

Automated Selection of Transcription Factor Binding Sites

Letha J. Sooter and Andrew D. Ellington*

Department of Chemistry and Biochemistry, Institute for Cell and Molecular Biology, University of Texas at Austin, Austin, TX

Keywords:

SELEX, in vitro selection, aptamer, transcription factor, NFκB

Double-stranded DNA binding sites that bound with high affinity to the nuclear factor kappa B (NFκB) p50 homodimer were selected using a Tecan Genesis workstation. The adaptation of the Tecan to automated selection required the integration of multiple devices and modifications to standard selection protocols, and resulted in a significant increase in throughput. The sequences obtained by automated selection strongly correlated with the well-known family of natural NFκB double-stranded DNA binding sites and with previous manual selection experiments. In addition, the selection experiments better defined the contributions of residues outside of the well-known, decameric core binding site for NFκB. (JALA 2004;9:277–84)

INTRODUCTION

Transcription factors are important regulatory components within a cell. In response to stimuli, they modulate gene expression both individually and in conjunction with other proteins. Insight into the complex regulatory networks governed by the plethora of transcription factors that are present in any given organism would provide important information about cellular functions. However, the

identification of transcription factor binding sites has generally required painstaking analysis of the sequences upstream of genes.

Automated selection methods can be patterned after similar efforts based on manual selection. Transcription factor binding sites have previously been isolated from random sequence libraries of double-stranded oligonucleotides. In short, the in vitro selection process begins with a synthetic oligonucleotide pool that has a random sequence core of from 20 to 100 residues, flanked by constant regions required for enzymatic manipulations. The initial pool contains from 10^{13} to 10^{15} different sequences, representing all possible sequences of length 22. The single-stranded DNA can be amplified via the polymerase chain reaction to form double-stranded libraries that are in turn incubated with a target protein. Binding species are partitioned from non-binding or weakly-binding species, eluted from the target protein, and amplified via the polymerase chain reaction (PCR). The amplified, double-stranded DNA product can be used for additional rounds of selection. Iterative rounds of selection and amplification result in the preferential enrichment of those binding species with the highest affinities for the protein target. Following selection, the delimited pools are cloned and individual binding sites sequenced. Typically, one or several families of sequences emerge after multiple rounds of selection.

Manual selection methods have previously yielded binding sequences for a variety of transcription factors, including NFκB,¹ estrogen receptor,² p53,³ myogenin,^{4,5} and CTF/NFI.⁶ Binding sites have been identified not only for purified proteins, but also for protein complexes.^{4,5} For example, selections against

*Correspondence: Andrew D. Ellington, Department of Chemistry and Biochemistry, Institute for Cell and Molecular Biology, The University of Texas at Austin, 1 University Station, A4800 Austin, TX 78712; Phone: +1.512.232.3424; Fax: +1.512.471.7014; E-mail: andy.ellington@mail.utexas.edu

1535-5535/\$30.00

Copyright © 2004 by The Association for Laboratory Automation
doi:10.1016/j.jala.2004.05.004

purified myogenin and myogenin in nuclear extracts each yielded the same sequence families. The selected protein binding sites corresponded to natural sequences known to bind to the myogenin homodimer.

Double-stranded DNA selections have been used not only to globally define DNA binding sites for proteins, but also to determine the relative contributions of individual residues within a given site to interactions with a target protein.⁶ For example, after four rounds of *in vitro* selection against the CTF/NFI transcription factor, Bucher and co-workers sequenced over 10,000 possible binding sites and constructed a binding model based on the sequence distribution. The binding model was used to predict natural binding sites, which were subsequently experimentally verified.

In general, the sequence data from *in vitro* selection experiments corresponds to identified, natural binding sites and hence can be matched to promoter regions within genomes and used to identify genes that may be regulated by a transcription factor. There are a number of programs and databases that are available to assist with transcription factor binding site identification. For example, TRANSFAC, The Transcription Factor Database, is a compilation of known transcription factor binding sites, and EPD, the Eukaryotic Promoter Database, is a collection of eukaryotic promoters for which the transcriptional start sites have been experimentally determined. There is also software, such as SiteSeer,⁷ that can interface directly with the TRANSFAC database and can aid in binding site identification. In addition, statistical weighting algorithms^{8,9} can assist in the identification of new sites or subtle alterations in specificity.

The identification of multiple transcription factor binding sites by a combination of selection, sequence analysis, and database mining can potentially lead to the construction of a full description of the regulatory pathways in a cell. Unfortunately, the *in vitro* selection process can be extremely time-consuming. To facilitate high-throughput binding site identification, we have attempted to automate the selection of transcription factor binding sites. The NFκB p50 homodimer was chosen as an initial target for the development of automated selection methods. This transcription factor is well-known to bind double-stranded DNA,¹⁰⁻¹² and has previously been a target for manual selection experiments.¹ In addition, a Tecan Genesis workstation was chosen for this project; the flexibility of the Tecan allowed the implementation of high-stringency panning and the separation of high-affinity binding sites from non-specific binding sequences.

MATERIALS AND METHODS

Liquid-Handling Robot

A Tecan Genesis workstation 200 was used as the platform for the automation of double-stranded DNA binding site selections. This robot has two pods, a liquid-handling (LiHa) pod and a robotic manipulator (RoMa) arm. The LiHa is composed of eight, independently

controlled pipetting tips that have liquid sensing capabilities and that can accurately pipette between 0.5 and 1000 μL. The RoMa arm can reach off the worksurface and was essential for integrating the auxiliary equipment necessary for automated selection.

The Tecan Genesis worksurface holds a number of items (all from Tecan, unless otherwise indicated), including a 12-position microplate carrier (MP-12), a solid-phase extraction unit (SPE) with an adapter for Qiagen (Valencia, CA) kits, a two position orbital shaker, a 4°C cooled microplate carrier with a recirculating temperature bath (Julabo, Allentown, PA), a -20°C cooled microplate carrier (Mecour, Groveland, MA) with a recirculating temperature bath (Neslab, Waltham, MA), disposable tips (DiTi's), and reservoirs for buffers, reagents, and other solutions. Items off the worksurface but accessible by the RoMa arm included a thermal cycler (MJ Research, Waltham, MA), Tecan microplate holders, and a Tecan 16-channel Columbus plate washer. The plate washer was essential to the success of the selection, and its operation is described here in some detail. The Columbus washes two columns of eight wells in parallel, 16 total wells. Two needles are inserted into each well: an aspiration needle and a dispense needle. A defined volume of liquid flows out of one of the four solution reservoirs and into the microplate well through the dispense needle. The liquid remains in the well for a defined time and is then removed by the aspiration needle. All liquid is finally deposited in a solution waste reservoir.

Oligonucleotides

The N30 pool contains 30 random nucleotides between a 5' constant region (5' GATAATACGACTCACTA-TAGGGAATGGATCCACATCTACGAATTC) and a 3' constant region (5' TTCACTGCAGACTTGACGAAGC-TT¹³). Following amplification, the double-stranded N30 pool (10¹³ molecules) was used in the first round of selection. A positive control for the double-stranded DNA selection was constructed by inserting a NFκB p50 homodimer binding sequence (5' TGACTGATTGGGGATTCCCGA-AGCTTATC¹) between the two constant regions.

Target Plate Preparation

Target plates were prepared by hydrophobic immobilization of NFκB p50 homodimer protein (0.3 μg per well; Sigma, St. Louis, MO) in wells in TopYield microtitre plates (Nunc, Rochester, NY). The NFκB was dissolved in 100 μL of 1× selection buffer (20 mM HEPES, pH 7.9, 100 mM KCl, 0.2mM EDTA, 5 mM DTT). The solution was added to wells, the wells were sealed, and the plates were incubated without agitation at 4°C for approximately 18 hours. Following incubation, the solution was removed and the wells were washed with a casein blocking solution (Pierce, Rockford, IL). Remaining hydrophobic sites were blocked by incubation with casein solution at 4°C for greater than

3 hours. Plates used for negative selections were prepared in an identical manner except that NF κ B was not added to the selection buffer.

Automated Selection

The selection process is diagrammed in Figure 1 and the details of the selection cycles are provided in Table 1. The negative selection plate and the target plate were placed on the MP-12 microplate carrier on the Tecan work surface. The casein block was removed from the negative plate and the plate was rinsed with 175 μ L selection buffer. The Round 0 double-stranded DNA pool (100 μ L; 1.5 μ g; 10^{13} molecules) was spiked with 10^9 molecules NF κ B p50 homodimer binding sequence positive control. The spiked pool was transferred from the 4°C cooled microplate carrier to the negative selection plate. The negative selection plate was transferred to the orbital shaker where it underwent four cycles of alternating incubations (3 minutes at 500 rpm and then five minutes stationary), and was then moved back to the MP-12. The casein block was removed from the target plate and it was transferred to the Columbus plate washer and sequentially washed with 1.5 mL selection buffer and 300 μ L dH $_2$ O. The spiked pool in the negative selection plate was transferred to the target plate, which was in turn transferred to the orbital shaker. After one to four cycles of alternating incubations (as described previously), the target plate was transferred to the Columbus plate washer. The microtitre plate wells were washed with seven or eleven wash cycles (10.5 or 16.5 mL total) of selection buffer, then 300 μ L dH $_2$ O. The target plate was moved back to the MP-12, and PCR master mix (100 μ L; 10mM Tris, pH 8.4; 50 mM KCl; 2.5 mM MgCl $_2$; 0.2 mM dNTPs; 0.4 μ M each of the 41.30 5' primer and the 24.30 3' primer) and 5U Taq polymerase were added. The target plate was transferred to the thermal cycler and 15 or 20 cycles of PCR amplification (denaturation for 10 minutes at 90°C, then cycled for 90 seconds at 90°C, 30 seconds at 60°C, and 90 seconds at 72°C; final extension for 3 minutes at 72°C) were carried out. During the thermal cycling procedure, the Columbus probes were cleaned with 6 mL of a 7 M urea solution followed by 6 mL of dH $_2$ O. Following DNA amplification, the plate was returned to the MP-12 and 15 μ L 3M sodium acetate (pH 5.2) was added to the well to lower the pH of the solution to pH 6–7. The PCR solution was then added to 345 μ L Qiagen Buffer PM in a 2 mL deepwell plate on the MP-12 worksurface. The contents were mixed and transferred to the Qiagen filter plate on the SPE. A 500 mbar vacuum was applied for 5 minutes to pull the solution through the filter. Then 900 μ L of Qiagen Buffer PE were added, followed again by application of a vacuum. The final wash was an addition of 900 μ L of Buffer PE. Following filtration the filter was dried (as required by the protocol), and 120 μ L of selection buffer was added to the well. For the collection of the DNA eluate, the RoMa arm transferred the SPE block to the second position on the manifold, and a vacuum of 500 mbar was applied for 5 minutes. The purified PCR product was ultimately eluted

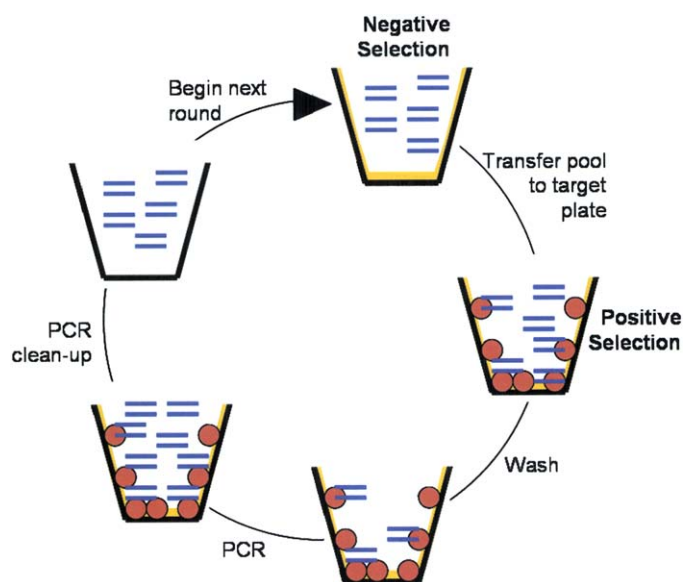


Figure 1. Automated panning protocol for in vitro, double-stranded DNA selections. This is a simple schematic of the protocol described in the “Materials and Methods” section. In short, first a negative selection of a double-stranded DNA library (dual lines) is carried out against a microtitre plate well containing only a casein block (light grey). Those DNA molecules that do not stick to the block or to the plate are then transferred to a different microtitre plate well containing target protein (grey circles). Loosely or non-specifically bound DNA species are washed away. PCR mix is added directly to the well and any remaining DNA molecules are amplified. The amplified products are purified and then added to a new negative selection well to begin the next cycle of selection and amplification.

into a Qiagen deepwell plate, and then the SPE block was transferred back to the first position on the manifold by the RoMa arm. The final 100 μ L of the DNA was then transferred from the SPE to a negative selection microtitre plate to begin the next round of selection and amplification. Following the negative selection, the pool was transferred to a new well coated with the target.

Sequencing

The double-stranded DNA pools from Rounds 0, 3, and 6 were cloned into TOPO TA vectors (Invitrogen, Carlsbad, CA) and transformed into Top 10 (Invitrogen) competent cells. Following transformation, cells were plated on Luria-Bertoni media (LB) plates supplemented with 50 μ g/mL kanamycin and 1600 μ g X-gal per plate. The plates were incubated at 37°C until small colonies were visible. White colonies were picked and used to inoculate 1.5 mL cultures of LB containing 50 μ g/mL ampicillin in a 2 mL 96-well deepwell plate (Corning, Acton, MA). The antibiotics used for growth were changed between plates and media in order to ensure that the transformants were derived from the original TOPO TA vector. Cell cultures were grown

Table 1. Selection conditions and stringency. In order to modulate the stringency of the selection through successive rounds, four different conditions were varied: length of incubation time for negative and positive selections, wash volumes, and the number of PCR cycles. Each of these variables is described in greater detail in the "Materials and Methods" section.

Round	Negative selection	Positive selection	Selection buffer wash (mL)	# PCR cycles
1	4 shaking/stationary incubation cycles	4 shaking/stationary incubation cycles	10.5	20
2	4 shaking/stationary incubation cycles	4 shaking/stationary incubation cycles	10.5	20
3	4 shaking/stationary incubation cycles + 1 hour stationary incubation	3 shaking/stationary incubation cycles	10.5	20
4	4 shaking/stationary incubation cycles + 1 hour stationary incubation	2 shaking/stationary incubation cycles	10.5	20
5	4 shaking/stationary incubation cycles + 24 hours stationary incubation	1 shaking/stationary incubation cycles	10.5	15
6	4 shaking/stationary incubation cycles + 24 hours stationary incubation	1 shaking/stationary incubation cycles	16.5	15

overnight at 37°C with shaking, and 2 μ L of cells were used directly as templates for PCR reactions. The 2 μ L of cells were boiled at 100°C in 78 μ L dH₂O for 10 minutes, then 19 μ L of PCR master mix (final concentrations 10 mM Tris, pH 8.4; 50 mM KCl; 2.5mM MgCl₂; 0.2mM dNTPs; 0.4 μ M each of the M13(-40)F and M13R primers) and 1 μ L (5U) of Taq polymerase were added. Following 15 thermal cycles (denaturation for 3 minutes at 95°C, then cycled 45 seconds at 95°C, 30 seconds at 45°C, and 90 seconds at 72°C; final extension for 3 minutes at 72°C), PCR products were purified with a Millipore (Billerica, MA) PCR clean-up kit and sequenced with Big Dye v3.0 mix (ABI, Foster City, CA).¹⁴ Sequencing reactions were analyzed on an ABI 3700 automated sequencer.

RESULTS AND DISCUSSION

The process of *in vitro* selection was automated by converting molecular biology steps that were normally carried out at the bench to steps that could be carried out by an automated workstation (Fig. 1). In order to carry out selection experiments a PCR machine, orbital shaker, solid phase extraction device (SPE), plate washer, and microplate carriers that maintained reagents at 4°C and -20°C had to be introduced on or adjacent to the worksurface.

As a test of the automated system, a double-stranded DNA selection was initiated against the transcription factor NF κ B p50 homodimer. Binding sites for this transcription factor had previously been identified by both the examination of promoter sequences and by manual selection experiments.¹ In order to determine whether the automated selection could potentially yield a NF κ B p50 homodimer binding site a previously selected high-affinity site (5' TGACTGATTGGGGGATCCCGAAGCTTATC) was doped into a double-stranded DNA pool that contained a similar sized random sequence region (N30). The high-

affinity site was included at a molar proportion of 1 to 10,000.

As with most molecular biology protocols, the primary steps in the automated selection protocol involved liquid handling. Initially, the DNA pool was moved from the 4°C microplate carrier to the MP-12 microplate carrier. Following incubation in the microtitre plates coated with NF κ B p50 homodimer, unbound aptamers were removed via a panning protocol. The advantage of using panning relative to other selection methods, such as filtration, is that the majority of the selection process (binding, washing, and PCR) could occur in the same well, reducing the number of liquid-handling steps and manipulations, and decreasing the possibility that rare binding sequences might be lost. Additionally, the amount of protein used to coat the microplate well in each round (0.3 μ g for the NF κ B p50 homodimer) is less than the amount that would be used in a round of a typical manual selection protocol (4.5 μ g). One potential disadvantage of using a panning protocol is that nucleic acid sequences might be selected that would bind to the hydrophobic surface of the plates, rather than to the immobilized target; for example, nucleic acids that bind to hydrophobic nitrocellulose filters frequently arise during filtration selections.¹⁵ To reduce the hydrophobic surface area that nucleic acids would be exposed to, the wells in the Top Yield plates were blocked following the immobilization of the NF κ B target. A wide variety of blocking agents were tested for their ability to reduce background binding, and Pierce Casein block was ultimately selected. To help prevent the selection of matrix-binding sequences, a negative selection was first carried out using blocked microtitre plates that did not contain NF κ B. Any nucleic acids that bound via non-specific hydrophobic interactions should have been lost from the selection at an early round. Such non-specific interactions would have been much more likely in the course of a single-stranded DNA selection, as the hydrophobicity of

single-stranded DNA is much greater than that of double-stranded DNA.

Following the negative selection, the DNA solution was transferred to wells containing NF κ B and thoroughly mixed via an orbital shaker. Stringency was varied by increasing the time allowed for plate-binding during the negative selection and decreasing the time allowed for target-binding during the positive selection (Table 1). Non-binding DNA species in solution were removed from the binding species immobilized on the surface of the plate. In manual selections this is one of the most critical but also one of the most tedious steps. In our panning protocol, a plate washer was used to rapidly wash the wells with 10.5–16.5 mL of buffer. In contrast, wash steps for bead- or filter-based selections typically rely upon repeatedly washing immobilized complexes with buffer aliquots of around 300 μ L. This slows the overall procedure and frequently results in researchers carrying out far less stringent selections than would otherwise be possible. While panning with the plate washer proved to be extremely efficient, one problem that was initially encountered was that pool DNA could stick to the aspiration needles on the plate washer, leading to cross-contamination between wells washed by the same needles. This problem was eliminated by washing the needles with 7 M urea after each use.

In order to amplify bound sequences, PCR reagents were added directly to the microtitre wells after the wash step. The PCR master mix was stored at 4°C and the Taq polymerase was stored at –20°C on the surface of the robot in cooled carriers connected to recirculating temperature baths. This allowed the automated selection to run essentially autonomously without the need for the addition of reagents at each step. Following reagent addition, the RoMa arm transfers the microtitre plate to the integrated thermal cycler, where the lid closes and a pre-set amplification program runs. The template DNA is conveniently eluted off of the target protein during the initial denaturation step of the PCR program. The number of thermal cycles required for amplification was determined by separating PCR products on agarose gels. After an initial optimization, it was determined that 15–20 thermal cycles would generally yield PCR products that could be carried forward into the next round. Following cycling, the plate is held at 4°C for 30 minutes to reduce aerosol formation, and then the lid of the PCR machine opens and the plate is transferred by the RoMa arm to the worksurface for the PCR product purification step.

Although automation of selection procedures helps to ensure reproducibility and increases throughput, consistency and attention to fine detail are essential for successful method development. As an example, one problem that initially arose was the cross-contamination of PCR products between different wells or cycles during or following amplification. In order to successfully integrate the thermal cycler and eliminate this problem, four separate optimizations were required. First, an MJ Research Microseal “P” adhesive-backed sealing pad was manually placed on the inside of the thermal cycler’s motorized Power Bonnet lid prior to the

selection, and the height of the lid was adjusted so that seals were formed around the top of each well. If the height was not adjusted properly, the PCR product was found to evaporate out of the well. Secondly, a heated lid was used to keep the PCR product from condensing on the lid. Third, an MJ Research Microseal 96 Plate Lifter was modified and placed in the thermal cycler so as to slightly lift the plate out of the unit when the lid of the thermal cycler opened. Without the Plate Lifter the lid pressed down on the plate so hard that it proved impossible for the RoMa arm to transfer the plate back to the Tecan work surface. The lid and Plate Lifter were adjusted until the plate was consistently available to the RoMa arm. Fourth, a slight vacuum was sometimes created in the wells during the heating and cooling process of the PCR program, causing the plate to stick to the lid of the thermal cycler when it opened. A final incubation at 4°C for 30 minutes not only minimized aerosol formation but also helped eliminate vacuum formation. All four of these optimizations were intertwined; for example, setting the Plate Lifter too high led to a greater probability of the plate sticking to the lid of the thermal cycler. These various improvements had to be iteratively implemented and tested in order to ensure that the final program would operate smoothly. As a standard of performance, if liquid was found to accumulate anywhere except in the microplate wells during a selection, the selection was terminated and the program was further modified.

Once amplification has been completed and the microtitre plate transferred to the Tecan worksurface, the PCR products were purified with a Qiagen kit. In order to automate this process, the two position SPE unit was equipped with an adapter that fit the Qiagen filter plate. The pure product was eluted into selection buffer, and was ready to initiate the next round of selection. Automation of the selection process provides a significant increase in throughput. While roughly the same amount of time is spent on reagent preparation and sequencing for manual and automated selections, there is a large difference in the time required for the selection process itself. Manual selections employ time-consuming techniques such as ethanol precipitation and purification via gel electrophoresis. Whereas these precautions can help to avoid the accumulation of amplification artifacts, they also typically extend a manual round of selection to several days, as opposed to the few hours required for an automated round. The total time savings over 6–18 cycles of selection (the number of rounds typically required for the purification of binding species) is therefore considerable, the difference between days and weeks.

Six rounds of automated selection were carried out, with the only human intervention being addition of PCR reagents to the target plate (although it should also be possible to automate this step, as well). The double-stranded DNA pool from Rounds 0, 3, and 6 were cloned and sequenced. The number of recognizable NF κ B binding sites progressively increased (Fig. 2). The results from the automated selection experiment were similar to those obtained from manual selection experiments. Rosen and co-workers identified

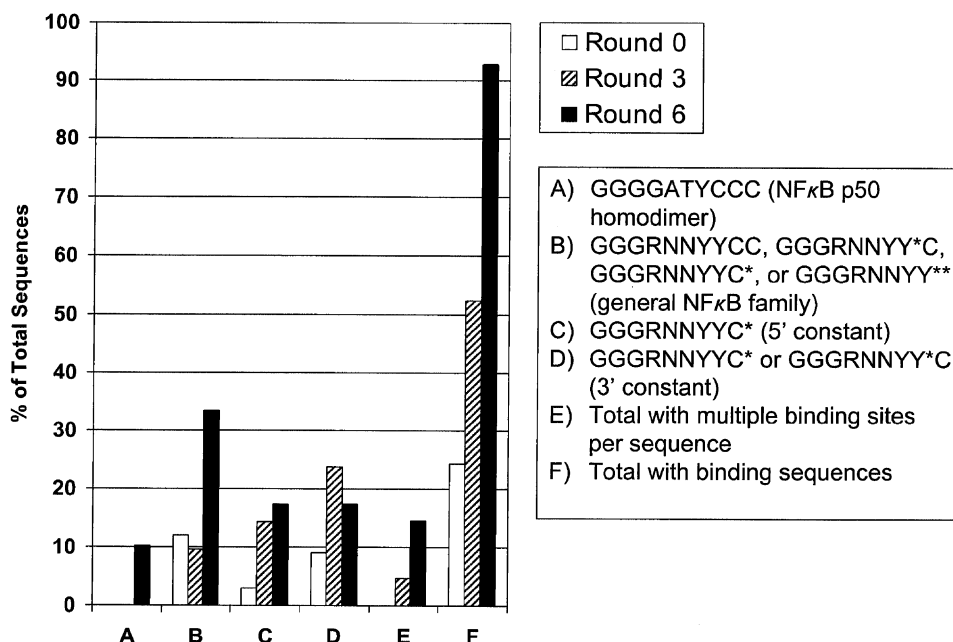


Figure 2. Frequencies of selected binding sequences. A variety of different NF κ B binding sequences were known prior to the start of the selection, and were recovered by the automated selection procedure. (A) Frequency of a core NF κ B p50 homodimer binding site. (B) Frequency of a more general representation of the NF κ B family binding site. (C and D) Frequency of the general NF κ B family binding sites that utilize either the 5' or 3' (respectively) constant regions from the original N30 pool. (E) Frequency of species which contain dimeric binding sites. The values in (E) are not included in the other tallies (A–D). (F) Total frequency of species with selected NF κ B p50 homodimer binding sequences, both core and more general family. In each instance, the white bar represents the 33 sequences derived from Round 0, the stippled bar represents the 21 sequences derived from Round 3, and the solid bar represents the 69 sequences derived from Round 6. Values in (A–E) sum to the value in (F).

a consensus, 10 base-pair (bp) NF κ B p50 homodimer binding sequence (5' GGGGATYCCC¹). Although the selected Rounds 0 and 3 did not contain the consensus NF κ B p50 binding sequence, it was present in 10% of the sequences by Round 6 (Fig. 2, A). Gorenstein and coworkers have suggested a more general consensus binding sequence (Fig. 2, B) (5' GGGRNNYYCC¹⁶). By Round 6 recognizable NF κ B binding sites were present in 93% of the sequenced clones (Fig. 2, F).

In addition to identifying NF κ B p50 homodimer binding sequences that corresponded to consensus binding sites, more subtle sequence contributions to protein recognition could also be discerned. For example, Kunsch et al.¹ had observed that the guanosine triplet at the 5' end of the 10 bp consensus NF κ B p50 homodimer binding sequence was essential for binding, whereas variations at the 3' cytidine doublet were tolerated. Similarly, we have found that additional sequence variations can occur in the 3' portion of this motif (5' GGGRNNYY*C, GGGRNNYYC*, and GGGRNNYY**, Fig. 2, B). Indeed, by Round 6, 62% of the binding sequences contained mutations in one or both of the 3' terminal cytidines. Kunsch et al.¹ also observed that there was frequently an additional guanine present at the 5' end of the consensus binding sites and an additional cytosine present at the 3' end (5' gGGGGATYCCCc). Similarly, all of the selected binding sites from Round 6 that contained the

10 bp consensus NF κ B p50 homodimer binding sequence also contained one or both of these additional 5' or 3' bases. Overall, 68% of the selected binding sites from Round 6 contained one or both of the additional bases; that is the core decamer binding site had apparently expanded to either an undecamer or dodecamer binding site.

The expansion of the previously determined core decamer binding site that is predicted by our selection experiments has recently been confirmed by other studies. The undecamer (5' GGGGATTCCCc) is palindromic about the central thymidine residue and is identical to the high affinity human major histocompatibility complex H-2 binding site for NF κ B.¹⁷ Crystal structures and DNA-protein crosslinking studies have shown that there are in fact specific base contacts between the NF κ B p50 homodimer and all four guanosine residues at the 5' end of the decamer core binding site sequence.¹ The palindromic undecamer therefore of necessity contains a guanosine quadruplet in each strand that each monomer of the homodimer can bind to.

Interestingly, selected DNAs that contained two NF κ B binding sites comprised 14% of the population by Round 6 (Fig. 2, E). The spacing between the two sites varied from zero (the core 10 bp consensus sites touched one another) to 22 nucleotides. Since selected DNAs that contained dimeric sites predominated only in the later rounds of selection, it seems likely that the presence of two sites resulted in

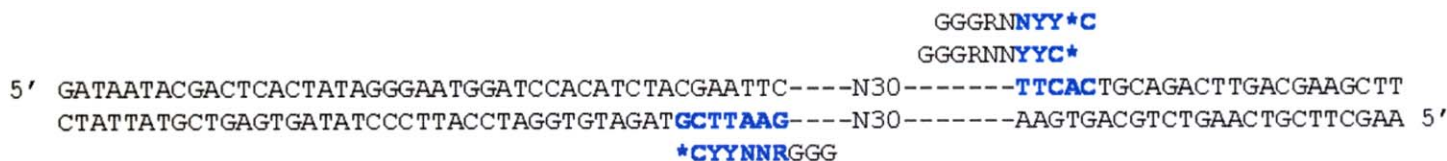


Figure 3. Utilization of constant regions in selected binding sequences. The original N30 pool is shown as a double-stranded DNA molecule. Superimposed on the pool are the locations of potential NF κ B binding sequences that utilize the 5' or 3' constant regions. These sequences correspond to the frequencies shown in Figure 2 (C and D).

a competitive advantage, by allowing multiple opportunities for interactions with the protein target.

The control sequence added to the pool did not take over the selection, as planned. Instead, the constant sequences in the N30 pool contributed to numerous NF κ B p50 homodimer binding sites (Fig. 3). Seven of the ten bases in the NF κ B family general consensus sequence were present in the 5' constant region, therefore only three bases were needed to complete the sequence. The probability of this was $1/4^3$ or 1 in every 64 molecules. Likewise, only five to six bases were needed to complete the consensus sequence utilizing the 3' constant region. The selection of N30 binding sites was also assisted by the previously mentioned sequence flexibility allowed at the 3' end of the consensus site, and again should have resulted in 1 in every 64 molecules containing a NF κ B binding site. Interestingly, by Round 6, utilization of the 5' and 3' constant regions to form NF κ B binding sites was equal (20 instances each), as predicted (Figs. 2, C and D). In contrast, the control sequences that were spiked (1 in 10,000 molecules) into the pool was much less populous than the binding sequences that occurred by chance in the N30 pool, and this helps to explain why they were not found by the conclusion of the selection. This tendency for selections, automated or manual, to utilize the most common, functional motifs has previously been observed and is known as the “tyranny of small motifs”.¹⁸ The fact that constant regions sometimes participate in binding sequences suggests that it may sometimes be desirable to compare the results of selection experiments with different pools, in order to more fully examine the range of binding sites that are possible.

Nonetheless, it should be noted that the same consensus NF κ B p50 homodimer binding site that was present on the control (5' GGGGATTCCC) was also recovered from the completely random region alone (Fig. 2, A). This site would have been present in the population at roughly the same frequency as the positive control (taking into account multiple possible registers), but by the conclusion of the selection it was present in 10% of the population. However, as we discussed above, many of the NF κ B binding sites that were recovered from the selection likely formed even more contacts with the protein than did the previously identified core decamer binding site, and thus would have enjoyed a selective advantage relative to the control. Therefore, our results actually highlight the extraordinary potential of robotic selection experiments to overcome even the tyranny of small motifs (in this instance, both degenerate but

populous binding motifs and the positive control). That is, with additional rounds of robotic selection, the best sequences can be culled from a population, even though they may be only incrementally better than a majority sequence.

CONCLUSIONS

Double-stranded DNA aptamer selections against the NF κ B p50 homodimer were successfully automated using a Tecan Genesis workstation. The consensus DNA binding site for NF κ B was isolated from a pool of 10^{13} double-stranded, random sequence oligonucleotides. Although the consensus binding sequence that was originally spiked into the pool did not rise to the fore in the automated selection, the fact that an unanticipated but more likely set of binding sequences was ultimately chosen was also proof that the automated method worked well. Moreover, the fact that variations observed between individual selected sequences could be largely explained based on the binding propensities of known NF κ B p50 homodimer binding sites indicated that the automated selection method should be capable of fully describing binding sites for other transcription factors. A significant increase in throughput was achieved, from several days for a round of manual selection to four hours for a round of automated selection.

ACKNOWLEDGMENTS

This research was funded by a grant from the Office of Naval Research (N00014-99-1-0861), a NIH grant (R21 AI053548-01A1), and a grant from the Department of Army Research (MURI) (DAAD19-99-1-0207).

REFERENCES

- Kunsch, C.; Ruben, S. M.; Rosen, C. A. Selection of Optimal kB/Rel DNA Binding Motifs: Interaction of Both Subunits of NF- κ B with DNA is Required for Transcriptional Activation. *Mol. Cell Biol.* **1992**, *12*, 4412–4421.
- Medici, N.; Abbondanza, C.; Nigro, V.; Rossi, V.; Piluso, G.; Belsito, A.; Gallo, L.; Roscigno, A.; Bontempo, P.; Puca, A. A.; Molinari, A. M.; Moncharmont, B.; Puca, G. A. Identification of a DNA Binding Protein Cooperating with Estrogen Receptor as RIZ (Retinoblastoma Interacting Zinc Finger Protein). *Biochem. Biophys. Res. Commun.* **1999**, *264*, 983–989.
- Funk, W. D.; Pak, D. T.; Karas, R. H.; Wright, W. E.; Shay, J. W. A Transcriptionally Active DNA Binding Site for Human p53 Protein Complexes. *Mol. Cell Biol.* **1992**, *12*, 2866–2871.

4. Wright, W. E.; Binder, M.; Funk, W. Cyclic Amplification and Selection of Targets (CASTing) for the Myogenin Consensus Binding Site. *Mol. Cell Biol.* **1991**, *11*, 4104–4110.
5. Funk, W. D.; Wright, W. E. Cyclic Amplification and Selection of Targets for Multicomponent Complexes: Myogenin Interacts with Factors Recognizing Binding Sites for Basic Helix-Turn-Helix, Nuclear Factor 1, Myocyte-Specific Enhancer-Binding Factor 2, and COMPI Factor. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 9484–9488.
6. Roulet, E.; Busso, S.; Camargo, A. A.; Simpson, A. J.; Mermod, N.; Bucher, P. High-Throughput SELEX SAGE Method for Quantitative Modeling of Transcription-Factor Binding Sites. *Nat. Biotechnol.* **2002**, *20*, 831–835.
7. Boardman, P. E.; Oliver, S. G.; Hubbard, S. J. SiteSeer: Visualisation and Analysis of Transcription Factor Binding Sites in Nucleotide Sequences. *Nucl. Acids Res.* **2003**, *31*, 3572–3575.
8. Sinha, S.; Tompa, M. Discovery of Novel Transcription Factor Binding Sites by Statistical Overrepresentation. *Nucl. Acids Res.* **2002**, *30*, 5549–5560.
9. Sinha, S.; Tompa, M. YMF: A Program for Discovery of Novel Transcription Factor Binding Sites by Statistical Overrepresentation. *Nucl. Acids Res.* **2003**, *31*, 3586–3588.
10. Muller, C. W.; Rey, F. A.; Sodeoka, M.; Verdine, G. L.; Harrison, S. C. Structure of the NF- κ B p50 Homodimer Bound to DNA. *Nature* **1995**, *373*, 311–317.
11. Ghosh, S.; May, M. J.; Kopp, E. B. Evolutionarily Conserved Mediators of Immune Responses. *Annu. Rev. Immunol.* **1998**, *16*, 225–260.
12. Ghosh, S.; Karin, M. Missing Pieces in the NF- κ B Puzzle. *Cell* **2002**, *109*, S81–S96.
13. Bell, S. D.; Denu, J. M.; Dixon, J. E.; Ellington, A. D. RNA Molecules that Bind to and Inhibit the Active Site of a Tyrosine Phosphatase. *J. Biol. Chem.* **1998**, *273*, 14309–14314.
14. Harkey, C. Sample prep for automated sequencing. <http://www.icmb.utexas.edu/core/DNAFacility/Sequencing%20Handout.pdf> (accessed Feb 2003).
15. Tuerk, C.; MacDougall, S.; Gold, L. RNA Pseudoknots that Inhibit Human Immunodeficiency Virus Type 1 Reverse Transcriptase. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 6988–6992.
16. King, D. J.; Bassett, S. E.; Li, X.; Fennewald, S. A.; Herzog, N. K.; Luxon, B. A.; Shope, R.; Gorenstein, D. G. Combinatorial Selection and Binding of Phosphorothioate Aptamers Targeting Human NF- κ B RelA(p65) and p50. *Biochemistry* **2002**, *41*, 9696–9706.
17. Angelov, D.; Charra, M.; Muller, C. W.; Cadet, J.; Dimitrov, S. Solution Study of the NF- κ B p50-DNA Complex by UV Laser Protein-DNA Cross-Linking. *Photochem. Photobiol.* **2003**, *77*, 592–596.
18. Ellington, A. D. Empirical Explorations of Sequence Space: Host-Guest Chemistry in the RNA World. *Ber. Bunsenges. Phys. Chem.* **1994**, *98*, 1115–1121.