Specificity landscapes of DNA binding molecules elucidate biological function

Clayton D. Carlson1, Christopher L. Warren1,1, Karl E. Hauschild1, Mary S. Ozers1, Naveeda Qadir1, Devesh Bhimsaria1, Youngsook Lee2, Franco Cerrina1, and Asseem Z. Ansariad,2

1Department of Biochemistry, University of Wisconsin, 433 Babcock Drive, Madison, WI 53706; 2Department of Anatomy, University of Wisconsin, 1300 University Avenue, Madison, WI 53706; 3Department of Electrical and Computer Engineering, University of Wisconsin, 1415 Engineering Drive, Madison, WI 53706; and 4The Genome Center of Wisconsin, University of Wisconsin, 425G Henry Mall, Madison, WI 53706

Communicated by Sankar Adhya, National Institutes of Health, NCI, Bethesda, MD, December 11, 2009 (received for review August 21, 2009)

Evaluating the specificity spectra of DNA binding molecules is a nontrivial challenge that hinders the ability to decipher gene regulatory networks or engineer molecules that act on genomes. Here we compare the DNA sequence specificities for different classes of proteins and engineered DNA binding molecules across the entire sequence space. These high-content data are visualized and interpreted using an interactive “specificity landscape” which simultaneously displays the affinity and specificity of a million-plus DNA sequences. Contrary to expectation, specificity landscapes reveal that synthetic DNA ligands match, and often surpass, the specificities of eukaryotic DNA binding proteins. The landscapes also identify differential specificity constraints imposed by diverse structural folds of natural and synthetic DNA binders. Importantly, the sequence context of a binding site significantly influences binding energetics, and utilizing the full contextual information permits greater accuracy in annotating regulatory elements within a given genome. Assigning such context-dependent binding values to every DNA sequence across the genome yields predictive genome-wide binding landscapes (genomescapes). A genomescape of a synthetic DNA binding molecule provided insight into its differential regulatory activity in cultured cells. The approach we describe will accelerate the creation of precision-tailored DNA therapeutics and uncover principles that govern sequence-specificity of DNA binding molecules.

A major goal at the interface of synthetic biology, chemistry, and personalized medicine is to create molecules that can specifically regulate genes and thereby dictate cell fate (1–4). An approach toward this goal has been to engineer natural DNA binding molecules to target desired DNA sequences. In a widely used strategy, the DNA binding residues of zinc-finger (ZF) proteins have been substituted to generate molecules with altered DNA specificities (5–7). Another approach has focused on engineering DNA binding small molecules, imidazole/pyrrole hairpin polyamides, to increase their sequence specificity (8–11). Both classes of engineered DNA binders have shown tremendous promise in regulating gene expression, but a critical hurdle in their application as therapeutic agents or genomics tools is controlling the precision with which engineered molecules target a desired site when faced with millions of binding sites of varying affinities within a genome (12–14).

As a representative of the synthetic DNA binders we examined the binding characteristics of a hairpin polyamide (PA-1, Fig. 1) (15, 16). This class of molecules has been shown to regulate target genes in human cells and in mouse models of disease (13, 17, 18). Hairpin polyamides are typically composed of pyrrole and imidazole rings and can be engineered, using simple recognition rules (16), to target specific 6–12 base pair (bp) sequences (Fig. L4 and Fig. S1A and B) (15, 16). These molecules bind in the minor groove of DNA with high affinity recognizing naked or chromatinized DNA with modest perturbation of the DNA groove geometry (19). We also examined the specificity of an engineered ZF protein (ZF-P1, Fig. 1B). This protein domain has been utilized, not only to create molecules that regulate gene expression (5–7), but also to generate site-specific nucleases that greatly improve homologous recombination in mammalian cells (20, 21). To directly compare engineered molecules with natural DNA binding proteins, we examined members of different classes of DNA binding domains, representing ~75% of the human DNA binding proteome (Fig. 1B) (22, 23). Gata4 (C2-type zinc finger), Nkx-2.5 (homeodomain), and Junonji/Jarid2 (AT-Rich Interaction Domain) play a central role in cardiac development and disease (24, 25); p53 (Immunoglobulin-fold) is an aniontagonist that regulates genes involved in apoptosis (26); c-Abi (HMG domain) is implicated in cellular signaling and cancer (27); and TBP with its unusual minor groove binding fold is an essential component of the transcriptional machinery (28).

To elucidate the specificity of engineered DNA binders and compare them with natural DNA binding proteins, we performed Cognate Site Identifier (CSI) analysis (Fig. 1) (29–31). With CSI, the sequence preferences of DNA binding molecules are comprehensively examined by determining their affinity for nearly half a million duplex DNA sequences that collectively display the entire sequence space of a binding site on a microarray (Fig. L4) (31). The highest intensity sequences are interrogated by motif-finding algorithms to derive consensus motifs which are based on underlying position weight matrices (PWMs) (29, 32). To determine if PWM-derived motifs fully described the DNA binding preferences, we displayed the entire dataset of binding intensities as unique Sequence Specificity Landscapes (SSLs). An important advantage of this display, as opposed to twodimensional bar graphs, is that the landscape can be adjusted to optimize the cognate site motif(s) and thereby uncover major binding motifs and effects of flanking sequence.

Using the CSI-SSL approach, we determined the optimal DNA binding motifs for several transcription factors (TFs) and compared them to engineered molecules. Our findings indicate that consensus motifs, derived from PWMs, often mask or compress important binding specificity information generated from large datasets. Solution-based measurements validate that the CSI intensity values are strongly correlated with binding affinities, thus SSLs also function as comprehensive energy landscapes. Moreover, the landscape displays reveal that 6-ring hairpin polyamides discriminate between optimal and suboptimal sequences as

Author contributions: C.D.C., C.L.W., K.E.H., M.S.O., D.B., and A.Z.A. designed research; C.D.C., C.L.W., K.E.H., M.S.O., and D.B. performed research and analyzed data; C.L.W., D.B., K.E.H., N.Q., Y.L., and F.C. contributed new reagents/analytic tools; C.D.C., C.L.W., K.E.H., M.S.O., and A.Z.A. wrote the paper.

The authors declare a conflict of interest. A.Z.A. is a founder/proprietor and C.L.W. and M.S.O. are part-time employees of VistaMotif, and M.S.O. owns Invitrogen stock.

1C.D.C. and C.L.W. contributed equally to this work.

2To whom correspondence should be addressed. E-mail: ansari@biochem.wisc.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0914023107/DCSupplemental.
Results

Cognate Site Identification and Sequence-Specificity Landscapes (CSI-SSL). CSI analysis of the natural and engineered DNA binding molecules yielded a comprehensive binding profile across the entire sequence space of a binding site (Fig. S1C). The highest intensity binding sites were compiled to identify consensus binding motifs (Fig. S1D) which are displayed as sequence Logos (33) in Fig. 1. This comparative analysis indicates that the consensus motif bound by polyamide PA-1 is similar to motifs recognized by natural proteins, both in length and information content.

While consensus motifs derived from PWMs summarize the best binding sequences, they overlook significant insights embedded within the complete sequence recognition profiles captured by CSI analysis (Fig. 2A left panel) (34-36). To surmount this limitation, we created the SSL as an adaptable tool to display and interpret the full recognition preferences of DNA binding molecules (Fig. 2). SSLs present the entire binding dataset through a series of concentric rings or a linear format (Fig. 2A).

In a circular SSL, the innermost ring displays sequences that contain a perfect match to a given seed motif (0-mismatch). The subsequent circles, going outward, represent increasing mismatches from the seed motif. The height of each color-coded peak on the individual rings corresponds to the CSI binding intensity data. In this way, the entire sequence space can be displayed in a comprehensive yet easily interpreted format. To generate an SSL, a seed motif is used as the starting input; this motif may be generated from a known PWM, although a motif from any source may be used to initialize the sequence alignment. Through iterative optimizations the seed motif is refined until one obtains an SSL with high affinity sequences restricted to the 0-mismatch ring, and moderate-to-low affinity sequences assigned to the appropriate outer mismatch rings. The sequences in the center ring are sorted first alphabetically by motif (or submotif) and then alphabetically by the flanking sequences. This provides a consistent ordering of sequences in the center ring. The sequences represented in the outer rings are organized first by the position of the mismatch and then alphabetically (e.g. in the motif G1T2A1T3, followed by A2T3G2T2, and onwards through all positions of the motif). In the linear format, the display permits vertical alignment of the sequences by mismatch ring (Fig. S2B).

In essence, the SSL format reduces the high-dimensionality problem of displaying nonlocal sequence interdependencies of nearly a million different permutations into a readily interpretable graph that reports the entire specificity-spectrum of a DNA binding molecule.

A key advantage of the landscape display is that all data points are displayed without sequence compressions that are necessary to generate PWMs and consensus motifs. The absence of sequence compression in SSL facilitates the identification of optimal motif(s). If the seed motif, used to initiate an SSL, is too restrictive then several high-intensity sequences appear in the mismatch rings, and conversely if the motif is too inclusive then low-intensity peaks (i.e. valleys) would invade the innermost match ring (Fig. S2C). Importantly, if a chosen seed motif ignores the existence of alternative modes of binding to different sequences, then outer mismatch rings would display clustered peaks which would reveal the unique binding motifs (see below). Moreover, because the motifs identified by CSI (Fig. 1) are shorter than the permuted DNA duplexes on the array (5-7 bp motifs embedded within a 14-20 bp duplex), the landscapes identify the contributions of flanking sequences on the binding energetics of a given motif. In contrast to binding profiles predicted by PWMs (Fig. S2A), the flanking sequence influences binding properties to an unexpected degree, yielding highly textured binding profiles and rugged landscapes even in the 0-mismatch ring (Fig. 2). This additional contextual information greatly improves the accuracy of regulatory element annotation across genomes.
Specificity Landscapes of Natural and Engineered DNA Binders. The SSLs of natural and engineered DNA binders immediately reveal several interesting results (Fig. 2B). Hairpin polyamides display a high-degree of sequence specificity as compared to natural DNA binding proteins. For PA-1, as well as previously studied PA-2 (chemical structure of PA-2 shown in Fig. S1B) (31), the highest intensity peaks are clustered within the perfect match ring (innermost circle). While few strong binding sites are present in the 1-mismatch ring, virtually none are present in the 2-mismatch ring. By contrast, widely differing specificity spectra are evident in both the perfect match and 1-mismatch rings. For PA-2, the highest intensity peaks are clustered within the perfect match ring (innermost circle). While few strong binding sites are present in the 1-mismatch ring, virtually none are present in the 2-mismatch ring.

Deconvoluting Submotifs That Are Compressed into a “Consensus.” The SSLs of a homeodomain protein, Nkx-2.5, shows that the consensus motif 5′-TGT1GTG-3′ provides an incomplete view of the true binding specificity (Fig. 1B). The submotifs T1,T2-AAGTG and N1,T2-AAGTG are strong binding sites with TTAA GTG being the best (29). These compressed submotifs are readily deconvoluted as high-intensity peaks in the 1-mismatch ring of an SSL generated with the TTAAGTG seed motif (compare Fig. 2A with dual motif landscape in Fig. 2B). To explore the generality of this phenomenon, we examined motifs of murine homeodomain proteins that were recently reported (42). Evaluating all 18 classes of homeodomains using the SSL displays revealed that nearly 40% of the motifs could be optimized to better define the
specificity of the proteins (Fig. S2E). For example, in the case of Barx1 the reported consensus motif compresses two submotifs and inaccurately includes poorly binding sequences within the motif. The SSLs of the same dataset yield a more accurate view of the cognate sites preferred by Barx1 as well as several other homeodomains (Fig. S2E).

**Precision-Tailoring the Specificity of Engineered DNA Ligands.** CSI analysis combined with SSL display permits the rapid and rigorous evaluation of engineered DNA ligands with altered chemical composition and architecture. We synthesized a 6-ring hairpin polyamide (PA-3) rather than the larger 8-ring hairpin polyamide (Fig. 3A and Fig. S3). The CSI-SSL analysis shows that this minimized molecule binds a shorter core sequence but its specificity profile is comparable to 8-ring hairpin polyamides. This has significant implications in engineering smaller molecules that retain specificity yet have improved cell permeability properties.

The SSL profiles also indicate that modular recognition between heterocycle rings of polyamides and individual base steps of DNA governs recognition in this class of molecules. The results strongly support the recognition rules described by Dervan and others (9, 11, 15). We also used the SSL format to display CSI data of a hairpin polyamide (PA-4) containing an atypical 3-chlorothiophene ring (Fig. 3A and Fig. S3) (30). The SSL of this bioactive molecule validates the increased preference for Thymine by the thiophene ring. This result highlights the importance of the CSI-SSL approach in evaluating specificities of new chemical entities that target DNA. Finally, the SSL of a linear polyamide, PA-5, (30) that targets sites abnormally repeated in Friedrich’s ataxia patients shows binding to a larger 9 bp site (Fig. 3A and Fig. S3) (17). In the landscape, significant binding is detected in the outer mismatch rings. Examination of the outer-ring peaks identified three related submotifs (Fig. 3B) that were concealed in consensus motifs derived from PWMs (30). This example of relaxed specificity displayed by a linear polyamide (PA-5), along with the proteins described above, serve to highlight the ability of SSLs to identify multiple binding submotifs that are compressed into a single consensus by motif-finding algorithms. These results also reemphasize the utility of the CSI-SSL approach in creating and evaluating molecules that target DNA with the desired degree of specificity.

**Landslides Define Binding Energetics Across the Entire Sequence Space.** SSLs directly translate to binding energy landscapes for hairpin polyamides because their CSI intensities correlate strongly with equilibrium binding energies measured in solution (29, 30). To evaluate if this relationship holds for a natural DNA binding protein, we measured affinities of Nkx-2.5 for six DNA sequences that span a range of CSI intensities. As a control we refined the previously reported (29) affinity of PA-1 for a similar range of sites identified by CSI. The binding isotherms for these sequences were determined by nuclease protection assays for PA-1 and by EMSA for Nkx-2.5 (Fig. 4 and Fig. S4). The CSI intensities of both molecules were highly correlated with equilibrium binding energy. Not only do we observe a clear correlation at strong-to-moderate binding sites, the CSI intensities indicate differential binding at suboptimal sites that are too weak to be evaluated by EMSA analysis. The linear correlation observed, indicates that the fractional occupancy on the array is sufficiently low so as to still correlate with solution binding affinity. The low fractional occupancy is likely due to a combination of multiple factors. Thus, SSLs directly translate to energy landscapes and provide unprecedented information on binding energetics across broad sequence space (see Fig. 5).

**Genomescapes: Genome-Wide Binding Landscapes Explain the Differential Bioactivity of an Engineered DNA Binder.** A compelling need for developing CSI-SSL is to map the regulatory elements across the genome. Based on the CSI scores, we assigned binding probabilities across the entire human genome (Fig. 5). We designate these genome-wide binding landscapes, as “genomescapes.” Here, we focused on the CSI-SSL scores of a hairpin polyamide (PA-4) that targets the binding sites of the hypoxia induced transcription factor, HIF-1α. PA-4 binding to HIF-responsive-elements (HREs) blocks the activation of the hypoxia induced genes (18), including VEGF, a cytokine that is implicated in cancer and angiogenesis (43). Blocking HIF-1α dependent expression of VEGF suffices for averting tumorigenesis, thus marking this binding event as a target for therapeutic intervention. Previous genome-wide expression profiles indicated that genes bearing an HRE could be targeted in cells by the designed polyamide (PA-4) (Fig. 5B) (18). However, in those studies it was not clear why the endothelin gene, ET-2, bearing a moderate HRE was robustly inhibited by the compound.
To resolve this outstanding issue, we examined the genomic DNA for nonobvious binding sites that may exert additional inhibition. The CSI genomescape of ET-2 promoter regions identifies moderate PA-4 sites across the known HIF-1α binding site. More importantly, in contrast to other HIF-1α regulated genes, ET-2 also has several moderate PA-4 binding sites across the Transcription Start Site (TSS) and into the coding region (Fig. 5C and Fig. S5). At 1 μM concentrations of the polyamide in the media, ET-2 gene is robustly down-regulated. CSI genomescapes suggest that this unusually high level of inhibition might be due to multiple synergistic inhibitory events where both the transcription factor and the transcriptional machinery are occluded from their binding sites. The binding energetics for this compound suggests that nanomolar binding affinities are required for effective action of the compound in living cells. We further anticipate that the CSI genomescapes will greatly aid in predicting how DNA binding molecules localize across the genome and in elucidating seemingly off-target events in living cells.

**Discussion**

The combined CSI and SSL analyses of natural and engineered molecules lead to the unexpected conclusion that synthetic DNA binders can achieve specificities that match or exceed natural DNA binding proteins. The SSL displays overcome the limitations of consensus motifs derived from PWMs and reveal an unprecedented view of the entire specificity and energy landscape of DNA binding molecules. The varying specificity constraints imposed by different protein or small molecule structural folds offer insights into the mechanisms for modulating specificity in molecular recognition. Elucidating the effects of sequence context on binding energetics permits an accurate annotation of regulatory elements across the genome. This is particularly true for synthetic DNA binders, and the resulting genomescapes predict the biological outcomes with a high degree of accuracy. The CSI-SSL approach will be invaluable in designing, evaluating, and refining unique chemical entities that target genomic sites with the desired degree of precision. This is a critical step in the generation of molecules that will have therapeutic utility and also serve as powerful genomic tools.

In the near future, high-throughput data for most DNA binding proteins from various organisms will become publicly available. As demonstrated with the homeodomain family of proteins (Fig. S2E), high-dimensionality data when viewed through SSLs reveal unique modes of sequence recognition. SSL analysis of such datasets will address how natural proteins target cognate...
sites in the genome, how sequence specificity evolved within different DNA binding folds, and how regulatory elements and networks are evolutionarily retained and organized.

The Sequence-Specificity Landscapes that we describe here are not limited to displaying DNA binding data. They can be readily applied to any high-dimensional data that define interaction interfaces of biopolymers or other molecules, including protein-peptide or protein-small molecule molecules. An integrative landscape that combines landscapes along with other physical and biological analyses will greatly enhance our understanding of the elusive principle of “specificity.”

Materials and Methods

CSI Microarray Analysis—Polyamide. The fluorescently-labeled polyamide was diluted to a final concentration (250 nM for PA-1, PA-2; 20 nM for PA-3; and 10 nM for PA-4) in hybridization buffer (1 M NaCl, 100 mM MES pH 7.5, 20 mM EDTA, 0.01% Tween-20). The polyamide was added to the hybridization chamber (Grace BioLabs) on the array and rotated for 1 h at 23 °C. The arrays were washed with nonstringent wash buffer (6X SSPE pH 7.5, 0.01% Tween-20), dried, and scanned using an Axon 4000B (Molecular Devices). Data was viewed using GenePix™ Pro 6.0 (Molecular Devices).  

CSI Microarray Analysis—Protein. Arrays were blocked with 2.5% nonfat dried milk for 1.5 h. The protein was diluted to a final concentration in buffer and mixed with a directly-labeled fluorescent antibody to either the protein or a tag on the protein. The protein-antibody mixture was added to the hybridization chamber on the array and rotated for 1 h. The arrays were washed, dried, and scanned as above.

Generation of Specificity Landscapes. To order the binding data for presentation in the SSL, the binding site on each sequence is determined using the best match to the seed motif. Landscapes are optimized by maximizing the number of high-intensity sequences and minimizing the number of low-intensity sequences in the 0-mismatch ring. The data is parsed according to the number of mismatches from the motif being plotted. If the seed motif has multiple submotifs, we use the minimum number of mismatches to the best submotif. Each ring is sorted first by the position of the mismatch and then alphabetically. The intensities are then plotted using MatLab R2006a (The MathWorks, Inc.). The intensities are smoothed using 0.5% of the data on each ring.

Additional details are included in SI Text.

ACKNOWLEDGMENTS. We thank P.B. Dervan, S. Keles, J. Mitchell, and S. Cunningham for invaluable discussions; S. Foister and P. Dervan for the gift of PA1 (SfiI-02); and J. Markley and the Center for Eukaryotic Structural Genomics for plasmids. This work was supported by grants from the National Institutes of Health (GM069420), March of Dimes, U.S. Department of Agriculture, Innovation and Economic Development Research Program, and Vilas Associate and Shaw scholar awards to (A.Z.A.). C.D.C. was supported by the American Heart Association predoctoral fellowship (0615615Z). C.L.W. was supported by a National Institutes of Health/National Library of Medicine predoctoral fellowship (T15LM007359). K.E.H. was supported by a Natural Sciences and Engineering Research Council fellowship. D.B. was supported by the Khorana program. Y.L. was supported by the National Institutes of Health (HL07050).

Supporting Information

Carlsen et al. 10.1073/pnas.0914023107

Methods. Protein purification. GST-tagged Gata4 (mouse) was over-expressed in BL21-lysS cells (plasmid from Professor Youngsook Lee) that were lysed by sonication and cleared by centrifugation. Gata4 was enriched using ammonium sulfate precipitation followed by purification on glutathione-sepharose beads. The eluted protein was further purified by ion exchange chromatography over an SP-sepharose column. His-tagged Nkx-2.5 (mouse) was overexpressed in BL21 as an MBP-His-Nkx-2.5 fusion protein. Cells were lysed by sonication and cleared by centrifugation. The fusion protein was then purified using amylase resin followed by MBP tag cleavage using AcTev (Invitrogen) overnight at 4 °C. The His-tagged Nkx-2.5 was then separated from the solution by ion exchange chromatography using SP-sepharose. Any residual undigested MBP-Nkx-2.5 was removed by using a second amylase resin column.

GST-Jumonji (mouse) DNA binding domain (aa 529 – 792) was over-expressed in BL21-lysS cells (plasmid from Professor Youngsook Lee) that were lysed by sonication and cleared by centrifugation. Fusion protein was captured on glutathione-sepharose and after elution was further purified by ion exchange over a SP-sepharose column.

ZFP-1 used for CSI analysis was generated using the wheat germ expression system (1). cDNA for each gene was cloned into the pE-Flag-His expression plasmid, modified to include an N-terminal 5x-Myc tag. The purified plasmid was then used in an in vitro transcription reaction with Stp6 RNA polymerase (Promega, Madison, WI) at 37 °C for 3 h. The mRNA was then ethanol precipitated and used in an in vitro translation reaction with wheat germ cell-free extract (CellFree Sciences, Matsuyama Ehime Japan) at 16 °C for 20 h. Cleared reactions containing soluble Myc-tagged proteins were used directly for CSI analysis.

Synthesis of PA-3. Polyamide PA-3 was synthesized by following well-documented solid-phase synthetic techniques using Boc protection chemistry (2). The final polyamide product was purified by HPLC and validated by its mass spectra. The monomer 2,3-Benzotriazol-1-yl 4-{(tert-Butoxycarbonyl)amino-1-methyl-pyrrole-2-carboxylate was synthesized according to Baird and Carlson et al. www.pnas.org/cgi/doi/10.1073/pnas.0914023107

Polynucleotide protection assay of PA-1. PA-1 Dnase I footprinting experiments were performed using the parent PA-1 (unlabeled) polyamide on two 223-bp DNA fragments, GGT1 and GGT2. GGT1 contained five DNA binding sites (G – AAGGTTA, H – AACGTTA, J – ATGGAAAT, K – ATCGTAT, L – AAGGT CA). Site G was present in two locations as an internal control to ensure binding site proximity did not affect the association constants (Kd) determined. GGT2 contained three DNA binding sites (G, J – CAGGTTA, and K). Sites G and K were used to normalize the Kd values determined on GGT2 with those on GGT1. For the footprints, 10 μL reactions volumes containing varying concentrations of the parent PA were mixed in the following buffer: 50 mM KCl, 50 mM Tris-HCl (pH 7.5), 2 mM CaCl2, 2 mM MgCl2, 5% Glycerol and 100 ng/μL BSA. Each lane contained approximately 10 K cpm of radiolabeled GGT1 or GGT2 DNA. The polyamide was allowed to equilibrate for 1 h at room temperature before 1 μL of Dnase I (Invitrogen part number 18047-019; diluted 1:50,000 prior to use) was added. After 30 s of digestion, 10 μL of a stop buffer containing 10 mM EDTA, 10 mM NaOH, 50% Formamide, 0.01% xylene cyanol and 0.01% bromphenol blue was added, and the reactions were immediately heated to 95°C for 5 min to quench the Dnase reaction. The reactions were chilled on ice and loaded onto a 8% acrylamide/7M Urea denaturing gel. The gels were run at 2,000 V in 0.5 × TBE until the bromphenol blue was run off the gel. The gels were dried, exposed to a phosphorimager screen overnight, and visualized using a Typhoon imager. Tripletic experiments were performed, and binding constant determinations were obtained using ImageQuant 5.2 to quantify the footprinting gel data and SigmаПlot 6.0 for nonlinear regression. Binding energies were determined using ΔG = −RTln Kd, with R = 1.98 cal · K−1 · mol−1 and T = 296.15 K.

GGT1 contained five DNA binding sites (G – GGT1 contained three DNA binding sites (G, J – ATCGTAT, K – ATCGTAT, and L – AAGGTCA). Site G was present in two locations as an internal control to ensure binding site proximity did not affect the association constants (Kd) determined. GGT2 contained three DNA binding sites (G, J – CAGGTTA, and K). Sites G and K were used to normalize the Kd values determined on GGT2 with those on GGT1. For the footprints, 10 μL reactions volumes containing varying concentrations of the parent PA were mixed in the following buffer: 50 mM KCl, 50 mM Tris-HCl (pH 7.5), 2 mM CaCl2, 2 mM MgCl2, 5% Glycerol and 100 ng/μL BSA. Each lane contained approximately 10 K cpm of radiolabeled GGT1 or GGT2 DNA. The polyamide was allowed to equilibrate for 1 h at room temperature before 1 μL of Dnase I (Invitrogen part number 18047-019; diluted 1:50,000 prior to use) was added. After 30 s of digestion, 10 μL of a stop buffer containing 10 mM EDTA, 10 mM NaOH, 50% Formamide, 0.01% xylene cyanol and 0.01% bromphenol blue was added, and the reactions were immediately heated to 95°C for 5 min to quench the Dnase reaction. The reactions were chilled on ice and loaded onto a 8% acrylamide/7M Urea denaturing gel. The gels were run at 2,000 V in 0.5 × TBE until the bromphenol blue was run off the gel. The gels were dried, exposed to a phosphorimager screen overnight, and visualized using a Typhoon imager. Tripletic experiments were performed, and binding constant determinations were obtained using ImageQuant 5.2 to quantify the footprinting gel data and SigmаПlot 6.0 for nonlinear regression. Binding energies were determined using ΔG = −RTln Kd, with R = 1.98 cal · K−1 · mol−1 and T = 296.15 K.

EMSAs of Nkx-2.5. Thirty-seven nucleotide DNA hairpins were labeled with 32P using standard protocols. Reactions were performed in binding buffer (50 mM NaCl, 10 mM Tris- HCl (pH 7.4), 1 mM MgCl2, 0.5 mM EDTA, 0.25% nonfat dried milk, and 10% glycerol). Binding reactions were incubated on ice for 1 h before loading onto the gel. The reactions were resolved through a prerun 10% polyacrylamide/3% glycerol gel in TBE until the bromphenol blue was run off the gel. The gels were dried, exposed to a phosphorimager screen overnight, and visualized using a Typhoon imager. Tripletic experiments were performed, and binding constant determinations were obtained using ImageQuant 5.2 to quantify the EMSA gel data and SigmаПlot 6.0 for nonlinear regression. Binding energies were determined using ΔG = −RTln Kd, with R = 1.98 cal · K−1 · mol−1, and T = 296.15 K.

CSI microarray analysis. Polynucleotide. The fluoresceingly-labeled polyamide was diluted to a final concentration (250 nM for PA-1, PA-2, PA-5; 20 nM for PA-3; and 10 nM for PA-4) in Hybridization buffer (1M NaCl, 100 mM MES pH 7.5, 20 mM EDTA, 0.01% Tween-20). The polyamide solution was added to the hybridization chamber (Grace Bio-Labs, Bend OR) on the array and incubated for 1 h at room temperature with constant rotation. The arrays were washed with non-stringent wash buffer (6X SSPE: pH 7.5, 0.01% Tween-20), dried, and scanned using an Axon 4000B (Molecular Devices, Sunnyvale CA). Data
was viewed using GenePix™ Pro 6.0 microarray analysis software (Molecular Devices).

**Protein.**

Arrays were blocked with 2.5% non-fat dried milk for 1.5 h prior to protein-antibody incubation. The protein was diluted to a final concentration in buffer and mixed with a directly-labeled fluorescent antibody to either the protein or a tag on the protein. The protein-antibody mixture was added to the hybridization chamber on the array and incubated for 1 h at either room temperature (or 4 °C) with constant rotation. The arrays were washed, dried, and scanned using an Axon 4000B. Data was viewed using GenePix™ Pro 6.0 microarray analysis software.

Protein concentrations and antibodies used were as follows: cell-free lysate and used at 10 μL with AlexaFluor 647-labeled α-myc antibody (AbD Serotec, Morphosys); purified mouse his-tagged Nkx2.5 at a final concentration of 50 nM was incubated with AlexaFluor 674-labeled α-his antibody (Qiagen, Valencia CA); purified GST-tagged mouse Gata4 at a final concentration of 50 nM was incubated with AlexaFluor 555-labeled α-GST antibody (Upstate, Millipore, Billerica, MA); purified mouse GST-tagged Jumonji at a final concentration of 200 nM was mixed with AlexaFluor 555-labeled α-GST antibody (Upstate, Millipore, Billerica, MA); purified human his-tagged p53 (ProteinOne) was used with AlexaFluor 647-labeled α-his antibody (Qiagen, Valencia CA); purified human his-tagged c-Abl (Invitrogen,) was used with AlexaFluor 647-labeled α-his antibody (Qiagen, Valencia CA); purified human his-tagged TBP (ProteinOne) was used with AlexaFluor 647-labeled α-his antibody (Qiagen).

**CSI data normalization.** For each array replicate, global mean normalization was used to equalize the mean intensity of each microarray. Local mean normalization (3) was then used to ensure the intensity was evenly distributed throughout each sector of the microarray surface. Outliers between replicate features were detected using the Q test at 90% confidence and filtered out. The replicates were then quantile normalized to account for any possible nonlinearity between arrays (4). Replicate features were averaged together. The center of the normal distribution (ignoring the right-handed tail) of the averaged features was subtracted to account for background.

Z-scores were calculated as | signal minus median/standard deviation. Due to the right-handed tail effect, standard deviation of the background signal was based upon the standard deviation from the median of all signals less than the median. Motif searching was conducted using MEME/MAST System Motif Discovery and Search (http://meme.sdsc.edu/meme/Intro.html) on sequences in the highest Z-score bin (5). Logos of each motif were created using sequences from the highest Z-score bins (6).

**Genomescapes of PA-6.** Based on several hairpin polyamide CSI profiles, we simulated the binding properties of the control polyamide PA-6 across the genome. Consistent with the reported genome-wide expression profiles, our simulated CSI genomescape for PA-6 does not identify strong-to-moderate binding sites at any of the four HIF-regulated genes (Fig. S5B). As the repertoire of binding profiles for different polyamide architectures grows, and the recognition properties are better refined by CSI-SSL analysis, the accuracy of predicting the binding potential and regulatory properties of newly engineered polyamides will sharpen.

**PDB codes for Figure 1.** Polyamide -1M19; C2H5 – 1ZAA; Nk2 Homeodomain – 1Nk2; C; ZFP – 1GAT; Ig Fold – 2ATA; ARID – 1KQQ; HMG – 2LEF; Beta Scaffold – 1TGH.

**Fig. S1.** (A) Polyamide pairing rules. (Right) Diagram representing polyamide binding to DNA. The C-G base pair is recognized by a pyrrole/imidazole ring pair, G-C by imidazole-pyrrole, T-A by hydroxypyrrole-pyrrole and TA or A-T by pyrrole-pyrrole. (Left) Schematic of polyamide. Filled circle: N-methyl imidazole; open circle: N-methyl pyrrole; gray circle: hydroxypyrrole; turn: \(\gamma\)-aminobutyric acid; diamond: \(\beta\)-alanine; half-circle with a positive charge: dimethylaminopropylamide. (B) CSI-determined PWM for hairpin polyamide PA-2. (Upper) Logo for PA-2. (Center) Schematic of PA-2. Each ring pair is aligned to the base pair in the PWM to which it binds. (Lower) Below the schematic is the full polyamide structure (in black) with attached Cy3 fluorescent dye (Green). Ring pairs in the full PA-2 structure are aligned with the ring pairs in the PA-2 schematic. Filled circle: N-methyl imidazole; open circle: N-methyl pyrrole; open circle with inner dot: pyrrole with attached Cy3 dye; turn: \(\gamma\)-aminobutyric acid, diamond: \(\beta\)-alanine; half circle with a positive charge: dimethylaminopropylamide. (C) Histograms of CSI data. Histograms of the Z-scores for all DNA binding molecules reported. (D) Position Specific Probability Matrices (PSPMs). Frequency matrices for all transcription factors from Fig. 1 as well as PA-2. (Left) Engineered modules (polyamides and C2H2 zinc finger). (Center and Right) Additional transcription factor DNA-binding folds. A value of 0.01 was used as a pseudocount where the frequency of a base at a specific position was zero to avoid calculation errors in the PWM-based SSL in Fig. S3B.
A

Nkx-2.5
TTAAGTG
Frequency Matrix

B

Nkx-2.5
TTAAGTG
Full Linear Landscape

C

Nkx-2.5
CTGNACG
D

Finger:

Predicted: GGA - GTT - AAG

Analysis:
High intensity peaks in the 1 - 3 mismatch rings indicate the seed motif is too restrictive in the last positions.

3 2 1
GGA - GTT - AA

Analysis:
The presence of only high intensity peaks in the match ring coupled with low peaks in the outer rings indicate seed motif is correct.

3 2
GGA - GTT

Analysis:
Low intensity peaks (valleys) in match ring indicate seed motif is too degenerate.

3 2
R - GGA - GT

Analysis:
Seed motif is incorrect because of poor binding in match ring and high binding in multiple mismatch rings.
Fig. S2. (A) Specificity landscapes for Nkx-2.5 using PSPM. Specificity landscape of Nkx-2.5 using the PSPM instead of CSI data. In this case, the frequency of each position in the binding site on a probe are multiplied together to yield a PSPM “intensity” for that probe. (B) Complete linear SSL for Nkx-2.5. The linear SSL for Nkx-2.5 was generated using the seed motif 5′-TTAAGTG-3′. All mismatch panels are displayed corresponding to the mismatch rings from the circular version. (C) Specificity landscapes for Nkx-2.5 using an incorrect binding motif. The SSL of Nkx-2.5 was created using an incorrect motif, resulting in peaks present only in the outer mismatch rings. (D) CSI-SSL analysis of engineered zinc finger ZFP-1. SSL using a seed motif representing most of the predicted motif (Upper Left) is too exclusive, while a seed motif based on the PWM-Logo (Upper Right) is too inclusive. An SSL using a seed motif from only the second and third fingers (Lower Left) represents the data well, whereas an SSL using a seed motif from the first two fingers (Lower Right) is clearly incorrect. (E) SSLs deconvolute complex DNA binding motifs that are compressed in PWMs. (a) Optimized specificity landscapes reveal more accurate binding motifs than PWMs for many transcription factor classes, such as in the data presented by Berger et al. The sequences (Arrows) for Barx1 partition into more optimal concentric rings when submotifs (Red) are plotted. (b) Representatives from all 18 murine homeodomain classes were analyzed using SSLs. Binding profiles for 40% of the classes were significantly improved by using multiple sub-motifs. The Logos from PWMs of high-throughput microarray binding data for these representatives are shown (Left). SSL (Center) using the entire binding data set of the given Logo are compared to optimized SSLs (Right) for each of the transcription factors. The reverse strand of the Nkx-2.2 motif is used in the SSL. (c) Logo and SSL of homeodomains that were not improved upon.
Fig. S3. Structures of the alternative polyamides and their corresponding CSI histograms. (Left) Chemical structures of PA-3, PA-4, and PA-5. (Right) CSI microarray histograms of the resulting Z-scores for each corresponding polyamide. PA-4 and PA-5 are from Puckett et al.
Fig. S4. Binding isotherms for PA-1 and Nkx-2.5. Binding isotherms for each individual sequence for PA-1 (A–F, Upper) and Nkx-2.5 (G–L, Lower) as determined from nuclease protection assay or EMSA, respectively. Each experiment was performed in triplicate, and a representative binding isotherm is shown. Tables below each set of isotherms display the sequence, experimentally determined KA (footprint or EMSA), and CSI intensity for each data point.
Additional correlation of gene expression microarray data with CSI microarray data. Genomescape data for the human ET-1 and ET-3 genes using PA-4. CSI data. The schematic of the polyamide is at the top of the column. Genomescapes are generated by assigning an intensity to every 10 bp sequence in the genome based on the CSI microarray data. The genomescape data for 100 Mbp of the chromosome surrounding either the ET-1 or ET-3 gene is shown. The genomescape is focused on two 100 bp regions containing the HRE and the TSS for each gene. The fold repression (Right) was reported by Olenyuk et al. using gene expression microarray data to compare the RNA expression levels from an untreated cell line to that of one incubated with the PA. (B) Correlation of gene expression microarray data with CSI microarray data for PA-6. Genomescape data for HIF regulated genes using predicted PA-6. CSI data. A schematic of the polyamide is at the top of the figure. Genomescapes are generated by assigning an intensity to every 10 bp sequence in the genome based on CSI microarray data. The genomescape data for 100 Mbp of the chromosome surrounding each target gene is shown. The genomescape is focused on two 100 bp regions containing the HRE and the TSS for each gene. The fold repression (Right) was reported by Olenyuk et al. using gene expression microarray data to compare the RNA expression levels from an untreated cell line to that of one incubated with the PA.