, about one-fourth of the nucleotides incorporated after the 5′-GGTA-3′ subsite were C, completing the restriction site for *Rsa*I. Therefore, loss of signal at this site was not due to nonspecific cutting; sites containing dsDNA sequences unrelated to the *Rsa*I site had even lower *z* scores.

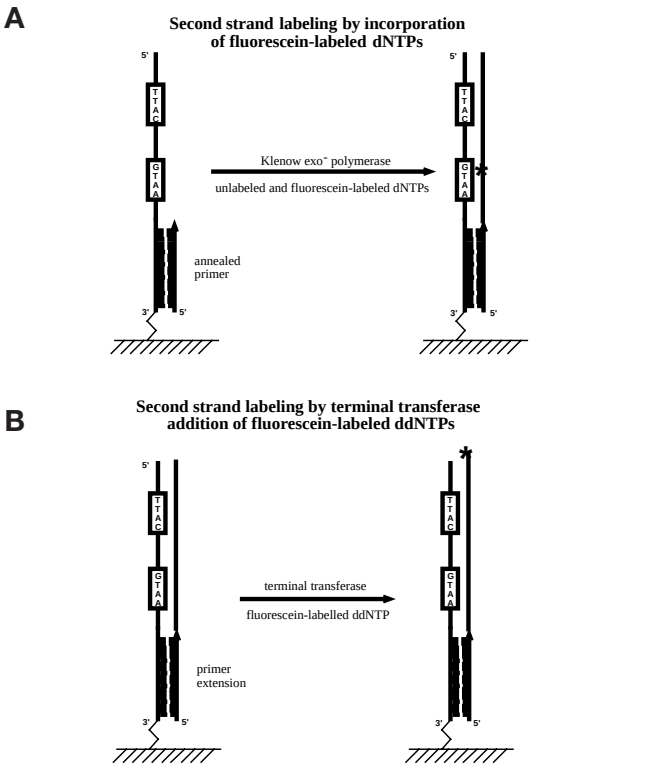

Figure 2. Alternative methods for second-strand labeling. (A) Schematic for labeling by incorporation of fluorescently labeled dNTPs and (B) schematic for end-labeling with fluorescently labeled ddNTPs.

The region surrounding the 5′-GTAC-3′ features (after digestion with *Rsa*I) is shown in Figure 4B. There was significant loss in signal intensity (average $z = 8$, $p = 4 \times 10^{-16}$) in the 30 features in which the first subsite was 5′-GTAC-3′. As for the 4-mer subsite 5′-GGTA-3′, about one-fourth of the nucleotides incorporated after the 5′-GTA-3′ subsite were C, completing the restriction site for *Rsa*I. Thus, in addition to a significant loss of signal intensity at the features corresponding to a 5′-GTAC-3′ first subsite, there was also a partial loss of intensity at those features that contained only 5′-GTA-3′ in the first subsite (average $z = 3$, $p = 1 \times 10^{-3}$).

Data from these array-based assays will be useful in modeling the binding mechanisms of restriction enzymes. Because restriction enzymes have very stringent sequence requirements, they have been used as model systems to study the sequence specificity of DNA–protein interactions[22–28]. However, experimental data on mutant proteins or DNA sites can be interpreted with greater accuracy if the interactions governing sequence specificity are understood at the atomic level. Such understanding comes from studying the structure of the protein bound to DNA. The set of restriction enzymes cocrystallized with DNA includes *Eco*RI[26], *Eco*RV[28], *Bam*HI[29], *Pvu*II[30], *Fok*I[27], and *Bgl*I[31]. Moreover, measuring DNA–pro-

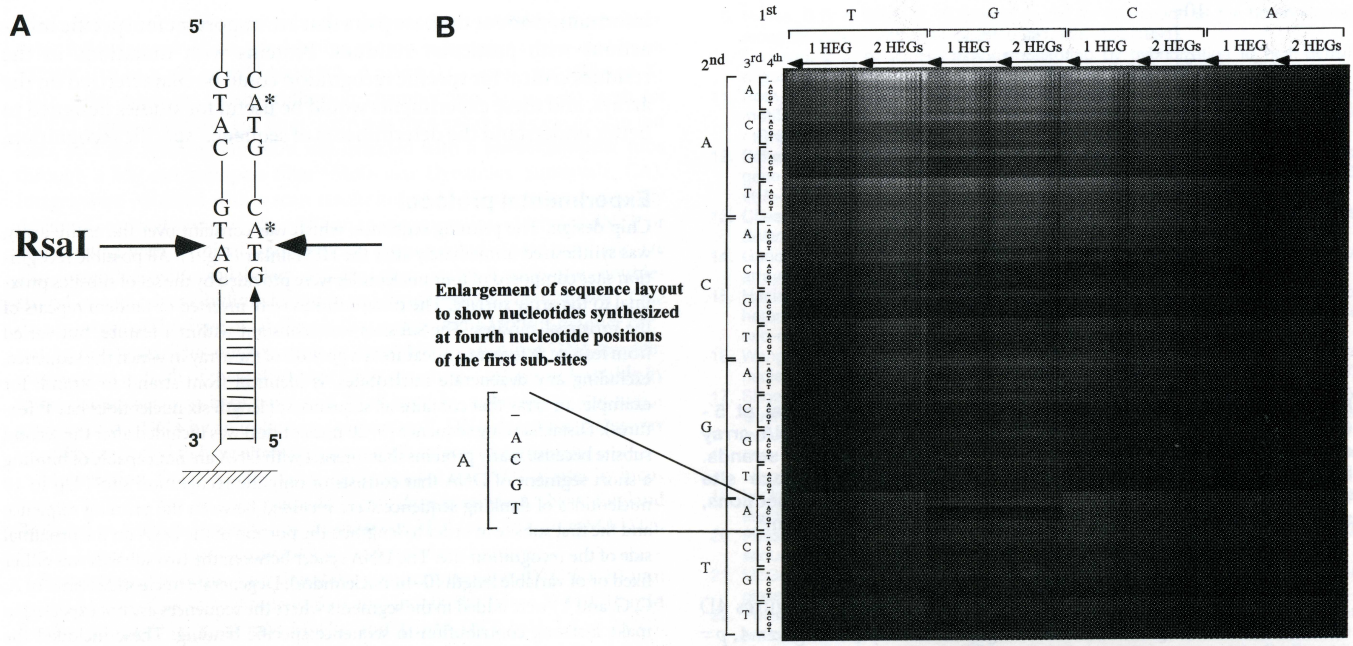

Figure 3. (A) Schematic of *Rsa*I digestion of an array of dsDNA oligonucleotides labeled with fluorescein-12-dATP. (B) *Rsa*I digestion of a variable strand density array labeled with fluorescein-12-dATP. The second strands of an array with both one and two HEG synthesis linkers were synthesized using primer 1s in the presence of a 3:1 ratio of unlabeled to labeled dATP in a primer-extension reaction. The subsites were in an inverted repeat orientation with no proximal flanking sequence, 5 bp distal flanking sequence, and a spacer of constant length (6 bp). The arrows under the labels "1 HEG" and "2 HEGs" indicate increasing first-strand density, except for the 15th and 30th cells from left to right, which are out of order in this density gradient. The 15th cell contains the densest strands with 1 HEG linker, and the 30th cell contains the densest strands with 2 HEG linkers. "1st" (A, C, G, T above the array image) refers to the sequence of the first nucleotide in the first subsite on the second strand, "2nd" (A, C, G, T to the far left of the array image) refers to the sequence of the second nucleotide, and so on. The symbol "-" for the fourth nucleotide indicates that no fourth nucleotide was synthesized, so the subsites for those strands were three, not four, nucleotides in length. The *Rsa*I recognition site is 5′-GTAC-3′ (indicated in red).

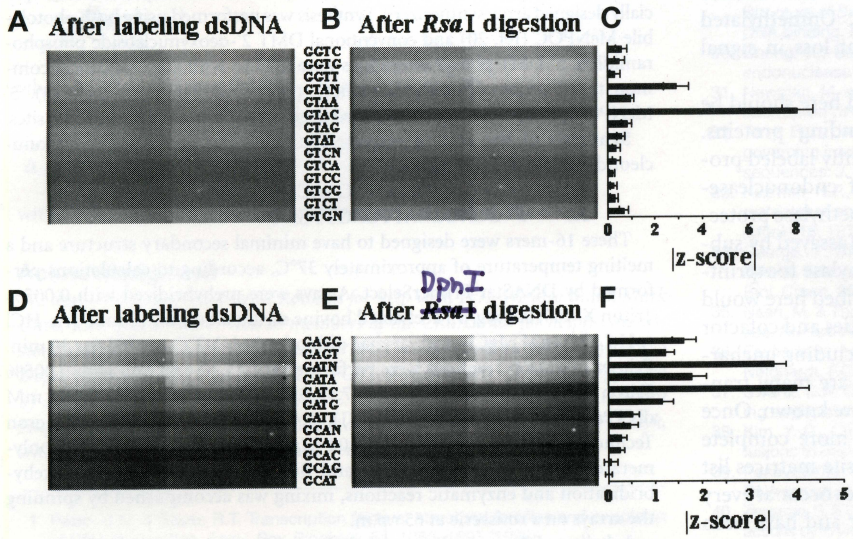

Figure 4. *Rsa*I and *Dpn*I digestions of labeled dsDNA arrays. (A–F) Inverted repeat arrays with no proximal flanking sequence, 20 bp distal flanking sequence, and 0–14 bp spacers. The sequences of the first subsites are listed between (A) and (B), and between (D) and (E). (A) Before and (B) after *Rsa*I digestion of a dsDNA array labeled by terminal transferase addition of fluorescein-12-ddCTP after primer extension. (C and F) Average *z* scores for the sequences shown. The length of the error bars is one standard deviation of the *z* score for each subsite. (D) Before and (E) after *Dpn*I digestion of a methylated dsDNA array labeled by incorporation of fluorescein-12-dCTP.

tein interactions in different buffer conditions allows the determination of the effects of factors such as ionic strength and pH on the sequence specificity of the interactions. For example, digests under different buffer conditions will induce "star" activity to varying degrees for *Eco*RI[22,32], *Eco*RV[24,33], *Bam*HI[34], and *Pvu*II[35]. Such digests performed on arrays will provide data that will aid in modeling interactions with many different DNA sequence variants under conditions of differing pH and ionic strength.

***dam* methylation.** A number of DNA–protein interactions require that the DNA be biochemically modified in some way. For example, a common DNA modification in bacteria is N^6-methyladenine at 5′-GATC-3′ sites. In *Escherichia coli* it is accomplished by *dam* methylase and used not only to distinguish self from foreign DNA, but also to identify the parent from the daughter strand at replication forks[36]. Methylated arrays could be used to conduct further study of proteins involved in the restriction system. They also have the potential to be used for studying proteins involved in discriminating between parent and daughter strands.

To determine whether dsDNA arrays could be used to investigate methylation-sensitive interactions, we digested deoxycytidine triphosphate (dCTP)-labeled dsDNA arrays with *Dpn*I, which cleaves only when the adenine within the restriction site 5′-GATC-3′ is methylated. After labeling, the array was methylated with *dam* methylase. The portion of the methylated array containing 5′-

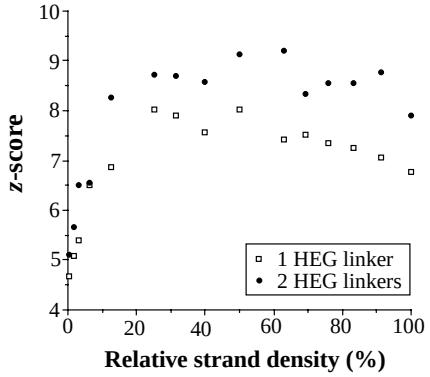

Figure 5. *z* scores (normalized signal intensity differences) at 5′-GTAC-3′ sites before and after *Rsa*I digestion of a dsDNA array designed to vary the density and accessibility of the DNA strands. Densities (not measured) are estimates based on surface site densities, chemical coupling yields, deprotection cross-sections, and half-lives. 100% corresponds to approximately 1 pmol/cm^2.

GATC-3′, before and after *Dpn*I digestion, is shown in Figures 4D and E. There was significant loss in signal intensity (average $z = 4$, $p = 3 \times 10^{-5}$) at the 30 features on the array where the first subsite is 5′-GATC-3′. As there was no proximal flanking sequence on this array, any first subsites that started with the dinucleotide 5′-TC-3′ also contained an intact recognition site for *Dpn*I as a result of the junction of this dinucleotide with the 16-mer used for priming (5′-GGACCGATTGACTTGA-3′). Excluding these subsites, the 4-mer subsite with the second highest *z* score on the methylated array after *Dpn*I digestion was 5′-GGAT-3′ ($z = 3$, $p = 1 \times 10^{-3}$). As the spacer (0–14 nucleotides) consisted of degenerate sequence, about one-fourth of the nucleotides incorporated after the 5′-GGAT-3′ subsite were C, completing the restriction site for *Dpn*I. Unmethylated arrays digested with *Dpn*I did not show significant loss in signal intensity ($z = 0.5$, $p = 0.3$) at the 5′-GATC-3′ subsite.

The DNA–protein interaction assay we described here should be generally useful for characterization of DNA-binding proteins. Potential variations include (1) binding of fluorescently labeled protein[37] to unlabeled dsDNA arrays, (2) binding of endonuclease-labeled proteins[38,39] to labeled dsDNA arrays, or (3) methylase protection of labeled dsDNA arrays by unlabeled proteins (assayed by subsequent restriction digests analogous to in vivo methylase footprinting[40]). The type of parallel-binding experiment described here would permit the characterization of the sequence specificities and cofactor requirements of potential DNA-binding proteins, including uncharacterized open reading frames. Furthermore, there are many transcription factors for which only a few operator sites are known. Once a greater number of binding sites are discovered, more complete recognition site matrices can be constructed. These site matrices list the frequencies with which each of the four nucleotides occur at every position in the binding site of a transcription factor and have been used to predict new binding sites in genomes[41]. After defining matrices from the sequences bound on dsDNA arrays, genomic locations of sites that match the matrix could be identified. Any genes downstream of these sites might be regulated by the putative transcription factor.

The method we present here is general with respect to the DNA component, and it could be used to study DNA-binding proteins of most structural classes from any organism. Most of the sequence specificity of binding is dependent upon two factors: a set of 3–5 bp subsites[42–44] and the length of the spacer separating the subsites[45]. For example, to study a protein that binds DNA via three zinc-fingers, arrays with three sets of 3 bp[1] subsites could be used. Because of the highly parallel nature of the arrays, it should be possible to derive information about the base pairs that are important for specific interactions with particular residues. Proteins with mutations in the residues critical for specific recognition could be characterized on the arrays, and these experiments would be useful for studies designed to better understand the determinants of sequence-specific recognition.

Experimental protocol

Chip design. The priming sequence, which was constant over the entire array, was synthesized immediately after the HEG linker (Fig. 1). All possible recognition sites composed of four nucleotides were provided by the set of subsites proximal to the array surface. The distal subsites were inverted or tandem repeats of the proximal subsites. The subsites were constant within a feature, but varied from feature to feature. (A feature is a portion of the array in which the sequence, excluding any degenerate nucleotides, is identical from strand to strand; for example, an array that contains all sequences of length six nucleotides has 4^6 features). Distal flanking sequence (5–20 nucleotides) was included after the second subsite because many proteins that interact with DNA are not capable of binding a short segment of DNA that consists of only the recognition site[46]. Up to 10 nucleotides of flanking sequence were included between the priming sequence and the first subsite in order to lengthen the portion of the DNA on the proximal side of the recognition site. The DNA spacer between the two subsites was either fixed or of variable length (0–14 nucleotides). Degenerate nucleotides (mix of A, C, G, and T) were added in the segments where the sequences are not expected to make a strong contribution to sequence-specific binding. These included the proximal and distal flanking sequences, as well as the spacer. Unlike the priming sequence and the subsites, the segments containing degenerate sequence varied from strand to strand within a feature. Certain arrays were synthesized in ways designed to vary the density and accessibility of the DNA strands. A first-strand density gradient was created by varying the time of photodeprotection after addition of the first photoprotected HEG synthesis linker. The photodeprotection was followed by coupling of *N*,*N*-diethyl *N*,*N*-diisopropyl phosphoramidite to "cap" the exposed sites before synthesis of the oligonucleotides. The distance separating the DNA strands from the glass surface of the array was varied by using either one or two HEG synthesis linkers.

DNA array synthesis. DNA arrays were synthesized at Affymetrix on a specially designed array synthesizer[21]. Synthesis was performed using both photolabile MeNPOC (ref. 20) and conventional DMT 2'-deoxynucleoside phosphoramidites. 5′ DMT phosphoramidites were used to synthesize sequences common to all probes (priming sequence, spacer, and flanking sequences). 5′ MeNPOC phosphoramidites were used to generate the unique recognition sites.

Synthesis of the second strand. The following two 16-mer oligonucleotides were used for priming:

1s: 5′-TCGACACGCTCCAACA-3′
2s: 5′-GGACCGATTGACTTGA-3′

These 16-mers were designed to have minimal secondary structure and a melting temperature of approximately 37°C, according to calculations performed by DNAStar PrimerSelect. Arrays were prehybridized with 0.005% Triton X-100, 0.2 mg/ml acetylated bovine serum albumin, 10 mM Tris-HCl (pH 7.5), 5 mM $MgCl_2$, and 7.5 mM dithiothreitol (DTT) at 37°C for 30 min. Primer extension reactions were performed at 37°C for 60 min with 0.005% Triton X-100, 10 mM Tris-HCl (pH 7.5), 5 mM $MgCl_2$, 7.5 mM DTT, 0.4 mM dNTPs (Pharmacia, Piscataway, NJ), 0.4 μM 1s or 2s primer (Operon Technologies, Alameda, CA), and 0.04 U/μl Klenow fragment of DNA polymerase I (3′ to 5′ *exo*$^-$; New England BioLabs, Beverly, MA). For prehybridization and enzymatic reactions, mixing was accomplished by spinning the arrays on a rotisserie at 65 r.p.m.

Labeling of the second strand. Labeling of the second strand by incorporation of fluorescently labeled dNTP was accomplished by addition of 0.4 μM of fluorescein-labeled dNTP (New England Nuclear, Boston, MA) to the primer extension reaction. A negative control consisting of a primer extension reaction lacking Klenow fragment of DNA polymerase I resulted in no gain in signal intensity (data not shown). This demonstrates that incorporation of fluorescent label and signal generation is polymerase dependent. End-labeling after synthesis of the second strand was carried out in 0.005% Triton X-100, 10 mM Tris-acetate (pH 7.5), 10 mM magnesium acetate, 50 mM potassium acetate, 0.005 mM fluorescein-12-ddNTP, 0.15 U/μl recombinant terminal deoxynucleotidyl transferase (Gibco-BRL, Gaithersburg, MD) at 37°C for 90 min. A negative control consisting of end-labeling an ssDNA array showed minimal signal incorporation (data not shown). This demonstrates that only duplex DNA was end-labeled with terminal deoxynucleotidyl transferase. After background subtraction, the signals in the different features on arrays

labeled by either of these protocols were within a factor of three of each other. After following one of the two labeling protocols, arrays were washed in 6× SSPE/T (0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA, and 0.005% Triton X-100). After washing, the hybridization chamber was filled with fresh 6× SSPE/T before scanning. The fluorescein was then excited using an argon ion laser, and the resulting emission was detected with a photomultiplier tube through a 530 nm bandpass filter (Molecular Dynamics, Sunnyvale, CA). Images were obtained with a scan resolution of 11.25 μm using a specially designed confocal scanner[47]. In addition to incorporation of labeled dNTP or end labeling with terminal transferase to assess second-strand synthesis, we also extended with unlabeled dNTPs, then stained with the double-strand–specific dyes ethidium bromide and PicoGreen (data not shown). One disadvantage of this type of labeling is that arrays stained in this manner cannot be destained. This renders them undesirable for studying DNA–protein interactions as both of these dyes alter DNA structure—ethidium bromide by intercalating between base pairs and PicoGreen by adhering to the major grooves of the DNA by an unknown mechanism[48]. DNA distorted in this way would be expected to have altered contacts with interacting proteins.

Methylation. DsDNA arrays were methylated at 5′-GATC-3′ sites by incubation at 37°C for 90 min with 0.2 U/μl *dam* methylase (New England BioLabs) in 0.005% Triton X-100, 50 mM Tris-HCl (pH 7.5), 10 mM EDTA, 5 mM 2-mercaptoethanol, and 80 μM *S*-adenosylmethionine.

Restriction digestion. *Rsa*I digestion was performed at 37°C for 30 min with 0.01 U/μl *Rsa*I (New England BioLabs) in 0.005% Triton X-100, 10 mM Bis-Tris-propane-HCl, 10 mM $MgCl_2$, and 1 mM DTT (pH 7.0 at 25°C). *Dpn*I digestion was performed at 37°C for 60 min with 0.5 U/μl *Dpn*I (New England BioLabs) in 0.005% Triton X-100, 20 mM Tris-acetate, 10 mM magnesium acetate, 50 mM potassium acetate, and 1 mM DTT (pH 7.9 at 25°C).

Data analysis. The mean signal intensities and the corresponding standard deviations and sample sizes were measured using GeneChip software version 2.0 (Affymetrix). The *z* scores were evaluated using customized Perl scripts. The software package *Mathematica* (Wolfram Research, Champaign, IL) was used to calculate *p* values for the *z* scores. The *z* score for each of the features on the array was calculated by normalizing the difference between the individual and the mean signal intensity losses by the standard deviation of the loss:

$$z_k = \frac{(\Delta_k - \Delta_{mean})}{SD_\Delta}$$

where $\Delta_k = (\text{signal intensity})_k^{before} - (\text{signal intensity})_k^{after}$,

$$\Delta_{mean} = \frac{\sum_{k=1}^{n} \Delta_k}{n}, \text{ and } SD_\Delta = \sqrt{\frac{\sum_{k=1}^{n} (\Delta_k - \Delta_{mean})^2}{n}},$$

with $k = 1, \ldots, n$ corresponding to each of the features on the array.

Acknowledgment

We thank John Aach and Keith Robison for help with Perl. We also thank Mark Chee, Rich Baldarelli, as well as members of the Church lab for helpful discussions and critical reading of the manuscript. This work was supported by the US Department of Energy (grant no. DE-FG02-87-ER60565). M.B. was supported by an NSF Graduate Fellowship. G.M.C. was partially supported by the Howard Hughes Medical Institute. This article is dedicated to my father, Roman P. Bulyk, who passed away while this work was being completed.

1. Pabo, C.O. & Sauer, R.T. Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61,** 1053–1095 (1992).
2. Craig, N.L. The mechanism of conservative site-specific recombination. *Annu. Rev. Genet.* **22,** 77–105 (1988).
3. Pingoud, A. & Jeltsch, A. Recognition and cleavage of DNA by type-II restriction endonucleases. *Eur. J. Biochem.* **246,** 1–22 (1997).
4. Margulies, C. & Kaguni, J.M. Ordered and sequential binding of DNA protein to oriC, the chromosomal origin of *Escherichia coli*. *J. Biol. Chem.* **271,** 17035–17040 (1996).
5. Woodbury, C.P. & Hippel, P.H.V. On the determination of deoxyribonucleic acid–protein interaction parameters using the nitrocellulose filter-binding assay. *Biochemistry* **22,** 4730–4737 (1983).
6. Jansen, C., Gronenborn, A.M. & Clore, G.M. The binding of the cyclic AMP receptor protein to synthetic DNA sites containing permutations in the consensus sequence TGTGA. *Biochem. J.* **246,** 227–232 (1987).
7. Bowen, B., Steinberg, J., Laemmli, U.K. & Weintraub, H. The detection of DNA-binding proteins by protein blotting. *Nucleic Acids Res.* **8,** 1–20 (1980).
8. Miskimins, W.K., Roberts, M.P., McClelland, A. & Ruddle, F.H. Use of a protein-blotting procedure and a specific DNA probe to identify nuclear proteins that recognize the promoter region of the transferrin receptor gene. *Proc. Natl. Acad. Sci. USA* **82,** 6741–6744 (1985).
9. Hanes, S.D. & Brent, R. A genetic model for interaction of the homeodomain recognition helix with DNA. *Science* **251,** 426–430 (1991).
10. Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14,** 1675–1680 (1996).
11. Roth, F.P., Hughes, J.D., Estep, P.W. & Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16,** 939–945 (1998).
12. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
13. Chee, M. et al. Accessing genetic information with high-density DNA arrays. *Science* **274,** 610–614 (1996).
14. Gunderson, K.L. et al. Mutation detection by ligation to complete N-mer DNA arrays. *Genome Res.* **8,** 1142–1153 (1998).
15. Hacia, J.G., Brody, L.C., Chee, M.S., Fodor, S.P.A. & Collins, F.S. Detection of heterozygous mutations in *BRCA1* using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat.Genet.* **14,** 441–447 (1996).
16. Wang, D.G. et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–1082 (1998).
17. Shoemaker, D.D., Lashkari, D.A., Morris, D., Mittmann, M. & Davis, R.W. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* **14,** 450–456 (1996).
18. Cho, R.J. et al. Parallel analysis of genetic selections using whole genome oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **95,** 3752–3757 (1998).
19. Lockhart, D.J., Vetter, D. & Diggelmann, M. Surface-bound, unimolecular, double-stranded DNA. (Affymetrix, Inc., USA). US patent # 5556752, issue date 9/17/96.
20. Pease, A.C. et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* **91,** 5022–5026 (1994).
21. McGall, G.H. et al. The efficiency of light-directed synthesis of DNA arrays on glass substrates. *J. Amer. Chem. Soc.* **119,** 5081–5090 (1997).
22. Polisky, B. et al. Specificity of substrate recognition by the *Eco*RI restriction endonuclease. *Proc. Natl. Acad. Sci. USA* **72,** 3310–3314 (1975).
23. Thielking, V., Alves, J., Fliess, A., Maass, G. & Pingoud, A. Accuracy of the *Eco*RI restriction endonuclease: binding and cleavage studies with oligodeoxynucleotide substrates containing degenerate recognition sequences. *Biochemistry* **29,** 4682–4691 (1990).
24. Engler, L.E., Welch, K.K. & Jen-Jacobson, L. Specific binding by *Eco*RV endonuclease to its DNA recognition site GATATC. *J. Mol. Biol.* **269,** 82–101 (1997).
25. Lesser, D.R., Kurpiewski, M.R. & Jen-Jacobson, L. The energetic basis of specificity in the *Eco*RI endonuclease–DNA interaction. *Science* **250,** 776–786 (1990).
26. McClarin, J.A. et al. Structure of the DNA–*Eco*RI endonuclease recognition complex at 3 A resolution. *Science* **234,** 1526–1541 (1986).
27. Wah, D.A., Hirsch, J.A., Dorner, L.F., Schildkraut, I. & Aggarwal, A.K. Structure of the multimodular endonuclease *Fok*I bound to DNA. *Nature* **388,** 97–100 (1997).
28. Winkler, F.K. et al. The crystal structure of *Eco*RV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J.* **12,** 1781–1795 (1993).
29. Newman, M., Strzelecka, T., Dorner, L.F., Schildkraut, I. & Aggarwal, A.K. Structure of *Bam*HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. *Science* **269,** 656–663 (1995).
30. Cheng, X., Balendiran, K., Schildkraut, I. & Anderson, J.E. Structure of *Pvu*II endonuclease with cognate DNA. *EMBO J.* **13,** 3927–3935 (1994).
31. Newman, M. *et al.* Crystal structure of restriction endonuclease *Bgl*I bound to its interrupted DNA recognition sequence. *EMBO J.* **17,** 5466–5476 (1998).
32. Terry, B.J., Jack, W.E., Rubin, R.A. & Modrich, P. Thermodynamic parameters governing interaction of *Eco*RI endonuclease with specific and nonspecific DNA sequences. *J. Biol. Chem.* **258,** 9820–9825 (1983).
33. Kuz'min, N.P., Loseva, S.P., Beliaeva, R.K., Kravets, A.N. & Solonin, A.S. EcoRV restrictase: physical and catalytic properties of homogenous enzyme. *Mol. Biol. (Mosk)* **18,** 197–204 (1984).
34. George, J., Blakesley, R.W. & Chirikjian, J.G. Sequence-specific endonuclease *Bam*HI. Effect of hydrophobic reagents on sequence recognition and catalysis. *J. Biol. Chem.* **255,** 6521–6524 (1980).
35. Nasri, M. & Thomas, D. Alteration of the specificity of *Pvu*II restriction endonuclease. *Nucleic Acids Res.* **15,** 7677–7687 (1987).
36. *Escherichia coli* and *salmonella.* Cellular and molecular biology Vol. 1. (ed. Neidhardt, F.C.) (ASM, Washington, DC; 1996).
37. Swartz, D.R. Covalent labeling of proteins with fluorescent compounds for imaging applications. *Scanning Microsc. Suppl.* **10,** 273–284 (1996).
38. Kim, Y.-G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to *Fok*I cleavage domain. *Proc. Natl. Acad. Sci. USA* **93,** 1156–1160 (1996).
39. Panayotatos, N. & Backman, S. A site-directed recombinant nuclease probe of DNA structure. *J. Biol. Chem.* **264,** 15070–15073 (1989).
40. Tavazoie, S. & Church, G.M. Quantitative whole-genome analysis of DNA–protein interactions by in vivo methylase protection in *E. coli*. *Nat. Biotechnol.* **16,** 566–571 (1998).
41. Robison, K., McGuire, A.M. & Church, G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* **284,** 241–254 (1998).
42. Desjarlais, J.R. & Berg, J.M. Length-encoded multiplex binding site determination: application to zinc finger proteins. *Proc. Natl. Acad. Sci. USA* **91,** 11099–11103 (1994).
43. Harrison, S.C. & Aggarwal, A.K. DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.* **59,** 933–969 (1990).
44. Suzuki, M., Yagi, N. & Gerstein, M. DNA recognition and superstructure formation by helix-turn-helix proteins. *Protein Eng.* **8,** 329–338 (1995).
45. Harrison, S.C. A structural taxonomy of DNA-binding domains. *Nature* **353,** 715–719 (1991).
46. Liu-Johnson, H.-N., Garterberg, M.R. & Crothers, D.M. The DNA binding domain and bending angle of *E. coli* CAP protein. *Cell* **47,** 995–1005 (1986).
47. Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H. & Lockhart, D.J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15,** 1359–1366 (1997).
48. *Molecular probes: handbook of fluorescent probes and research chemicals.* (ed.