Defining the sequence-recognition profile of DNA-binding molecules

Christopher L. Warren[†‡], Natasha C. S. Kratochvil[†], Karl E. Hauschild[†], Shane Foister[§], Mary L. Brezinski[†], Peter B. Dervan[§], George N. Phillips, Jr.[†‡], and Aseem Z. Ansari[†‡¶]

[†]Department of Biochemistry and [‡]Genome Center, University of Wisconsin, Madison, WI 53706; and [§]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125

Contributed by Peter B. Dervan, November 11, 2005

Determining the sequence-recognition properties of DNA-binding proteins and small molecules remains a major challenge. To address this need, we have developed a high-throughput approach that provides a comprehensive profile of the binding properties of DNA-binding molecules. The approach is based on displaying every permutation of a duplex DNA sequence (up to 10 positional variants) on a microfabricated array. The entire sequence space is interrogated simultaneously, and the affinity of a DNA-binding molecule for every sequence is obtained in a rapid, unbiased, and unsupervised manner. Using this platform, we have determined the full molecular recognition profile of an engineered small molecule and a eukaryotic transcription factor. The approach also yielded unique insights into the altered sequence-recognition landscapes as a result of cooperative assembly of DNA-binding molecules in a ternary complex. Solution studies strongly corroborated the sequence preferences identified by the array analysis.

chemical genomics | ligand–DNA recognition

A central goal of synthetic biology, chemical biology, and molecular medicine is the design and creation of synthetic molecules that can target specific DNA sites in the genome (1, 2). Such molecules can be harnessed to regulate biological processes such as transcription, recombination, and DNA repair (1–4). The greatest success in designing molecules with programmable DNA-binding specificity has been with polyamides (2). However, a major hurdle in the design of new classes of sequence-specific DNA-binding molecules is the inability to comprehensively define the full range of their DNA sequence-recognition properties, and therefore, the inability to predict all their potential target sites in the genome.

Given the importance of understanding the basis of molecular recognition between DNA and its ligands, several methods have been developed to determine the sequence specificity of DNA-binding molecules (small molecules as well as proteins). The most frequently used approach is the systematic evolution of ligands by exponential enrichment (SELEX), which utilizes selection and enrichment of the DNA sequences that bind with the highest affinity to a molecule of interest (4). This assay, although highly informative, identifies only the best binding sequences, whereas the less optimal, and often biologically relevant, sequences are missed. Other commonly used biochemical or biophysical approaches are labor-intensive and can be used only to study a limited set of sequence variants (5–10). Medium-throughput microarrays have also been developed in which duplex DNA molecules are immobilized on surfaces and protein binding is detected by surface plasmon resonance (11) or fluorescence (12, 13). Despite such demonstrations of feasibility, technical challenges have hindered the general application of these array platforms. A solution-phase medium-throughput assay utilizes DNA sequence variants presented in distinct wells and protein or small molecule binding detected by displacement of a DNA-intercalating fluorescent dye (14). Each of these medium-throughput approaches, however, is limited to querying DNA sequences with only three, four, or five permuted positions.

In a recent approach, a biased microarray bearing only the intergenic regions of yeast chromosome was used to map transcription factor binding sites *in vitro* (15). These arrays provide a biased binding profile and are limited to organisms with small and well annotated genomes. Another technique that circumvents this problem relies on sonicating genomic DNA into small fragments and adding a transcription factor to isolate putative binding sites (16). However, this method, like SELEX, is likely to overrepresent strong binding sites, thereby providing biased sequence-recognition profiles. These methods are not amenable to an unbiased analysis of the binding properties of small molecule DNA ligands.

Chromatin immunoprecipitated (ChIP) DNA analyzed on oligonucleotide microarrays (chip) has also been used to map binding sites for DNA-binding transcription factors (17–19). Importantly, ChIP-chip studies have suggested that *in vitro* affinity of cooperatively binding transcription factors for specific DNA sequences is often recapitulated in the relative occupancy of these sequences *in vivo* (20, 21). This observation suggests that for a given transcription factor (or a set of cooperatively binding factors), the knowledge of its full sequence-recognition profile, measured *in vitro*, can be highly instructive in computationally identifying binding sites in the genome. Thus far, in the absence of genome-wide binding and expression data, computational approaches to identifying regulatory sites have been limited to phylogenetic comparisons of conserved noncoding sequences (22). However, unlike proteins, for most DNA-binding small molecules with unknown DNA-binding properties, ChIP-chip analysis is nontrivial, and phylogenetic comparisons are irrelevant.

To bridge this gap between computational methods and molecular recognition properties of DNA ligands, we have developed a comprehensive high-throughput platform that can rapidly and reliably identify the cognate sites of DNA-binding molecules. This platform provides an unbiased analysis because it consists of a double-stranded DNA array that displays the entire sequence space represented by 8 bp (all possible permutations equal 32,896 molecules) and can currently be extended to as many as 10 variable base pair positions. We have also developed a systematic approach for treating the array data that can be applied to arrays of greater complexity. Because most metazoan DNA-binding proteins target 6–10 bp (23), and because DNA-binding small molecules rarely exceed 8 bp (24), our cognate site identifier (CSI) arrays should be capable of identifying and ranking sequences preferred by almost any DNA-binding ligand by itself, or, in many cases, in cooperatively binding pairs. Our approach derives comprehensive binding profiles from a rapid, unbiased, and unsupervised examination of the entire

Conflict of interest statement: No conflicts declared.

Abbreviations: ChIP-chip, analysis of chromatin-immunoprecipitated DNA on oligonucleotide microarrays; CSI, cognate site identifier; PA1, polyamide 1; PA2, polyamide 2; PA3, polyamide 3; Exd, extradenticle; Hox, homeobox transcription factors; Dp, dimethylaminopripylamide; Py, *N*-methylpyrrole; Py*, Cy3-Py; Im, *N*-methylimidazole.

¶To whom correspondence should be addressed at: Department of Biochemistry and Genome Center of Wisconsin, University of Wisconsin, 433 Babcock Drive, Madison, WI 53706. E-mail: ansari@biochem.wisc.edu.

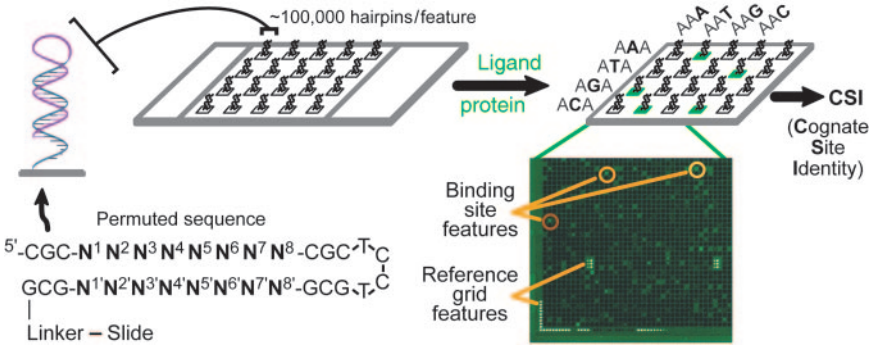


Fig. 1. Illustration of a CSI microarray and the experimental approach. Each hairpin probe is composed of a permuted hairpin stem (N^1–N^8) with a 3-bp flanking sequence (CGC) on either side. N′ represents the exact complement to the permuted (N) forward sequence. A fluorescently tagged ligand is applied to the microarray to obtain a comprehensive ligand-binding profile. In addition to reference grid features, high intensity features are circled, indicating tight binding of the ligand to that specific probe sequence.

DNA sequence space. These analyses can be extended to DNA-binding proteins from any organism or, in the case of small molecules, used to predict binding sites in any genome.

Results

Array Design. The duplex DNA sequences are designed as self-complementary palindromes interrupted at the center by a TCCT sequence to facilitate the formation of DNA hairpins (Fig. 1). The 34-residue oligonucleotide is synthesized directly on the glass surface by using a maskless array synthesizer (25) that can readily create up to 786,000 spatially resolved features. After inducing hairpin formation, we found that 95% of the oligonucleotides in the array form duplexes (see *Materials and Methods*). In our hairpin design, we added three constant base pairs on either side of the 8 bp that were permuted (N^1–N^8 in Fig. 1). Previous work shows that this addition is sufficient to buffer the core of the hairpin stem against thermal end-fraying of the duplex and against deviations from B-form DNA resulting from the presence of the loop (26). There is good evidence that the core of a hairpin stem interacts with proteins and small molecule ligands indistinguishably from DNA duplexes composed of two individual complementary strands (27, 28).

Array Validation Using an Engineered Small Molecule. To test the accuracy and fidelity of the CSI array, we used a polyamide engineered to target a specific DNA sequence (PA1, Fig. 2*A*). Polyamides are DNA-binding small molecules composed of *N*-methylpyrrole (Py) and *N*-methylimidazole (Im) heterocycle rings. The arrangement of the heterocycles (Im or Py) can be programmed to create polyamides that target most naturally occurring 6- to 8-bp DNA sequences (2). PA1, in particular, was designed to target the sequence 5′-WWGWWCWW-3′ (W = A or T) (Fig. 2) (29). A Cy3 fluorescent dye is conjugated to the *N*-methyl position of an internal pyrrole (Py*). Such conjugation does not meaningfully alter the DNA-binding properties of the polyamides (30). Previous solution-based footprinting (29) and dye displacement assays (28) have shown that polyamides discriminate very highly between their targeted cognate site and sites that differ by a single base pair. Thus, PA1, a well characterized DNA-binding molecule, serves as a stringent test for the ability of the CSI array to accurately identify its sequence recognition landscape.

PA1 was incubated with the array and a distinct pattern of fluorescent binding features was readily discernible, and the pattern did not change over a broad range of PA1 concentrations (0.5–500 nM). The array-to-array variability was very low, with an average correlation coefficient of 0.88 (Fig. 6, which is published as supporting information on the PNAS web site). A majority of the features showed low background fluorescence, and a small subset of the features were of high intensity (Fig. 2*B* and Fig. 7, which is published as supporting information on the PNAS web site). The duplicate features within an array and replicate features between arrays were averaged together to give finalized intensities. These averaged intensities were then converted into *Z* scores [Z = |signal − mean|/standard deviation] to reflect the signal-to-noise ratio (Fig. 2*B*).

Sequences in the highest *Z* score bin (≥25) were subjected to several motif-searching algorithms (31–33), which identified 5′-

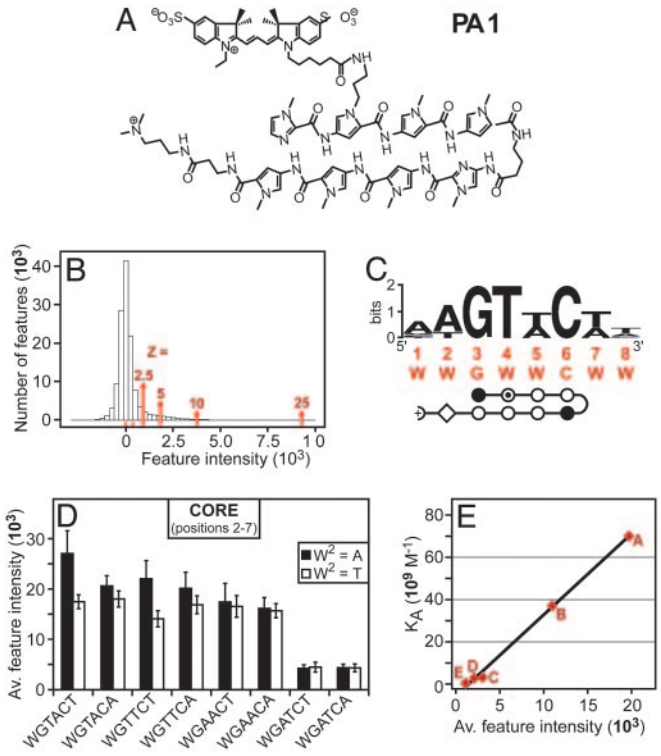


Fig. 2. CSI profile of PA1. (*A*) Structure of polyamide-Cy3 conjugate PA1 (ImPy*PyPy-γ-ImPyPyPy-β-Dp). (*B*) Histogram of averaged intensities of all replicate features. Intensities are background-subtracted so that the mean intensity is zero. Red numbers indicate *Z* scores. (*C*) (*Top*) Logo (53) based on the sequences from the top *Z* score bin ($Z > 25$). (*Middle*) DNA sequence that would be targeted by PA1 based on the ring pairing rules for polyamides; an Im/Py ring pair targets G·C, and a Py/Py pair targets either A·T or T·A (2). Numbers indicate base pair positions. (*Bottom*) A ball-and-stick schematic of PA1. Im or open circle, *N*-methylimidazole; Py or filled circle, *N*-methylpyrrole ring; Py* or open circle with inner dot, *N*-methylpyrrole ring with a Cy3 dye attached; β or diamond, β-alanine; Dp or a half circle with a positive charge, dimethylaminopripylamide; γ or turn, γ-aminobutyric acid. (*D*) Intensity profile of all sequence permutations of the core consensus sequence 5′-WGWWCW-3′. The intensities of all probes that contain a specific permutation of the core consensus sequence are averaged together. (*E*) Plot of the correlation between CSI intensities and equilibrium association constants (K_a) determined from nuclease protection (DNase I footprinting) experiments (Table 1). The intensities of all CSI probe sequences that contain a particular footprinted sequence are averaged together.

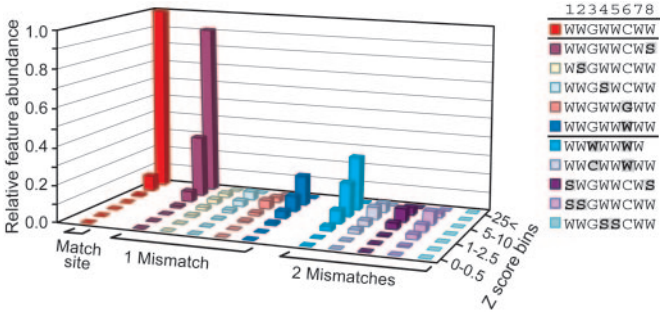


Fig. 3. Comprehensive mutational analysis plot of PA1. (*Left*) Plot of the relative abundance of each sequence motif in each *Z* score bin. Relative abundance is calculated as the number of sequences in each *Z* score bin that contain a particular sequence motif divided by the number of total sequences in that *Z* score bin. These abundances are then scaled to one. (*Right*) Sequences. S = G or C, W = A or T.

$W^1W^2G^3T^4W^5C^6W^7W^8$-3′, a motif that is nearly identical to the predicted binding site for the polyamide 5′-WWGWWCWN-3′ (Fig. 2*C*). Parsing of the core sequences (N^2–N^7) showed that not all permutations of the consensus are bound equally well. In particular, all sequences that contained the sequence 5′-WWGATCWW-3′ had significantly lower intensities than other permutations of the consensus sequence (Fig. 2*D*). This observation is consistent with previous solution studies (34). Furthermore, the flanking sequence (N^1, N^8) showed an equally strong preference for a W (A/T) in both positions (Fig. 8, which is published as supporting information on the PNAS web site). This observation is also in agreement with the preference of the polyamide γ-aminobutyric acid turn and Dp tail for A/T residues (35). Finally, we found that the cognate site preferences identified by the array were entirely consistent with reported solution binding studies of this polyamide for five different sequences (Fig. 2*E* and Table 1, which is published as supporting information on the PNAS web site). The high correlation (r^2 = 0.997) of feature intensity on the array with affinity for different cognate sites in solution provides significant confidence in the veracity of the cognate site preferences identified by the array. Taken together, these correlations demonstrate that the CSI array correctly identifies the cognate sites of a DNA-binding molecule, and that the CSI array accurately ranks each cognate site in the order of increasing affinity.

Comprehensive Mutational Analysis. In essence, the array performs a comprehensive ''mutational'' analysis as it queries the entire sequence space (within a defined size) to determine the contribution of every base pair in the cognate site for molecular recognition (Fig. 3). By examining the array data, it is apparent that substituting an S (G or C) at position 8 only subtly decreases binding by PA1. This finding is consistent with the ability of this symmetric polyamide to bind the sequence in only one orientation. Replacing one of the S residues at positions 6 (or 3′) with a W significantly attenuates, but does not abolish, binding. However, substituting any other position that prefers a W in the motif with an S residue nearly abolishes binding by PA1 (Fig. 3). The data also show that despite a double substitution at positions 3 and 6 to W, the resulting A/T stretch retains its ability to bind PA1. This observation is likely a result of the inherent affinity of polyamides for A/T-rich sequences (36).

Transcription Factor Binding and Cooperative Assembly. Having demonstrated the accuracy of the CSI arrays, we probed the sequence preferences of molecules that bind DNA cooperatively. We examined the cognate site preference of Exd, a transcription factor that plays an essential role in *Drosophila* development and is highly conserved across species, including humans (37). Exd binds DNA cooperatively with transcription factors from the Hox (homeobox) family (37, 38). Individual Hox proteins, as well as Exd, bind DNA with very low affinity and with poor specificity (38). Cooperative binding dramatically increases the affinity of Exd and Hox proteins for DNA and strongly influences DNA sequence specificity such that different Hox–Exd complexes target different genes (38). We have generated synthetic molecules (polyamide-peptide conjugates) that can mimic two key functions of the Hox family of transcription factors (39). First, they can bind sites targeted by specific Hox proteins, and second, they can cooperatively recruit Exd to an adjacent cognate site.

To determine the sequence specificity of Exd, we labeled it with Cy3 at a unique cysteine residue on an unstructured portion of the protein (Fig. 9*A*, which is published as supporting information on the PNAS web site.) (40). The modified protein does not differ in its ability to bind cooperatively with ultrabithorax (Ubx), a member of the Hox family, or with a synthetic Hox mimic to a known cognate DNA site. When tested on the CSI array, Exd alone, as expected, demonstrated little sequence-specific binding at concentrations ranging from 0.2 to 200 nM. It does, however, show an unexpected preference for stretches of consecutive G residues (Fig. 9*B*). Initial studies suggest that these sequences can form non-B-form, likely G-quadruplex (41), structures (Fig. 9*D*). The physiological importance of this binding interaction remains to be investigated.

When incubated with two different synthetic Hox mimics (PA2 and PA3), the Cy3-labeled Exd displayed an unambiguous pattern of feature binding in both sets of experiments (Fig. 4). PA2 and PA3 (ImImPy*Py-γ-ImPyPyPy-β-Dp) are designed to target the sequence 5′-WGWCCW-3′. Furthermore, instead of a Cy3 dye, PA2 and PA3 do not bear any dye but are conjugated to an Exd-binding peptide (N-FYPWMK-C). PA2 and PA3 differ solely by a single methylene in the linker connecting the Exd-binding peptide to the polyamide (Fig. 4*A*) (42). Because PA2 and PA3 are not fluorescently labeled, we detected cognate sites bound cooperatively by these synthetic Hox mimics and Exd, as well as sites bound by Exd alone.

The raw array data for the above experiments were treated as described for Fig. 2 (Fig. 10, which is published as supporting information on the PNAS web site). In addition to the G-stretches that Exd binds in the absence of any partner, three clear motifs emerged from the PA2–Exd data, whereas only two of those motifs were found in the PA3–Exd data (Fig. 4*B*). The Exd binding motif is 5′-NGAN-3′, which is consistent with the structural and genetic studies of Hox–Exd cognate sites (43). In other words, the 5′-GA-3′ dinucleotide is the only required sequence determinant for Exd binding to DNA. Remarkably, the array identified the differences in the arrangement of polyamide and Exd binding sites because of an ≈1.25-Å difference in the linker length between PA2 and PA3 (Fig. 4*C*). The other important result that emerged is that cooperative ternary assembly with Exd stabilizes binding of synthetic Hox mimics to truncated sites (5′-WGWC-3′). This stabilization is often seen in nature, where cooperative assembly of transcription factors utilizes suboptimal binding sites to ensure that only a higher order complex can efficiently bind to a regulatory element (44, 45).

Solution Binding and Molecular Modeling. To validate the unexpected differences in the motifs identified by each polyamide with Exd, we performed EMSAs. These studies with Exd and the two Hox mimics strongly support the cognate site preferences identified by the array (Fig. 5 *A* and *C*). Furthermore, molecular modeling (46, 47) analyses of PA2 and PA3 with Exd (with a docked Hox hexapeptide) agree well with the CSI array data. Both demonstrate that the linkers for PA2 and PA3 (9.98 and 11.25 Å, respectively) are able to deliver the hexapeptide to Exd at the composite consensus site (Fig. 5*B*). The array data indicate that both PA2 and PA3 reach Exd at the gapped composite site (consensus +1); however, simple geometric measurements with some energy min-

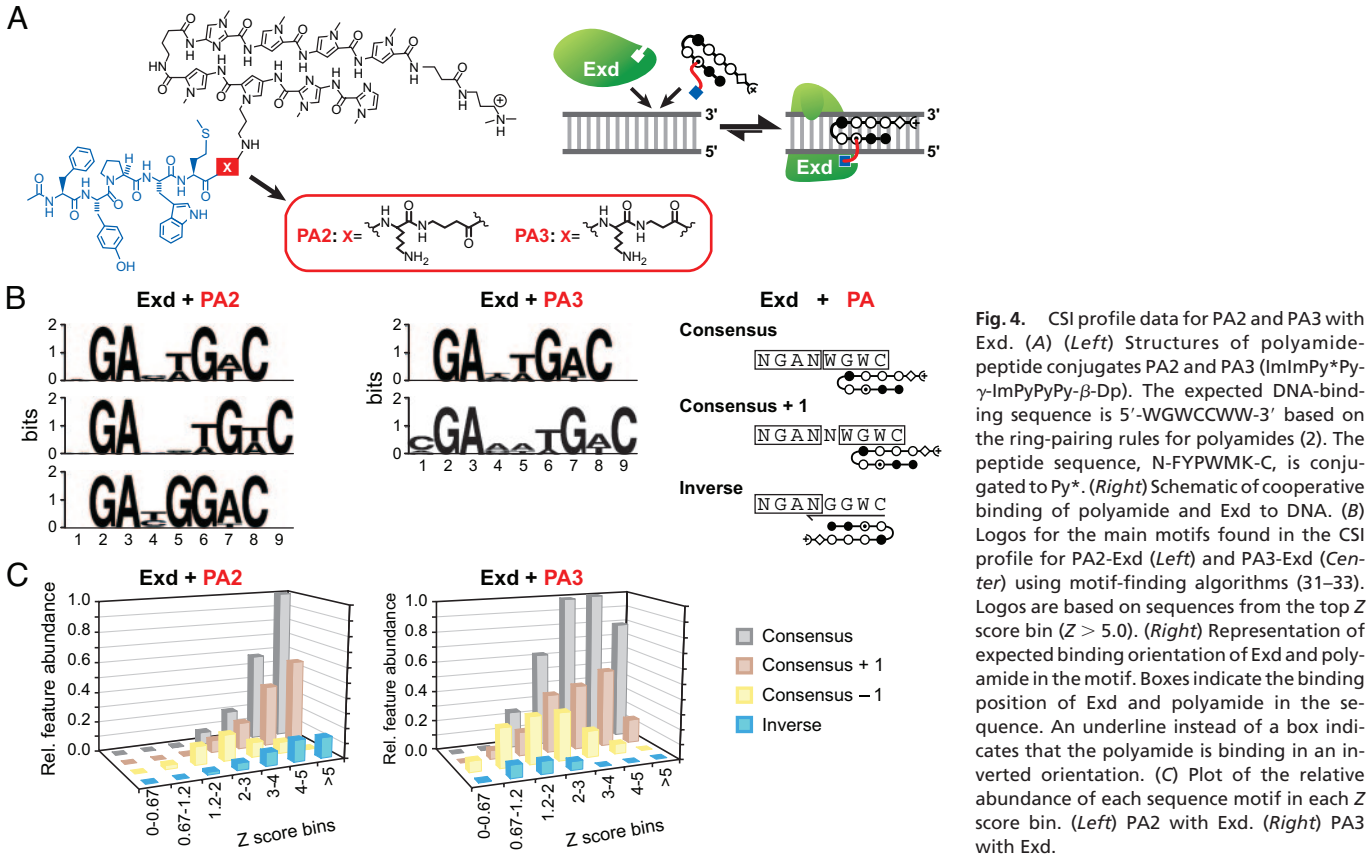


Fig. 4. CSI profile data for PA2 and PA3 with Exd. (*A*) (*Left*) Structures of polyamide-peptide conjugates PA2 and PA3 (ImImPy*Py-γ-ImPyPyPy-β-Dp). The expected DNA-binding sequence is 5′-WGWCCWW-3′ based on the ring-pairing rules for polyamides (2). The peptide sequence, N-FYPWMK-C, is conjugated to Py*. (*Right*) Schematic of cooperative binding of polyamide and Exd to DNA. (*B*) Logos for the main motifs found in the CSI profile for PA2-Exd (*Left*) and PA3-Exd (*Center*) using motif-finding algorithms (31–33). Logos are based on sequences from the top *Z* score bin ($Z > 5.0$). (*Right*) Representation of expected binding orientation of Exd and polyamide in the motif. Boxes indicate the binding position of Exd and polyamide in the sequence. An underline instead of a box indicates that the polyamide is binding in an inverted orientation. (*C*) Plot of the relative abundance of each sequence motif in each *Z* score bin. (*Left*) PA2 with Exd. (*Right*) PA3 with Exd.

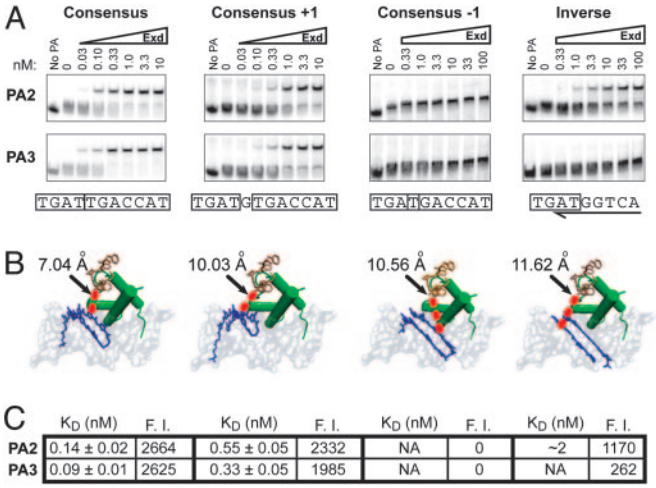
imization (48) suggest that the linker of PA3 should not span the distance. In the case of inverted binding sites, it is clear from modeling that the linker of PA3 is incapable of reaching Exd, and that the linker of PA2, even when fully extended, would be suboptimal, yielding an unstable ternary complex with Exd. These predictions are in good agreement with the observed CSI array binding data and EMSA results (Fig. 5*C*). However, the array data also demonstrate that a single base overlap (consensus −1) in the binding sites is not able to support binding of the complex, despite the fact that modeling indicates that the distance is similar to that of the consensus +1 site (Fig. 5*B*). The binding of either partner to overlapping sites may deform the DNA and prevent complex formation, even though modeling studies suggest that polyamide or Exd binding to the consensus −1 site should not disfavor complex formation (40, 49). Therefore, the ambiguities in molecular docking and energy minimization methods prevent precise prediction of the geometry of DNA grooves and distances between the interacting partners. In other words, the dramatic consequences on cognate site preference because of subtle, seemingly trivial, alterations in the linker length would not be readily apparent without the CSI array analysis. Therefore, this approach provides unexpected insight into molecular recognition properties of DNA-binding molecules when they bind individually or in cooperative pairs.

Discussion

In the CSI approach, the comprehensive sequence recognition landscape of DNA-binding molecules is determined in a rapid, unbiased, and unsupervised manner. Although we have directly labeled the DNA-binding molecules with a fluorescent dye, we anticipate that fluorescently labeled antibodies would serve equally well in detecting target proteins bound to their cognate sites. Because of the display of entire sequence space (within a certain size) on the array, there is no limitation on the use of proteins of a specific organism (as would be the case with ChIP-chip) or a specific class of small molecule ligands.

Because the CSI array analysis examines the entire sequence space at once, it also performs a comprehensive mutational analysis in a single experiment. Thus, one obtains information on the contribution of each nucleotide residue to the molecular recognition event between the DNA ligand and its cognate site(s). Moreover, DNA-binding preferences of the ligands (proteins or small molecules) are queried under identical conditions, yielding high-quality information. In the future, the accumulation of binding data from CSI analysis of different molecules will lead to the elucidation of the molecular recognition by a cluster of residues displayed on the surface of DNA-binding molecules. Such integrative perspective may be necessary to decipher the principles of molecular recognition displayed by DNA-binding molecules.

By determining the complete sequence recognition profile of DNA-binding molecules, the CSI array analysis also bridges the gap between the ChIP-chip approach and bioinformatic approaches of identifying regulatory DNA elements in genomes. For example, from a single CSI array experiment, one can validate (and order by affinity) all binding sites identified by ChIP-chip assays. Furthermore, the rank-order of the sequences can be used to computationally mine the genome for possible binding sites. The CSI array analysis would enable a coherent analysis of transcriptome studies by scanning for the presence of a range of possible binding sites in coregulated genes. Thus, including a CSI analysis in conjunction with other approaches will greatly aid in reducing the discrepancies between and absence of discernible binding sites in coregulated genes or the inability to detect protein binding at all biologically relevant sites *in vivo* by ChIP-chip analysis. The CSI array also provides a much-needed high-throughput approach for the design and development of novel classes of sequence-specific DNA-binding molecules.



C

	K_D (nM)	F. I.	K_D (nM)	F. I.	K_D (nM)	F. I.	K_D (nM)	F. I.
PA2	0.14 ± 0.02	2664	0.55 ± 0.05	2332	NA	0	~2	1170
PA3	0.09 ± 0.01	2625	0.33 ± 0.05	1985	NA	0	NA	262

Fig. 5. Solution binding and molecular modeling data. (*A*) EMSA. (*Upper*) PA2 (50 nM) incubated with increasing concentrations of Exd (in nM). (*Lower*) PA3 (50 nM) with an Exd titration. Labels above each pair of EMSAs indicate the binding motif used. The sequences used are shown below each pair of EMSAs. Boxes indicate the Exd- and polyamide-binding sites. An underline instead of a box indicates that the polyamide is binding in an inverted orientation. (*B*) Molecular models (46, 47) of Exd and polyamide bound in consensus, consensus +1, consensus −1, and inverse orientations. Models are based on aligning the DNA from the Protein Data Bank files 1B8I and 1M18 (47). Distances are calculated from the *N*-methyl group of the analogous ring to which the linker is connected on PA2 and PA3 to the carboxyl carbon of the methionine of the Hox docking peptide (FYPWM) bound to Exd in the crystal structure. (*C*) Table listing the K_d calculated from the EMSA and the fluorescence intensity (F.I.) extracted from the CSI profile for each polyamide–Exd complex.

As the ability to display more oligonucleotide features on a surface increases, the CSI approach can be easily scaled to represent larger sets of sequence variants. The most current technology (up to 10 positional variants), which is accessible to the entire scientific community, is sufficient to determine the binding preferences of nearly all metazoan DNA-binding proteins and engineered DNA-binding small molecules. Thus, this platform provides a powerful tool for tackling the important challenge of deciphering the DNA recognition code of DNA-binding molecules, individually or in cooperatively assembling complexes.

Materials and Methods

Duplex DNA Arrays. Microarrays were synthesized by using a Maskless Array Synthesizer (NimbleGen Systems, Madison, WI) (25). Homopolymer (T_{10}) linkers were covalently attached to monohydroxysilane glass slides. Oligonucleotides were then synthesized on the homopolymers to create a high-density oligonucleotide microarray. The array surface was derivatized such that the density of oligonucleotides was sufficiently low within the same feature so that no one oligonucleotide would hybridize with its neighbors. Four copies of every sequence required a total of 131,584 features per array.

Hairpin Formation Percentage. In two distinct features on the array we present two sequences: one that forms a hairpin (5′-CGC-TTAGTTCA-CGC-TCCT-GCG-TGAACTAA-GCG-3′) and one that does not (5′-CGC-TTAGTTCA-CGC-3′). By using a Cy3-labeled DNA probe that is complementary to the core sequence (5′-CGC-TTAGTTCA-CGC-3′) present in both oligonucleotides, we determined the ability of the complementary strand to bind the hairpin versus the single-stranded DNA molecules. The fluorescence intensity of the hairpin sequence was divided by the fluorescence intensity of the single-stranded sequence. The averaged background-subtracted intensity ratio of the double-stranded versus the single-stranded features indicated 95.6% hairpin formation.

Polyamide Synthesis. Polyamide-Cy3 conjugate PA1 was prepared by employing an orthogonally protected *N*-(phthalimidopropyl)pyrrole building block in standard Boc-based solid-phase synthesis (50). Cleavage of the polyamide from 100 mg of phenylacetamidomethyl (PAM) resin by treatment with 1 ml of dimethylaminopropylamine also removed the phthalimide protecting group to give the free base. The crude cleavage mixture was diluted with 0.1% trifluoroacetic acid (aq) and acetonitrile to a final volume of 5 ml and loaded onto a preconditioned solid-phase extraction column (C_{18} bonded phase). After washing with a 4:1 (vol/vol) solution of 0.1% trifluoroacetic acid (aq) and acetonitrile, product was eluted with methanol, and solvents were removed by azeotropic distillation from toluene. The resulting aminopropyl precursor of PA1 was a slightly yellow solid. Analytical HPLC and MALDI-TOF MS verified the identity and purity of this intermediate, and it was used without further manipulation.

The intermediate free base (0.5 μmol) was dissolved in 0.45 ml of anhydrous dimethylformamide and 0.05 ml of diisopropylethylamine. An amine-reactive Cy3 fluorophore (1 mg) (Amersham Pharmacia) was added to this solution, and the resulting mixture was agitated for 4 h. Crude products were purified by preparative HPLC. The purity and identity of the product was confirmed by analytical HPLC and MALDI-TOF MS.

PA1. UV-Vis (H_2O) λ_{max} in nm (ε in $M^{-1}\cdot cm^{-1}$): 313 (69,500), 555 (75,000). MALDI-TOF MS (monoisotopic) [M + H] 1,877.60 (1,877.81 calculated for $C_{91}H_{112}N_{24}O_{17}S_2$).

Binding Assay. Microarray slides were immersed in 1× PBS and placed in a 90°C water bath for 30 min to induce hairpin formation of the oligonucleotides. Slides were then transferred to a tube of nonstringent wash buffer (saline/sodium phosphate/EDTA buffer, pH 7.5/0.01% Tween 20) and scanned to check for low background (<200 intensity). Microarrays were scanned by using a ScanArray 5000 (GSI Lumonics, Billerica, MA), and the image files were extracted with GENEPIX PRO Version 3.0 (Axon Instruments, Foster City, CA).

Polyamide binding. Microarrays prepared as above were placed in the microarray hybridization chamber and washed twice with nonstringent wash buffer. Polyamide was diluted to 5 nM in Hyb buffer (100 mM Mes/1 M NaCl/20 mM EDTA, pH 7.5/0.01% Tween 20). Polyamide (5 nM) was then added to the hybridization chamber and incubated at room temperature overnight for 16 h. Finally, the microarrays were washed twice with nonstringent wash buffer and scanned.

Protein binding. The microarrays were washed with reaction buffer containing 150 mM potassium glutamate, 50 mM Hepes (pH 7.5), and 5% glycerol for 5 min. Cy3-labeled Exd (extradenticle) protein and polyamide were diluted in reaction buffer to a final concentration of 20 nM and 50 nM, respectively. This solution was added to the hybridization chamber and incubated for 30 min. Subsequently, the microarrays were washed with reaction buffer and scanned.

Data Processing. For each replicate, global mean normalization was used to ensure the mean intensity of each microarray was the same. Local mean normalization (51) was then used to ensure that the intensity was evenly distributed throughout each sector of the microarray surface. Outliers between replicate features were detected by using the *Q* test at 90% confidence and filtered out. The replicates were then quantile-normalized (52) to account for any possible nonlinearity between arrays. Duplicate features were then

averaged together. The median of the averaged features was subtracted to account for background.

Z scores were calculated as |signal − median|/standard deviation. Because of the right-handed tail effect, standard deviation of the background signal was on the basis of the standard deviation from the median of all signals less than the median. The relationship of *Z* scores to *P* values can be found in Table 2, which is published as supporting information on the PNAS web site. Motifs were then found by running several motif-finding algorithms (31–33) on sequences in the highest *Z* score bin. Logos (53) of each motif were then created by using sequences from the highest *Z* score bin that contained the motif.

Molecular Modeling. Molecular models were created by aligning the coordinates of Exd crystallized with DNA (Protein Data Bank ID code 1B8I) with DNA of hairpin polyamide crystallized with DNA (Protein Data Bank ID code 1M18) (54). The DNA was aligned at four different positions by using structural alignment software (46), creating consensus, consensus +1, consensus −1, and inverse binding of the polyamide relative to Exd. The distance from the *N*-methyl group of the heterocycle ring, which is analogous to the *N*-methyl group of the ring of our polyamide that bears the hexapeptide, to the carboxyl carbon of the methionine of the recruitment peptide bound to Exd in the crystal structure was then calculated for each of the four alignments. This calculation demonstrated the distance that the linker in our polyamide (PA2 and PA3) would have to reach to recruit Exd to DNA (46). The alignments were visualized by using VMD (Visual Molecular Dynamics) software (47). The linkers for PA2 and PA3 were then drawn and energy was minimized to estimate how far each linker could likely reach (48).

Dye Conjugation to Exd. A pET3A vector containing the Exd sequence (residues 1–88) was mutated by using standard quick-change mutagenesis procedures to replace the cysteine with a serine (C41S), and an arginine was replaced with a cysteine (R2C) to generate Exd R2C (see Fig. 9*A* and *Supporting Methods*, which are published as supporting information on the PNAS web site). This Exd mutant was found to be stable, and the mutation had a minimal effect on DNA binding affinity. Exd R2C was then labeled with Cy3 by using a Cy3 maleimide Mono-Reactive dye pack (no. PA23031; Amersham Pharmacia Biosciences). The molar dye/protein ratio was determined to be 0.96 (quantified as follows): [Cy3] = (A_{552} × dilution factor)/150,000 $M^{-1}\cdot cm^{-1}$; [Exd-R2C] = [A_{280} − (0.08 × A_{552})]/12,090 $M^{-1}\cdot cm^{-1}$; Dye/protein = [Cy3]/[Exd-R2C]; Exd sequence: ^{1}A(R→C)RKRRNFSK ^{11}QASEILNEYF ^{21}YSHLSNPYPS ^{31}EEAKEELARK ^{41}(C→S)GITVSQVSN ^{51}WFGNKRIRYK ^{61}KNI.

EMSAs. Forty-mer DNA sequences labeled with ^{32}P (as per standard methods) were used in all reactions. Reactions were performed in a buffer containing 150 mM potassium glutamate, 50 mM Hepes (pH 7.5), 2 mM DTT, 100 ng/μl BSA, 10% DMSO, and 10% glycerol. For the binding tests, polyamide-peptide conjugates (50 nM final concentration) and [^{32}P]DNA were incubated together for 30 min at 4°C. Exd was then added to bring the reaction volume to 20 μl. Exd final concentrations were 0.033, 0.1, 0.33, 1, 3.3, 10, 33, and 100 nM. These reactions were incubated at 4°C for 1 h, and 15 μl was loaded onto a prerun 10% acrylamide/3% glycerol gel (1× TBE: 90 mM Tris/64.6 mM boric acid/2.5 mM EDTA, pH 8.3).

We thank D. Page and C. Kendziorski for helpful discussions and L. Vanderploeg for help with the figures. We gratefully acknowledge the support of the University of Wisconsin Industrial and Economic Development Research Program and the National Foundation–March of Dimes (A.Z.A.), Computation and Informatics in Biology and Medicine Training Grant T15LM007359 (to C.L.W.), and National Institutes of Health Grant GM51747 (to P.B.D.).

1. Ansari, A. Z. & Mapp, A. K. (2002) *Curr. Opin. Chem. Biol.* **6,** 765–772.
2. Dervan, P. B. & Edelson, B. S. (2003) *Curr. Opin. Struct. Biol.* **13,** 284–299.
3. Darnell, J. E., Jr. (2002) *Nat. Rev. Cancer* **2,** 740–749.
4. Tuerk, C. & Gold, L. (1990) *Science* **249,** 505–510.
5. Fried, M. & Crothers, D. M. (1981) *Nucleic Acids Res.* **9,** 6505–6525.
6. Garner, M. & Revzin, A. (1981) *Nucleic Acids Res.* **9,** 3047–3060.
7. Galas, D. J. & Schmitz, A. (1978) *Nucleic Acids Res.* **5,** 3157–3170.
8. Heyduk, T., Ma, Y., Tang, H. & Ebright, R. H. (1996) *Methods Enzymol.* **274,** 492–503.
9. Heyduk, T. & Heyduk, E. (2002) *Nat. Biotechnol.* **20,** 171–176.
10. Strauss, H. S., Boston, R. S., Record, M. T., Jr., & Burgess, R. R. (1981) *Gene* **13,** 75–87.
11. Brockman, J. M., Frutos, A. G. & Corn, R. M. (1999) *J. Am. Chem. Soc.* **121,** 8044–8051.
12. Bulyk, M. L., Huang, X. H., Choo, Y. & Church, G. M. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 7158–7163.
13. Wang, J. K., Li, T. X. & Lu, Z. H. (2005) *J. Biochem. Biophys. Methods* **63,** 100–110.
14. Boger, D. L., Fink, B. E., Brunette, S. R., Tse, W. C. & Hedrick, M. P. (2001) *J. Am. Chem. Soc.* **123,** 5878–5891.
15. Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A. & Bulyk, M. L. (2004) *Nat. Genet.* **36,** 1331–1339.
16. Liu, X., Noll, D. M., Lieb, J. D. & Clarke, N. D. (2005) *Genome Res.* **15,** 421–427.
17. Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000) *Science* **290,** 2306–2309.
18. Iyer, V., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M. & Brown, P. O. (2001) *Nature* **409,** 533–538.
19. Sikder, D. & Kodadek, T. (2005) *Curr. Opin. Chem. Biol.* **9,** 38–45.
20. Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., *et al.* (2004) *Nature* **431,** 99–104.
21. Horak, C. E., Mahajan, M. C., Luscombe, N. M., Gerstein, M., Weissman, S. M. & Snyder, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 2924–2929.
22. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003) *Nature* **423,** 241–254.
23. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., *et al.* (2001) *Nucleic Acids Res.* **29,** 281–283.
24. Neidle, S. (2001) *Nat. Prod. Rep.* **18,** 291–309.
25. Singh-Gasson, S., Green, R. D., Yeu, Y., Nelson, C., Blattner, F., Sussman, M. R. & Cerrina, F. (1999) *Nat. Biotechnol.* **17,** 974–978.
26. Ansari, A., Kuznetsov, S. V. & Shen, Y. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 7771–7776.
27. Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993) *Nature* **365,** 512–520.
28. Tse, W. C., Ishii, T. & Boger, D. L. (2003) *Bioorg. Med. Chem.* **11,** 4479–4486.
29. Trauger, J. W., Baird, E. E. & Dervan, P. B. (1996) *Nature* **382,** 559–561.
30. Rucker, V. C., Foister, S., Melander, C. & Dervan, P. B. (2003) *J. Am. Chem. Soc.* **125,** 1195–1202.
31. Bailey, T. L. & Elkan, C. (1994) *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*, eds. Altman, R., Brutlag, D., Karp, P., Lathrop, R. & Searls, D. (AAAI Press, Menlo Park, CA), pp. 28–36.
32. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. (2000) *J. Mol. Biol.* **296,** 1205–1214.
33. Liu, X. S., Brutlag, D. L. & Liu, J. S. (2002) *Nat. Biotechnol.* **20,** 835–839.
34. White, S., Baird, E. E. & Dervan, P. B. (1996) *Biochemistry* **35,** 12532–12537.
35. Swalley, S. E., Baird, E. E. & Dervan, P. B. (1999) *J. Am. Chem. Soc.* **121,** 1113–1120.
36. Kielkopf, C. L., White, S., Szewczyk, J. W., Turner, J. M., Baird, E. E., Dervan, P. B. & Rees, D. C. (1998) *Science* **282,** 111–115.
37. Rauskolb, C., Peifer, M. & Wieschaus, E. (1993) *Cell* **74,** 1101–1112.
38. Mann, R. S. & Chan, S. K. (1996) *Trends Genet.* **12,** 258–262.
39. Arndt, H., Hauschild, K., Sullivan, D., Lake, K., Dervan, P. & Ansari, A. Z. (2003) *J. Am. Chem. Soc.* **125,** 13322–13323.
40. Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S. & Aggarwal, A. K. (1999) *Nature* **397,** 714–719.
41. Sen, D. & Gilbert, W. (1992) *Methods Enzymol.* **211,** 191–199.
42. Hauschild, K. E., Metzler, R. E., Arndt, H. D., Moretti, R., Raffaelle, M., Dervan, P. B. & Ansari, A. Z. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 5008–5013.
43. White, R. A., Aspland, S. E., Brookman, J. J., Clayton, L. & Sproat, G. (2000) *Mech. Dev.* **91,** 217–226.
44. Thanos, D. & Maniatis, T. (1996) *Methods Enzymol.* **274,** 162–173.
45. Ptashne, M. & Gann, A. (2002) *Genes and Signals* (Cold Spring Harbor Lab. Press, Woodbury, NY).
46. Guex, N. & Peitsch, M. C. (1997) *Electrophoresis* **18,** 2714–2723.
47. Humphrey, W., Dalke, A. & Schulten, K. (1996) *J. Mol. Graphics* **14,** 33–38.
48. (2005) CHEMDRAW ULTRA (CambridgeSoft, Cambridge, MA), Version 9.0.
49. LaRonde-LeBlanc, N. A. & Wolberger, C. (2003) *Genes Dev.* **17,** 2060–2072.
50. Baird, E. E. & Dervan, P. B. (1996) *J. Am. Chem. Soc.* **118,** 6141–6146.
51. Colantuoni, C., Henry, G., Zeger, S. & Pevsner, J. (2002) *Bioinformatics* **18,** 1540–1541.
52. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. (2003) *Bioinformatics* **19,** 185–193.
53. Schneider, T. D. & Stephens, R. M. (1990) *Nucleic Acids Res.* **18,** 6097–6100.
54. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242.