CHAPTER ONE

SEQUENCE-SPECIFICITY AND ENERGY LANDSCAPES OF DNA-BINDING MOLECULES

Joshua R. Tietjen, Leslie J. Donato, Devesh Bhimisaria, and Aseem Z. Ansari

Contents
1. Introduction 4
2. Array-Based Cognate Sequence Identification 6
   2.1. Labeling DNA-binding molecules 6
   2.2. Label-free detection 12
3. Solution-Based Cognate Sequence Identification 13
4. Data Analysis and Visualization of Specificity, Binding Energy, and Genomic-Association Landscapes 18
   4.1. Sequence-specificity landscapes 18
   4.2. Binding energy landscapes 23
   4.3. Genomescapes 25
References 27

Abstract
A central goal of biology is to understand how transcription factors target and regulate specific genes and networks to control cell fate and function. An equally important goal of synthetic biology, chemical biology, and personalized medicine is to devise molecules that can regulate genes and networks in a programmable manner. To achieve these goals, it is necessary to chart the sequence specificity of natural and engineered DNA-binding molecules. Cognate site identification (CSI) is now achieved via unbiased, high-throughput platforms that interrogate an entire sequence space bound by typical DNA-binding molecules. Analysis of these comprehensive specificity profiles is facilitated through the use of sequence-specificity landscapes (SSLs). SSLs reveal new modes of sequence cognition and overcome the limitations of current approaches that yield amalgamated “consensus” motifs. The landscapes also reveal the impact of nonconserved flanking sequences on binding to cognate sites. SSLs also serve as comprehensive binding energy landscapes that provide insights into the energetic thresholds at which natural and engineered molecules function

Department of Biochemistry, The Genome Center of Wisconsin, University of Wisconsin-Madison, Madison, Wisconsin, USA

ISSN 0076-6879, DOI: 10.1016/B978-0-12-385075-1.00001-9 All rights reserved.
within cells. Furthermore, applying the CSI binding data to genomic sequence (genomescapes) provides a powerful tool for identification of potential in vivo binding sites of a given DNA ligand, and can provide insight into differential regulation of gene networks. These tools can be directly applied to the design and development of synthetic therapeutic molecules and to expand our knowledge of the basic principles of molecular recognition.

1. Introduction

To understand signal-responsive regulation of gene networks, it is crucial to obtain a comprehensive view of the molecular recognition events that underlie transcription factor–DNA interactions. Similarly, from a bioengineering perspective, designing programmable regulators of gene expression requires a detailed understanding of the molecular recognition events between DNA and the engineered protein or small molecule. However, for the majority of the natural transcription factors, there is limited knowledge of where these proteins bind across the genome to regulate specific genes and networks in response to cell-signaling events. In an effort to understand and control desired gene networks, synthetic and chemical biologists have engineered small molecules and proteins to mimic the sequence specificity and regulatory properties of natural DNA-binding factors (Ansari and Mapp, 2002; Dervan et al., 2005; Gonzalez et al., 2010; Gottesfeld et al., 2000; Lee and Mapp, 2010; Mapp and Ansari, 2007; Rodríguez-Martínez et al., 2010; Sera, 2010; Wolfe et al., 2000). Defining the DNA sequence specificity of these engineered molecules is an important step in targeted regulation of genes and networks. Moreover, defining entire gene networks that coordinately execute a cellular response can reveal key nodes to which such engineered regulators can be targeted.

The importance of understanding the specificity of DNA-binding molecules has stimulated the development of several experimental methods over the years. Most methods for determining the protein–DNA or small molecule–DNA-binding signatures have been quite apt at identifying the highest affinity sequences. These include methods like the systematic evolution of ligands by exponential enrichment (SELEX) and cyclic amplification and selection of targets (CASTing) (Blackwell and Weintraub, 1990; Oliphant et al., 1989; Tuerk and Gold, 1990; Wright et al., 1991). While valuable, such motifs often display limited ability to predict sequences that are bound in a biological context, because physiologically relevant binding sites frequently deviate from the highest affinity sequences (Farnham, 2009; Ward and Bussemaker, 2008). Often, moderate or even low-affinity sites regulate expression in a switch-like manner, and cooperative binding of multiple transcription factors is required (Moretti and Ansari, 2008; Ptashne and
Gann, 2002). In fact, cooperative assembly integrates multiple signals combinatorially to fine-tune the regulation of genes, which outnumber the transcription factors encoded in genomes by an order of magnitude (Lander et al., 2001; Maniatis et al., 1998; Ptashne and Gann, 2002; Venter et al., 2001; Yamamoto et al., 1998). Studies that comprehensively define these binding properties in cells usually correlate with their corresponding binding patterns in cells (Harbison et al., 2004; Horak et al., 2002). However, when they do not correlate, unexpected mechanistic insights can be obtained.

Recently, methods such as the cognate site identification (CSI) have been developed that can reveal the entire spectrum of DNA-binding specificities of a protein or a small molecule binding individually or in cooperative complexes (Carlson et al., 2010; Stormo and Zhao, 2010; Warren et al., 2006). Array-based methods display millions of unique sequences of B-form DNA in spatially resolved microfabricated arrays on glass slides. A DNA-binding factor is applied to the surface and allowed to query the entire sequence space. The factor is either directly fluorescently labeled or can be indirectly detected. After incubation, the unbound factor is washed away and molecules that remain bound to DNA sequences on the array are visualized. Alternative methods, such as fluorescence intercalator displacement (FID), surface plasmon resonance (SPR), or gated microfluidic chambers (MITOMI), have also been used to determine specificity and affinity of binding (Brockman et al., 1999; Campbell and Kim, 2007; Hauschild et al., 2009; Maerkl and Quake, 2007).

In each of the array-based methods, the complexity of the DNA library is limited by the number of sequences that can be displayed, which in turn limits the size of the variable region in the DNA sequences queried. Currently, the available technology limits the number of features on an array to 2.1 (for NimbleGen arrays) or 6 millions (for Affymetrix arrays). This restricts the size of the variable region to every sequence permutation of up to 10–12 contiguous base pairs or the same number of base pairs distributed within a larger DNA duplex. The complexity of the sequences displayed on an array can be expanded through the use of De Bruijn sequences, but this approach decreases the quality of the data obtained. For example, arrays that rely on the De Bruijn approach to display every 10mer only yield optimal data for 8mer binding sites. In addition to loss of information, the approach also increases the complexity of analyzing the resulting data (Philippakis et al., 2008).

To expand beyond the limitations on the sequence space queried by array-based methods, cognate site identity can be determined by directly sequencing DNA that binds to a protein of interest (Jolma et al., 2010; Stormo and Zhao, 2010; Zhao et al., 2009; Zykovich et al., 2009). In these experiments, a protein of interest is incubated with a library of DNA sequences that is not subject to the limitations on length or number of features that can be spatially resolved on a surface. The micromole-scale
DNA synthesis permits the examination of sequence space of 14–25mer within a much larger duplex. In these solution-based assays, the protein–DNA complexes are isolated and the DNA is sequenced by next-generation DNA sequencers. The frequency of detecting a certain sequence directly correlates with the affinity of the protein for that sequence. As costs for direct sequencing decrease and the ease of use increases, we anticipate that this approach to defining the sequence specificity of natural or engineered DNA-binding molecules will become more prevalent.

With the magnitude of such comprehensive molecular recognition data comes the challenge of mining that rich vein of information and identifying the full spectrum of binding specificities. The processing and visualization of such multidimensional data can be achieved using sequence-specificity landscapes (SSLs), which are designed to allow user-friendly visualization of DNA-binding characteristics of a protein or a small molecule of interest (Carlson et al., 2010). SSLs provide a convenient method for displaying and analyzing the binding specificity profile for natural or engineered DNA-binding molecules across the entire sequence space mapped by CSI. This displays both high- and low-affinity binding sites, rather than only the best bound sequences. Importantly, since binding intensities on the array correlate with affinities in solution, SSLs also yield comprehensive energy landscapes for DNA-binding molecules (Carlson et al., 2010; Keles et al., 2008; Puckett et al., 2007; Warren et al., 2006). The CSI specificity and energy landscapes can also be used to accurately annotate binding sites across the genome (genomescapes) and provide probabilities of binding to sequences that deviate from the consensus (Carlson et al., 2010).

### 2. Array-Based Cognate Sequence Identification

#### 2.1. Labeling DNA-binding molecules

To determine DNA-binding specificities of molecules, several high-throughput array-based methods have been developed (Berger and Bulyk, 2009; Liu et al., 2005; Ragoussis et al., 2006; Wang et al., 2005; Warren et al., 2006). One such technique is the CSI approach. CSI is an effective and unbiased method for determining the comprehensive DNA-binding specificities of small molecules or proteins. This technique involves the use of hairpin DNA, consisting of a variable region flanked by two constant regions: a hairpin loop and a complementary sequence. This design places the variable region in the stem of the hairpin. Previous studies confirm that this region of the stem adopts B-form structure under standard buffers that closely approximate physiological conditions (Kuznetsov et al., 2007). The sequences can be synthesized by photolithography on a silanized glass surface similar to standard fabrication procedures for DNA microarrays,
and these microarrays can be used to probe the sequence specificity of a DNA-binding molecule. The range of sequences that can be probed is only limited by the length of the DNA and the number of unique features that can be effectively synthesized by the microarray synthesis technology chosen. In other words, approximately half a million features bearing unique sequences are required to display the entire sequence space covered by 10 bp and each additional variable base pair requires approximately fourfold more features. The De Bruijn approach offers to extend the complexity without greatly expanding the number of features. However, this approach fails to capture the full spectrum of binding and it appears that differences between closely related sites are often lost.

A first step in a CSI experiment involves the design of the hairpin DNA probes to be synthesized on the silanized microarray surface. A commonly used design is 5'-GCGC-N₁N₂N₃N₄N₅N₆N₇N₈N₉N₁₀-GCGC-3', synthesized 3'-5', where N can be any of the four common bases (A, T, C, G), the GNA trinucleotide forms the hairpin loop, and the GCGC sequences form stabilizing constant regions flanking the variable region of the hairpin (Fig. 1.1). The 5'-GNA-3' sequence is ideal for loop formation in DNA hairpins (Hirao et al., 1993).

Once the microarray is designed and synthesized, the sequences are induced to form hairpins before performing a CSI experiment:

1. Incubate the array in 7 M urea for 30 min at 65 °C.
2. Transfer the array to a 1× phosphate-buffered saline (PBS) solution at 65 °C for 15 min to induce the duplex formation.

Next, the surface of the array is blocked or pacified to prevent adventitious binding to the surface from interfering with the detection of the sequence-specific DNA-binding events. This step may not be necessary in all cases, but is especially important if a given DNA-binding molecule has a propensity to bind the glass surface. Blocking the array is performed as follows:

1. Solubilize powdered milk to a final concentration of 2–3% (20–30 mg/mL) in sterile H₂O.
2. Apply the chamber to the microarray (i.e., Grace Bio-Labs, Item No. 623503).
3. Rinse the chamber twice with distilled H₂O, allowing the surface wetting during the second wash to proceed for 1–2 min.
4. Remove all liquid from the chamber and inject a sufficient volume of the milk solution into the chamber such that approximately three-fourth of the chamber is filled. The remaining one-fourth volume is occupied by an air bubble that rotates as the slide is rotated on a Southern blot hybridization wheel or another similar device. This constant rotation of the air bubble within the chamber mixes the solution evenly across the array.
Figure 1.1  The format of cognate sequence identification (CSI) microarrays. Self-complementary, single-stranded DNA is synthesized 3′–5′ on a glass slide. Hairpin formation is then induced, which leads to B-form DNA formation. Within each hairpin is a variable region (the Nmer) covering the entire sequence space up to 10 positional variants within a larger DNA duplex. Each “feature” of the microarray contains approximately 100,000 hairpins of a specific sequence (up to 2–6 million different sequences can be displayed currently). Once the DNA-binding protein or small molecule, or a combination of the two, is incubated on the microarray, fluorescence intensity can be used to quantitate the sequence affinity and specificity of the molecule.
(5) Once the samples are injected, seal the chamber as required by the manufacturer (typically done with a removable sticker) and spin the array on a rotor wheel for 1.5 h at 25 °C.

(6) Remove all of the liquid and repeat the rinsing from step (3).

Once the array is blocked, it can be incubated with the protein or small molecule of interest. These experiments can be performed using a fluorescently labeled DNA-binding protein or small molecule, or through indirect detection as discussed below.

2.1.1. Synthetic DNA-binding molecules

Of the many DNA-binding small molecules, polyamides are particularly important, as they can be specifically designed to target a variety of sequences. Polyamides bind DNA in a modular manner through various combinations of N-methylpyrrole and N-methylimidazole rings, even when the DNA is wrapped around nucleosomes (Dervan and Bürl, 1999; Dervan and Edelson, 2003; Suto et al., 2003; White et al., 1998). Hairpin polyamides have been used to regulate targets involved in many diseases (Burnett et al., 2006; Nickols et al., 2007; Olenyuk et al., 2004). Engineering these molecules to bind specific genomic sites requires a detailed understanding of the molecular recognition rules. To evaluate the DNA-binding specificity of polyamides and other engineered molecules, we developed the CSI approach on a microarray platform. The specific protocol is described below.

Buffer and hybridization mix preparation:

(1) Prepare wash buffer (1.22 M MES + 0.89 M [Na+] at pH 6.7–7.2), making sure that the MES stays covered at all times to prevent exposure to light.

(2) Thaw the control oligonucleotide that hybridizes to the control sequences in specific locations on the array. These control features permit data extraction after the hybridization.

(3) Prepare the small molecule/oligo/buffer mix with 5 nM control oligonucleotide and an amount of the DNA-binding molecule consistent with its binding affinity in solution.

After the reagents are prepared and the array is ready, the CSI array experiment can begin:

(1) Clean the inside of the chamber + array setup twice with wash buffer. Pipette the solution in and out of the chamber gently several times, allowing the second wash to wet the array surface for several minutes.

(2) Remove the wash buffer from the chamber and then fill the chamber completely with the small molecule/oligonucleotide/buffer mix.

(3) Remove approximately 20% of the solution to create one air bubble within the chamber for mixing.
Seal the chamber as before, attach the microarray to the rotor wheel, and cover it with aluminum foil to protect it from light exposure.

Begin rotating the array and observe the air bubble to ensure it moves smoothly around the array surface to prevent the array from drying out.

The final stage of the process before data analysis consists of washing, drying, and scanning the array:

1. Prepare a 1× final wash buffer (FWB) solution in a 50-mL falcon tube from the 10× stock (NimbleGen Systems Inc., Cat. No. R10019-2).

2. Unseal the chamber by removing the adhesive sticker from the holes, and remove all of the liquid. Then wash the array twice with wash buffer, allowing the buffer to sit in the chamber for a few minutes the second time.

3. Remove the chamber from the microarray surface and immerse the slide in the FWB solution for 30 s.

4. Remove the slide and dry using a microarray centrifuge.

5. Scan the array using a standard microarray scanner (≥ 5 μm resolution) with filters at wavelengths that permit the detection of fluorescent markers (i.e., 532 and 635 nm for Cy3 and Cy5 dyes, respectively). For our analysis, we use the Axon GenePix 4000B Microarray Scanner and the GenePix Pro software to scan the arrays.

6. Extract the array feature intensity data using compatible software, such as the NimbleScan software from Roche-NimbleGen.

2.1.2. Natural or engineered DNA-binding proteins

Similar to small molecule analysis described above, the comprehensive sequence-specificity profiles of proteins can also be determined using high-throughput DNA arrays. Unlike small molecules, each protein requires its own optimal binding conditions, depending on its biochemical properties. These binding conditions should be determined prior to performing the CSI array experiment. One should start with buffer conditions in which the protein of interest will bind to DNA in other experiments such as electrophoretic mobility shift assays (EMSAs), nuclease protection (Footprinting), or fluorescence polarization (FP) (Anderson et al., 2008; Fried and Crothers, 1981; Galas and Schmitz, 1978; Garner and Revzin, 1981; Heyduk and Heyduk, 2002; Heyduk et al., 1996). The binding reaction will contain buffer, salts, small molecules, and other cofactors that are required for the protein to bind DNA.

Additionally, the detection method will need to be determined. Proteins can be detected on the surface of the arrays by a variety of different methods. The protein can be conjugated directly to a fluorescent moiety using common commercially available compounds. If the direct labeling method is employed, the protein should be minimally labeled to limit the effect the
label has on its DNA recognition or functional conformation. Thus, proteins that are directly conjugated to a fluorophore require verification that the label does not adversely affect the specificity profiles of the conjugated protein. A less invasive approach for detecting the DNA-binding preferences of a protein involves using a fluorescently labeled antibody that is specific to the protein. In this indirect labeling approach, the antibody specifically recognizes the protein of interest or a peptide tag engineered onto the protein (such as 6–10 histidine residues or Flag, Myc, or GST tags). It is therefore the detection of the fluorescently labeled antibody that identifies the DNA sequences that are bound by the protein of interest. If antibody detection is to be used, it is important to ensure that the antibody does not interfere with the DNA-binding function of the protein of interest.

Here, we describe a protein CSI array experiment as a sample for adaptation to other proteins and conditions. In this example, we measure the binding profile of an unlabeled His-tagged fusion protein, with a binding reaction composed of a buffer (50 mM NaCl, 10 mM Tris, pH 7.4, 1 mM MgCl₂, 0.5 mM EDTA), the DNA-binding protein (one may use between 1 and 10 nM as a starting point depending on the affinity of the protein to DNA), 0.25% milk for protein stabilization, and a 1:500 dilution of a Cy5-labeled anti-His antibody.

Some things to remember when optimizing array–protein interactions:

- The protein concentration that will be optimal on the array (i.e., give the best binding profile with highest specific DNA binding and lowest non-specific or surface binding) will need to be determined for each protein.
- Additives can greatly increase the likelihood of obtaining a good binding profile. Such additives may include bovine serum albumin (BSA) as a molecular crowding agent, milk as a nonspecific competitor, small amounts of detergent (0.002%) to prevent protein aggregation, or dithiothreitol (DTT) to maintain a reducing environment.
- The stability of the protein of interest may be optimized at different temperatures. Therefore, the experiments must be performed at a temperature that is optimal for protein–DNA binding.
- If chemically prelabeling the protein of interest with a fluorophore, the added fluorophore must not interfere with the DNA recognition properties of the protein.
- If detecting the protein–DNA interaction using a fluorescently labeled antibody (such as a Cy5-labeled anti-His antibody to detect a 6-histidine-tagged fusion protein, or a Cy3-labeled anti-Myc antibody to detect a Myc-tagged fusion protein), the antibody must not interfere with the DNA recognition properties of the protein.
- Some commonly used fusion proteins are 6 histidine, GST, Myc, Flag, etcetera. Interestingly, we have found mixed results using a fusion protein of maltose binding protein (MBP).
After blocking of the array with milk, the protein is hybridized to the DNA microarray under optimal binding conditions (in a process similar to the small molecule procedure above). In brief, the steps are:

1. Clean the inside of the chamber + array setup twice with the prede
termined buffer by pipetting the solution in and out of the chamber gently several times, allowing the second wash to wet the array surface for several minutes.
2. Remove the buffer from the chamber and then fill the chamber completely with the protein/buffer mix.
3. Remove approximately 20% of the solution to create one air bubble within the chamber for mixing.
4. Seal the chamber as before, attach the microarray to the rotor wheel, and cover it with aluminum foil to protect it from light exposure.
5. Begin rotating the array and observe the air bubble to ensure it moves smoothly around the array surface so that none of the array dries out.
6. Store any remaining buffer until the run is completed (typically 1–2 h).

The final stage of the process before data analysis consists of washing, drying, and scanning the array:

1. Prepare a 1× FWB solution in a 50-mL falcon tube from the 10× stock (NimbleGen Systems Inc., Cat. No. R10019-2).
2. Unseal the chamber and remove all of the liquid, then wash twice with buffer, allowing the buffer to remain in the chamber for a few minutes the second time.
3. Remove the chamber from the microarray surface and immerse the slide in the FWB solution for 30 s.
4. Remove the slide and dry using a microarray centrifuge.
5. Scan the array and extract the data as mentioned in the previous small-molecule CSI protocol section (steps 5 and 6).

2.2. Label-free detection

In cases where a DNA-binding molecule cannot be fluorescently labeled, or in situations where this labeling perturbs its DNA-binding characteristics, an alternative application of CSI may be used. This method combines the use of FID with CSI microarrays. FID examines the binding affinity and specificity of small molecules to DNA through the displacement of a small fluorescent DNA intercalator, such as ethidium bromide. This has been used previously as a plate-based assay (Boger et al., 2001), where a small library of DNA hairpin oligonucleotides in individual wells is initially bound by the intercalator and the fluorescence for each sequence is measured. Then, the DNA-binding molecule of interest is applied to the wells and the binding displaces the intercalator from the DNA. Thus, binding is measured through
the decrease in fluorescence intensity within each well. One major limitation of the FID assay itself is the library complexity that can be achieved in a plate-based system. Thus, the hybrid CSI–FID approach was developed to overcome the library size limitations, allowing for a more complex library—one with a diversity of sequences as much as three to four orders of magnitude larger (Fig. 1.2; Hauschild et al., 2009).

Similar to the CSI array, this assay uses DNA microarrays having DNA hairpin oligonucleotides synthesized on the surface of the array. These arrays also require hairpin formation prior to the first use:

1. Incubate the array in 7 M urea for 30 min at 65 °C.
2. Transfer the array to a 1× PBS solution at 65 °C for 15 min to induce the duplex formation.

After hairpin formation, the DNA intercalator can be applied to the array. This procedure is performed on two arrays: one as the control for baseline EtBr binding and the other for the intercalation displacement experiment. The control array is necessary to prevent disruption of the intercalation equilibrium, which would occur if only one array were used and if the array were dried and scanned prior to incubation with the DNA-binding molecule:

1. Seal the arrays with the appropriate chamber (as mentioned above), add a solution of 6 μM EtBr in binding buffer (100 mM NaCl + 100 mM Tris, pH 8.0), and then seal the chambers.
2. Incubate the arrays at room temperature for 1 h while on the rotor wheel.
3. Apply the DNA-binding molecule of interest to one of the arrays (concentration to be determined empirically) and incubate at optimal binding temperature while rotating.
4. Wash both arrays twice with binding buffer and then dry and scan as described above for CSI experiments. EtBr can be detected using the same array scanner filter as used for the Cy3 dye.

Alternative methods that utilize high-throughput SPR as well as other highly sensitive detection technologies are being developed by several laboratories. In the future, such alternate detection strategies will obviate the need to label DNA-binding molecules directly or indirectly.

3. SOLUTION-BASED COGNATE SEQUENCE IDENTIFICATION

The application of high-throughput sequencing technology allows for the expansion of the sequence space of the library beyond the restrictions imposed by microarray-based assays. In addition to increasing the library
Figure 1.2  CSI–FID provides an important label-free tool for DNA-binding affinity and specificity determination. (A) Left column displays a standard CSI microarray experiment (as described in Fig. 1.1) showing the application of a fluorescently tagged DNA-binding molecule to a microarray containing approximately 2 million double-stranded
complexity, new sequencing technologies, such as that used by the Illumina Genome Analyzer, can sequence many samples in parallel. This is accomplished through the use of unique barcode sequences, which are included in the design of the sequences in the library being examined. Multiple binding experiments can be performed in parallel, each with a unique barcode, and the resulting samples can be pooled and sequenced simultaneously (Jolma et al., 2010; Stormo and Zhao, 2010; Zykovich et al., 2009).

As with the microarray-based methods, CSI by sequencing begins with the design of the DNA library to be synthesized. The library design is tailored to the sequencing technology that will be used in the experiment. For Illumina sequencing, the DNA must have a constant primer at the 5' and 3' ends of each piece of DNA. In addition, unique barcodes should be designed to identify the sample being sequenced. These barcodes can be as short or long as required to uniquely identify each sample. It is recommended that these barcodes vary by more than a single nucleotide from one another in order to prevent saturation of the detectors during sequencing. Finally, the central region of the DNA sequence should contain the randomized sequences to be bound by the protein or small molecule of interest. These randomized regions are no longer limited to the 10–12mer sequence space used in microarray-based assays. As many as 22 randomized nucleotides have been used to date (Zykovich et al., 2009). However, a balance must be established between the synthetic possibilities and the ability to perform binding in the appropriate concentration regimes. The greater the complexity of the library, the greater the challenges of retaining sufficient molecules of each sequence at high enough concentrations to permit equilibrium binding with the DNA-binding molecules of interest. A particularly thoughtful example is the use of a 14mer library, rather than a larger library, to interrogate the binding by several eukaryotic transcription factors (Jolma, et al. 2010).

Following the design and synthesis of the DNA libraries is the binding experiment itself. This can be performed with limited rounds of selection and enrichment to avoid loss of moderate-to-low-affinity binding sequences (i.e., SELEX; Jolma et al., 2010; Zhao et al., 2009) or with a single binding event followed by purification on a column or via EMSA.
Once the DNA–protein or DNA–small molecule complexes are isolated, the DNA can be released and purified for sequencing. One caveat to consider is that it may be necessary at this point to preamplify the bound DNA sequences to have enough material for the sequencing reaction. If this becomes necessary, one must take great care in selecting a method of amplification which will cause the least bias in the resulting DNA samples. It may be possible to avoid amplification if there are a sufficient number of samples being pooled to yield the required amount of DNA for sequencing.

The above protocol can be executed as follows (Fig. 1.3):

1. Design the library to be used for each experiment and the unique barcode used to identify the sample. As an example, we use a library synthesized by Integrated DNA Technologies (IDT) containing a 15mer randomized library adjacent to a 6-bp barcode and flanked by two adapter sequences required by the Illumina sequencing technology.

2. Generate dsDNA from the ssDNA library: mix 1 μM of the ssDNA library with 10 μM of the reverse primer in EconoTaq Plus 2× master mix (Lucigen, Cat. No. 30035-1). Run the PCR reaction as follows:
   (a) 94 °C for 2 min 30 s.
   (b) 55 °C for 2 min.
   (c) 72 °C for 30 min.

3. Perform the binding reaction by incubating the protein of interest with the DNA library: combine 50 mM biotin-labeled protein, 0.01 μM dsDNA library, 50 ng/μL polydI·dC in a total volume of 25 μL and incubate at optimal binding conditions for 2 h.

4. Pull down biotin–labeled protein–bound DNA using Dyna beads M280 (Invitrogen, Cat. No. 112-05D) according to manufacturer’s protocol.

5. Wash the bead-bound complexes six times with 1× PBS + 0.1% BSA solution (100-fold excess buffer).

6. Transfer the beads to PCR strip tubes and add 20 μM forward and reverse primers with a standard PCR reaction mix to each tube and run the following thermal cycle:
   (a) 94 °C for 2 min.
   (b) 94 °C for 30 s.
   (c) 55 °C for 30 s.
   (d) 72 °C for 30 s.
   (e) Repeat steps (b) through (d) 19 more times for a total of 20 cycles.

7. Place the samples on a magnetic block and manually remove the supernatant to separate the samples from the beads.

8. Purify the amplified DNA from the samples using the QIAquick PCR Purification Kit (QIAGEN, Cat. No. 28104).
Repeat steps (3) through (8) two more times to select for optimally bound sequences and reduce background binding noise.

Submit the samples for Illumina sequencing.

The binding data obtained by microarrays or by high-throughput sequencing can be analyzed in a similar manner to identify the consensus binding motif as well as the full spectrum of binding affinities of a natural or a synthetic molecule. These high-content data with multidimensional relationships require new computational tools to visualize and comprehend in their entirety. Below, we describe our approach to this challenge.
4. Data Analysis and Visualization of Specificity, Binding Energy, and Genomic-Association Landscapes

4.1. Sequence-specificity landscapes

At present, the sequence preferences of DNA-binding proteins are described by consensus motifs. Typically, a limited set of sequences with high affinity for a particular protein or small molecule are aligned to identify a consensus motif and the frequency of occurrence of a given nucleotide at a specific position is used as a measure of its contribution to the binding event. Such approaches yield a position matrix and several variations have been developed to account for the contributions of adjacent and nonadjacent nucleotides to binding. Collectively, such position weight matrices (PWMs) or position-specific scoring matrices (PSSMs) are then displayed using an intuitive Logo display that scales the size of a nucleotide to the information content at each position (Berg and von Hippel, 1987; Schneider and Stephens, 1990; Stormo and Zhao, 2010; Von Hippel and Berg, 1986).

The highly textured binding data across the entire sequence space from CSI and related methods are not adequately described using a simple PWM (Carlson et al., 2010; Keles et al., 2008; Maerkl and Quake, 2007). PWMs can only display limited information about the binding specificities of DNA-binding molecules. Often, PWMs do not capture the effects of neighboring nucleotides on a given nucleotide’s binding probability because they are typically generated with limited binding data and under the assumption that each nucleotide contributes independently to the binding of a protein to its cognate site (Benos et al., 2002). Moreover, the underlying focus on indentifying a “consensus motif” leads to amalgamation of several related motifs into a single consensus. Unrelated motifs that do not fit the consensus are often ignored or obfuscated by the consensus, leading to an inaccurate representation of the binding preferences of a given molecule (Carlson et al., 2010). Finally, the current motif search algorithms do not consider the impact of nonconserved sequences that flank a “consensus” motif on the overall binding affinity for an otherwise optimal site. While PWM/PSSMs are very useful and often suffice in describing the binding preferences of a DNA-binding molecule, they fail to accurately describe binding to suboptimal sites that are often used to regulate biological outputs. Recent efforts to consider di- or trinucleotides in developing PWMs, or to search for more than one motif in the binding data, have resolved some of the limitations of past motif searching algorithms (Stormo and Zhao, 2010). Nevertheless, these matrices and their Logo displays are inherently limited and, in the context of the comprehensive binding data obtained from CSI and related methods, it is important to view the entire
specificity landscape to determine the range of activity of a natural or engineered DNA-binding molecule.

Recently, a novel 3D method of displaying this type of data has been developed. SSLs can capture the full binding landscape of an Nmer sequence space for a given binding motif (Carlson et al., 2010). An SSL is generated from a seed motif and represents the data in concentric rings with the z-axis displaying the relative binding affinity—either as the fluorescence intensity on the array or as the number of reads of a given sequence from high-throughput sequencing methods. The center ring contains the data for all the sequences bound which contain the seed motif exactly. Each successive ring moving outward has an increasing number of mismatches to the seed motif (Fig. 1.4). The data within the “zero mismatch” ring are sorted alphabetically by the sequences flanking the seed motif. The mismatch rings are sorted first by the position of the mismatch within the motif, then alphabetically by the mismatched base(s), and finally sorted alphabetically by the sequences flanking the seed motif. Each ring is aligned in the clockwise direction starting with the first (5') residue of the motif. The impact of changing any single residue becomes apparent in quadrants that represent variations at specific positions. Thus, all changes at the first position of the motif are aligned at the beginning of each mismatch ring. Multidimensional relationships between several positions can be readily visualized in specific quadrants along the rings. In the example, the importance of changing the first two residues of the Nkx2.5 binding site is readily visualized as peaks in circles that are one or two

Figure 1.4  Sequence-specificity landscapes (SSLs). SSL generated from CSI intensity data for the Nkx-2.5 protein. The best bound sequences were used to generate the seed motif (shown in the LOGO and above the landscape). Color scale indicates normalized fluorescence intensity corresponding to how well a given sequence is bound. The sequences are distributed among the rings as described in the text. The distance between each ring and the next concentric ring outward is equivalent to a Hamming distance of 1. (Adapted from Carlson et al., 2010).
mismatches (Hamming distances) from the seed motif (Fig. 1.4). The data also reveal that the PWM compresses two motifs—TNAAGTG and NTAAGTG into a single consensus wherein the first two T residues are considered to have low information content. This is not the case, as an arginine on an unstructured arm of Nkx2.5 interacts with one T residue either at position 1 or at position 2 and this interaction contributes to high-affinity binding (Carlson et al., 2010; Keles et al., 2008). Similarly, low-to-moderate affinity binding sites with no sequence similarity to the accepted consensus can be deconvoluted by SSL display of the entire data. Using the SSL tool, we found that nearly 40% of the current motifs reported appear to be a forced compression of related motifs (Carlson et al., 2010).

To emphasize the importance of viewing the entire SSL, rather than focusing on the PWM or using the PWM to predict binding properties, we display the CSI data for GATA4 (Fig. 1.5). GATA4 is a nuclear receptor type of zinc-finger transcription factor. It regulates genes involved in embryogenesis and in myocardial differentiation and function, and malfunction of GATA4 has been shown to cause septal defects. Comprehensive GATA4 sequence-specificity profiles were obtained via CSI platform bearing the entire 8mer sequence space. The top 300 binding sequences yielded a PWM and a consensus motif of 5′-GATA-3′. This consensus motif agrees

![Figure 1.5 SSLs for GATA4](image)

Figure 1.5 SSLs for GATA4. Panel (A) displays the specificity landscape predicted from the position weight matrix (PWM) generated from the CSI data. The PWM-based landscape predicts uniform binding wherever the exact GATA sequence is found. Panel (B) shows the actual intensity of every sequence on the array that contains the GATA motif. The data clearly indicate that unlike the prediction from the PWM, the binding is not uniform in every instance where this sequence is found. Panel (C) displays the SSL obtained after iterative refinement of the motif permitted by landscapes. Inspection of binding data in this manner reveals that many of the motifs of DNA-binding proteins and molecules compress the data and do not properly account for the influences of flanking sequences. The color scale reflects the fluorescence intensity of each sequence on the z-axis and the seed motif is shown above each landscape.
with the “known” binding motif for this protein. As expected, when the underlying PWM is used to predict protein binding to sequences that vary from the consensus, we see little potential binding to mismatch sites (Fig. 1.5A). Moreover, the PWM-based landscape would predict equivalent binding to the consensus motif irrespective of its flanking sequence context (Fig. 1.5A). On the contrary, when the entire binding data (32,896 sequences) are displayed on a specificity landscape, several unanticipated properties emerge (Fig. 1.5B). First, the central “zero mismatch” ring containing all DNA hairpins with the perfect consensus core motif does not show equivalent binding. Second, the SSL for this consensus motif has very obvious regions of suboptimal binding within the zero mismatch ring, indicating that this motif is not restrictive enough to exclude poorly bound sequences. Third, the outer rings show some low-affinity binding that may have biological relevance. The actual CSI data displayed as a specificity landscape highlights the limitations of the PWMs and Logo displays in describing the full spectrum of molecular recognition properties of a DNA-binding molecule.

The specificity landscapes overcome these limitations and also reveal the significant impact of flanking sequences on the binding potential. It is important to note that in organizing the specificity landscape, we used the consensus motif obtained from the PWM. However, regardless of the source of the seed motif, SSLs can iteratively refine the motif for a given DNA-binding molecule. The quality of the seed motif can be assessed by inspecting where the highest binding regions are in the landscape and adjusting the sequence accordingly to bring all the best binding sequences into the center ring and move poorly bound sequences out. Using this iterative application of SSLs, the optimal binding motif can be refined from the original seed motif. In the case of GATA4, such iterations reveal a larger motif of 5′-WGATAA-3′ as the optimal binding motif (Fig. 1.5C). The data also reveal the contribution of nonconserved flanking sequences on the differential binding even if the optimal motif is present. In the earlier example of Nkx2.5, SSLs revealed the existence of closely related motifs (TNAAGTG and NTAAGTG) with high-information content residues (T at position 1 or 2) that were compressed by standard motif searching algorithms that seek to find a consensus site. In fact, as mentioned earlier, SSL evaluation of the current high-throughput binding data revealed that approximately 40% of reported motifs have two or more related motifs amalgamated into a single consensus (Carlson et al., 2010).

To facilitate the evaluation of binding data by SSL, we have developed a user-friendly software package. The software used to generate these SSLs will be available online at http://www.biochem.wisc.edu/faculty/ansari/ and it can generate SSLs, assist with seed motif identification, and even overlay CSI data on genomic DNA (see genomescape description below; Fig. 1.6). This software requires the user to input the data file containing a
forward sequence, the reverse compliment of that sequence, and the data value (intensity, Z-score, sequence read count) for that sequence in a tab-delimited text file. In addition, the program requires a seed motif in order to generate the SSLs. There are a number of ways to generate such a motif if the information is not already available from biochemical assays (EMSAs, SELEX, etc.). One such method is provided by the SSL software and uses “mean sorting” to find the optimal motif (1b in Fig. 1.6). This method calculates the mean and standard deviation of all possible sequences of a given length from CSI data and sorts them to show those with the highest mean intensity first. These motifs are then used as seed motifs and are iteratively refined into the optimally bound sequence motif.

Another method for selecting a seed motif involves the use of Multiple EM for Motif Elicitation (MEME) software, which is freely available online at http://meme.nbcr.net/meme4_4_0/intro.html. This program searches for motifs among a set of related DNA or protein sequences and represents the output with Logos and PWMs (1a in Fig. 1.6). The program requires a FASTA formatted list of sequences, weights for the sequences (if applicable), and input parameters such as number of motifs to find and the size of the motifs. One limitation to this method is the limitation on the number and length of sequences that can be included in the input. The program cannot
handle all of the sequences from a CSI experiment, and the typical input is the top 300 sequences bound. However, this program can still provide a reasonable starting seed motif from which one can refine an optimal motif through iterative use of SSLs.

Both microarray and sequencing data, bearing millions of different sequences, can be used to obtain a comprehensive view of the molecular recognition properties of DNA-binding molecules. The multidimensional relationships between individual residues within the motif can be intuitively comprehended and used to evaluate the specificity determinants of natural or engineered binding molecules.

4.2. Binding energy landscapes

The specificity landscapes also serve as comprehensive binding energy landscapes. The fluorescence intensities obtained from microarrays are often subject to the caveat that binding at the surface may not reflect true equilibrium due to mass transport problems at surfaces and perturbation of equilibrium during the washing steps. To address this challenge, microfluidic devices, high-throughput SPR, and even fluorescence displacement approaches are being actively developed, but they are still limited to 100- to 1000-fold fewer sequences than could be displayed on a microarray.

Based on systematic solution binding measurements, we find that the fluorescence intensity observed when labeled molecules bind a DNA sequence on the microarray is directly proportional to its binding affinity in solution (Fig. 1.7; Carlson et al., 2010; Puckett et al., 2007; Warren et al., 2006). CSI analysis of a synthetic polyamide and Nkx-2.5, a homeodomain protein, was performed via microarrays. Based on the resulting SSLs, specific sequences were chosen that span the range of fluorescence intensities (shown by the color scale). These intensities were compared to affinities measured by two different solution-based equilibrium binding methods. The binding affinities of the polyamide to these selected sequences were determined via DNaseI nuclease protection, whereas EMSA was used to measure the solution binding of Nkx-2.5. In both cases, we observed strong correlation between binding on the array and binding in solution ($R^2 > 0.99$). Therefore, the binding affinity is directly measurable from the range of fluorescence intensities on the microarray itself. This platform allows one to capture binding affinities of a given DNA-binding molecule or complex across millions of DNA sequences simultaneously. Thus, an SSL also provides a comprehensive binding energy landscape for the entire sequence space. This permits an unprecedented understanding of the energy and specificity thresholds that define the function of natural DNA-binding molecules in living cells. It also provides a framework to engineer artificial molecules that have desired biological potential.
Figure 1.7  Sequence-specificity landscapes also function as comprehensive binding energy landscapes. (A) Polyamide PA-1 SSL as determined by CSI. Six sequences with intensities spanning a broad range were chosen (labeled A–F) and nuclease protection was performed to examine the correlation between $K_A$ and CSI intensities. The color scale (center) shows the relationship of SSL–CSI intensities to their corresponding $\Delta G$ values ($\Delta G = -RT \ln K_A$). (B) Nkx-2.5 SSL as determined by CSI microarray. Six sequences were also chosen from the Nkx-2.5 CSI data (labeled G–L) and tested using EMSA to measure their $K_A$. The linear relationship between CSI intensity and DNA-binding characteristics measured in solution is readily detected (adapted from Carlson et al., 2010).
4.3. Genomescapes

Another powerful feature of the SSL software is the ability to overlay CSI data on genomic DNA sequences (Fig. 1.8). The CSI data capture not only the affinity of a given motif but also the contribution of the flanking (contextual) sequences. This entire sequence, core along with flanking residues, can be used to annotate the genome. Thus, a highly textured annotation of potential binding sites across the genome can be readily obtained. This comprehensive and unbiased “genomescape” further reduces the dependence on PWMs to evaluate genomes and identify potential binding sites. One must be cognizant of the fact that the occurrence of a binding site is not sufficient to assume binding of factors to those sites in living cells. This is in part due to the presence of chromatin and multiple epigenetic marks that are superimposed on the underlying sequence. Moreover, it is quite possible that proteins might associate with specific sites in the genome via interactions with other DNA-binding proteins, thus functioning in the absence of discernable cognate sites. These additional layers can be revealed by comparing the CSI-determined genomescape with the newly developed methods for determining transcription factor binding patterns in cells. Thus, when combined with chromatin immunoprecipitation on microarrays (ChIP-chip), sequencing (CSI-seq), or gene expression data, the “genomescape” can reveal new mechanism of regulation (Iyer et al., 2001; Johnson et al., 2007; Ren et al., 2000).

The genomescapes are represented as bar graphs across a given stretch of genomic DNA and can be useful in identifying possible in vivo binding sites for a given DNA-binding molecule. The software is capable of generating these genomescapes when provided with the SSL input data file (mentioned above) and a FASTA format text file containing the genomic sequence of interest (up to, and including the full length of a chromosome). The user can even provide boundaries for a region of the chromosome that was provided and expand that region for closer study (Fig. 1.8).

As an example, CSI binding data were obtained for an engineered DNA-binding small molecule (polyamide) and used to generate a genomescape across chromosome 6 of the human genome (Carlson et al., 2010; Puckett et al., 2007). This particular polyamide was designed to target and inhibit the binding of the Hif1α transcription factor to its cognate sites (Olenyuk et al., 2004). HIF1α is known to stimulate the expression of VEGF, a protein closely associated with angiogenesis and cancer (Kim and Kaelin, 2006; Underiner et al., 2004). The engineered polyamide targets the Hif1α binding sites at VEGF and other genes, prevents HIF1 from binding and activating transcription, and counteracts its action in oncogenesis (Kageyama et al., 2006; Nickols et al., 2007; Olenyuk et al., 2004). While the polyamide achieves its function at VEGF, it regulates Hif1α function differently at sites that have similar cognate sites (Olenyuk et al., 2004).
Figure 1.8 Genomescapes from CSI binding data. Schematic showing the parallel workflow of determining DNA-binding profiles in vitro (via CSI), and in live cells via genome-wide expression profiling, for a polyamide. The purple diamond represents a β-alanine, the orange square represents a chlorothiophene, the open circle represents an N-methylpyrrole, and the filled circle represents an N-methylimidazole. In genomescapes, the CSI data are used to annotate the genome with potential binding sites for a given DNA-binding molecule. This analysis, when combined with genome-wide expression or binding data, can reveal possible in vivo regulatory sites—as reflected by the transcription repression or activation seen in the expression profiling (adapted from Carlson et al., 2010).
This difference cannot be simply explained by differential affinity to different Hif1α binding sites. The genomescapes revealed multiple moderate-to-low binding sites for the polyamide at the endothelin 2 (ET-2) gene, where the synthetic molecule was particularly effective in inhibiting gene expression (Fig. 1.8). Thus, genomescapes reveal the cumulative effect of a cluster of weak binding sites that overlap with functional elements in the genome. The application of genomescapes, alongside ChIP-seq or transcriptomics analyses, will provide a far more mechanistic view of gene regulatory patterns.

CSI is a powerful approach for measuring the specificity profile of any natural or synthetic DNA-binding molecule, and is not limited to just proteins. Complexes between proteins and engineered small molecules have also been examined by CSI to reveal the changes in sequence specificity of individual molecules due to cooperative assembly on DNA (Warren et al., 2006). The ability to simultaneously interrogate the entire sequence space of a typical DNA-binding site yields a comprehensive view of the specificity spectrum of a DNA-binding molecule, while also uncovering the context effects of neighboring bases on binding affinity. When displayed as SSLs, the data reveal, in unprecedented detail, the different modes of sequence cognition and overcome the limitations of current motif searching algorithms that amalgamate several cognate sites into a single consensus motif. Moreover, the direct relationship observed between binding on microarrays and binding in solution allows SSLs to serve as binding energy landscapes. These specificity-energy landscapes help define the molecular recognition properties of DNA-binding molecules and facilitate the design of synthetic molecules to target genomic sites. Finally, overlaying the CSI data across genomic sequence gives a more accurate annotation of potential binding sites across the genome. Such CSI genomescapes not only complement the current approaches to genome-wide expression and binding studies but also facilitate the precision-tailoring of artificial molecules to target and regulate desired genes and networks. The applications of this methodology range from biophysical explorations of the principles that underlie specificity of molecular interactions, to evolutionary drift of regulatory motifs, to functional genomics, to synthetic regulation of gene networks and cellular fate.

REFERENCES


