Nucleic Acids Research, 2005, Vol. 33, No. 17 ***e141***
doi:10.1093/nar/gni139

Combining SELEX with quantitative assays to rapidly obtain accurate models of protein–DNA interactions

Jiajian Liu and Gary D. Stormo*

Department of Genetics, Washington University School of Medicine, 660 S Euclid, Box 8232, St Louis, MO 63110, USA

Received March 23, 2005; Revised April 25, 2005; Accepted August 26, 2005

ABSTRACT

Models for the specificity of DNA-binding transcription factors are often based on small amounts of qualitative data and therefore have limited accuracy. In this study we demonstrate a simple and efficient method of affinity chromatography-SELEX followed by a quantitative binding (QuMFRA) assay to rapidly collect the data necessary for more accurate models. Using the zinc finger protein EGR as an e.g. we show that many bindings sites can be obtained efficiently with affinity chromatography-SELEX, but those sequences alone provide a weight matrix model with limited accuracy. Using a QuMFRA assay to determine the quantitative relative affinity for only a subset of the sequences obtained by SELEX leads to a much more accurate model. Application of this method to variants of a transcription factor would allow us to generate a large collection of quantitative data for modeling protein–DNA interactions that could facilitate the determination of recognition codes for different transcription factor families.

INTRODUCTION

Searches of genome sequences for potential transcription factor binding sites (TFBS) typically use some form of weight matrix as the model for the TFBS (1). Databases, such as TRANSFAC and JASPAR (2,3), build the matrices from known binding sites. But if there are few known sites or if the collection of them is biased in any way, the resulting matrix may not be a good representation of the true specificity of the TF. There are a number of experimental methods for studying protein–DNA interactions. For example, one might use yeast the one-hybrid system or ChIP-chip experiments to identify protein–DNA interactions *in vivo* (4,5). DNA microarrays have also been used to determine the specificity of DNA-binding proteins (6). Probably the most versatile method is SELEX [(7), also referred to as SAAB (8) and CASTing (9)], in which a purified protein is used to isolate high affinity binding sites through several rounds of *in vitro* selection and amplification (10–12). The power of this method lies in its ability to isolate a small set of specific binding sites from a very large pool of random sequences. One important aspect of a SELEX experiment is how DNA–protein complexes are physically sieved from free DNA. Traditionally, the DNA–protein complexes are separated from free DNA by the gel mobility shift method where the DNA is radiolabeled (7,10,11). In this study, we used affinity chromatography to modify the conventional SELEX procedure to save time and labor. A set of preferred DNA-binding sites for Zif268 (6,10) finger 1 are highly enriched by two-selection rounds, as determined by sequencing only ~20 selected products after each round. The weight matrix model determined by the selected DNA sites by the SELEX provides a good initial estimate of the binding preference of the protein. However, to obtain a more accurate model we would either have to sequence many more products (13) or do some quantitative measurements (14–17). We further examine the quantitative affinities for some of the selected sites with the QuMFRA assay (15) and used these data to build a quantitative model. Using an independent dataset we compared these two models. The results show that the model derived from the SELEX performed moderately well only for predicting DNA-binding affinities, with the quantitative model significantly improving the prediction performance. The method developed in this study, combining SELEX with quantitative specificity measurements, provides a general and effective method that can be used for any DNA-binding protein even if nothing is known about its specificity.

MATERIALS AND METHODS

Preparation of GST-zif268 fusion

To create GST-Zif268 fusion, the previously constructed pET-18a-Zif268 plasmid (17) bearing the DNA-binding domain (DBD) of Zif268 was cleaved with BamH1 and EcoRI.

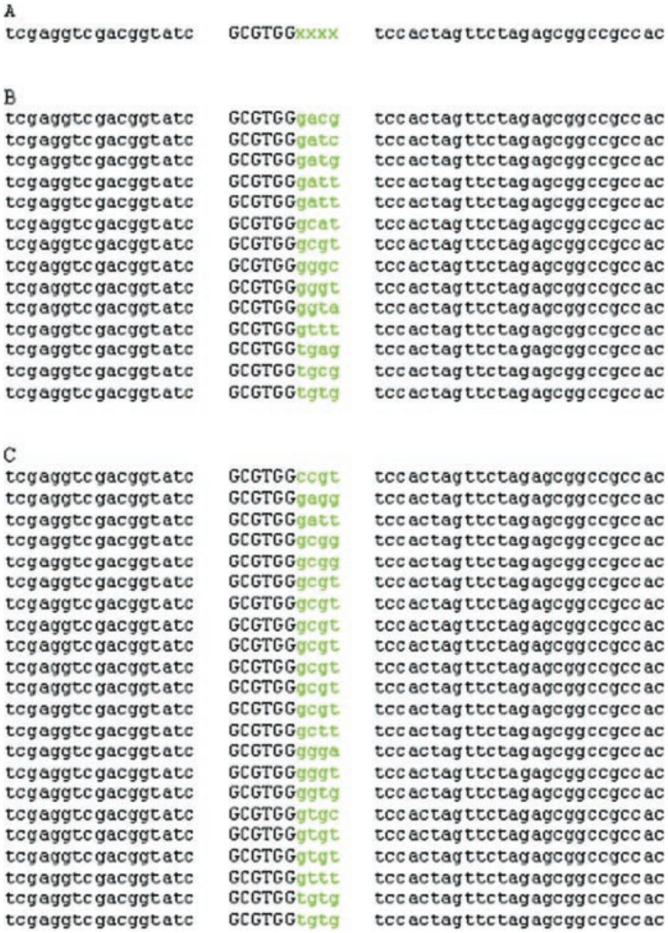
*To whom correspondence should be addressed. Tel: +1 314 747 5534; Fax: +1 314 362 2156; Email: stormo@genetics.wustl.edu

The fragment of Zif268 was inserted into the pGEX-4T-1 vector (Amersham Pharmacia Biotech) digested with BamH1 and EcoRI. The resulting plasmid pGST-Zif268 was verified by DNA sequencing. *Escherichia coli* BL21 cells bearing pGST-Zif268 were grown in 2× YT medium at 37°C with constant shaking. Isopropyl-β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 1 mM when OD_{600} reached 0.6–1.0. Cells were harvested 3 h after IPTG induction by centrifugation at 4000 r.p.m. (Sorvall SLA-3000 rotor) for 20 min. The pellets were then resuspended in 15 ml of lysis buffer [50 mM Tris–HCl (pH 8.0), 300 mM NaCl, 10 mM DDT and 1 tablet of protease inhibitor cocktail tablets (Roche)] and lysed with sonication. The pellets were then separated by centrifugation at 6000 r.p.m. (Sorvall SS-34 rotor) for 20 min and insoluble material was removed. The GST-fusion was purified with glutathione Sepharose (glutathione Separose™ 4B) chromatography using the procedures as suggested by the manufacturer (Amersham Pharmacia). The GST-Zif268 was eluted with the elution buffer [50 mM Tris (pH 8.0), 0.25 M KCl and 10 mM glutathione]. Elution fractions were analyzed on 12% SDS–PAGE gel, followed by silver staining to determine their purity. Finally, different elution fractions were pooled and dialysed against the dialysis buffer [30 mM Tris–HCl (pH 8.0), 50 mM NaCl and 3 mM DTT] at 4°C, followed by concentration with Amicon Ultra-4 filters (MWCO 5K) and kept at −80°C until usage. The protein concentration was determined with Bio-Rad protein assay kit.

SELEX procedure

The oligonucleotide pool contains a 53 base template strand with the following segments (Figure 1A): 6 bases of the wild-type binding site for Zif268 (the positions that interact with fingers 2 and 3); a 4 base randomized region to be selected for sites interacting with finger 1; two fixed sequences of DNA for PCR amplification and with SalI and XbalI restriction sites for cloning into the sequencing vector. Double-stranded DNAs employed for the SELEX experiment were created by PCR amplification. The pooled PCR products were phenol/chloroform treated and separated in a 2% agarose gel to create a pool of 53 bp double-stranded DNA. The DNA concentration was determined with PicoGreen dsDNA quantitation kit (Molecular Probes).

The resultant PCR products (∼10^{-8} M) were incubated with the purified GST-tagged Zif268 protein (∼10^{-8} M) in 1× reaction buffer [30 mM Tris–HCl (pH 8.0), 50 mM NaCl, 0.1 mg/ml BSA, 3 mM DTT, 20 μM ZnSO4, salmon sperm DNA 25 μg/ml] with the final volume of 100 μl at room temperature for about 1 h. In order to separate GST-tagged Zif268-DNA complex from free DNA, 20 μl of glutathione Sepharose slurry equilibrated with 1× reaction buffer was added into the reaction. The mixture was kept at room temperature for 1 h. After washing three times with 500 μl of 1× reaction buffer, the glutathione Sepharose slurry bearing bound DNA sites were transferred onto a small column. The column was washed again with 1 ml of 1× reaction buffer. The protein–DNA complexes were eluted with 40 μl of elution buffer [50 mM Tris (pH 8.0), 0.25 M KCl and 10 mM glutathione]. The bound DNAs in the eluted fraction were amplified by the PCR and used for the next round of selection after purification with 2% agarose gel. The entire selection procedure was performed for two times in this study. At the end of each selection cycle, the isolated DNA sites were cut with SalI and XbaI and inserted into pBluscript II KS vectors (Stratagene) digested with the same enzymes. These plasmids were then transformed into *E.coli* 5Dα competent cells. Individual clones were selected for sequencing by a standard dideoxy procedure.



```
A
tcgaggtcgacggtatc    GCGTGGxxxx    tccactagttctagagcggccgccac

B
tcgaggtcgacggtatc    GCGTGGgacg    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgatc    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgatg    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgatt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgatt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcat    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgggc    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgggt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGggta    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgttt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGtgag    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGtgcg    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGtgtg    tccactagttctagagcggccgccac

C
tcgaggtcgacggtatc    GCGTGGccgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgagg    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgatt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgg    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgg    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgcgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgctt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGggga    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgggt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGggtg    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgtgc    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgtgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgtgt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGgttt    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGtgtg    tccactagttctagagcggccgccac
tcgaggtcgacggtatc    GCGTGGtgtg    tccactagttctagagcggccgccac
```

Figure 1. *In vitro* selection for Zif268 finger 1 with the affinity chromatography-SELEX procedure. (**A**) The DNA template used for *in vitro* selection in this study. The fixed flanking sequences bind to the PCR primers and contain restriction sites for cloning. The capitalized sequence in the center is the consensus for interacting with fingers 2 and 3 of Zif268. The 'xxxx' are the randomized positions. (**B**) The sequences for 14 selected DNA sites obtained from the first-round of selection; (**C**) The sequences for 22 selected DNA sites obtained from the second-round of selection.

QuMFRA assay to determine the relative binding constants

Fifteen DNA sites obtained from the SELEX procedure were examined to determine their relative binding affinities using QuMFRA assay developed by Man and Stormo (15) with some modifications. Double-stranded oligonucleotide binding sites used in this study were generated by PCR. In each PCR, the plasmid pBluscript bearing selected Zif268 binding sites obtained from the SELEX was used as template. The two primers are KS (TCGAGGTCGACGGTATC) and SK

(GTGGCGGCCGCTCTAGAACTAGTGGA). The SK primer was labeled with one of the following three flurophores: FAM, TAMRA or ROX. The PCR products were dissolved in TS buffer [10 mM Tris–HCl (pH 8.0) and 50 mM NaCl] after purification and precipitated with 1/10 vol of 3 M NaAc and an equal volume of isopropanol.

The competitive binding assay was performed by mixing three different fluorophore-labeled DNA-binding sites with a certain amount of GST-tagged Zif268 ($\sim 10^{-8}$ M) in 1× reaction buffer [30 mM Tris–HCl (pH 8.0), 50 mM NaCl, 0.1 mg/ml BSA, 3 mM DTT, 20 µM ZnSO4 and poly (dI–dC) 5 µg/ml). The oligo with GGGT in the randomized positions was included in each reaction as the reference site to which all other affinities were compared. The reaction was equilibrated for 1 h at room temperature before being electrophoresed on a 10% polyacrylamide gel. Each of the three fluorophore-labeled PCR products was loaded individually onto the same gel. After electrophoresis, the gels were scanned by a Typhoon Variable Scanner (Molecular Dynamics, Sunnyvale, CA) to obtain the fluorescent intensities of separated bands at three different emission wavelengths using the same machine settings as employed by Man and Stormo (15). The resultant fluorescence intensities of a separated band for a reaction mixture at emission wavelengths make up the output vector $\vec{o}$, the fluorescence intensities of the three individual fluorophore-labeled DNA at three emission wavelengths constitute the emission matrix E. From the output vector and the emission matrix, the intensities of each DNA in a separated band, represented as a vector $\vec{x}$, were obtained from the relationship $E \times \vec{x} = \vec{o}$ (17).

Once the intensities of all three fluorophore-labeled oligos are known for both the bound and unbound fractions, the relative binding constant of a test DNA site in the reaction to the reference binding site can be calculated as follows (15):

$$K_{a(rel)} = \frac{[PD_{test}][D_{ref}]}{[PD_{ref}][D_{test}]}$$

where $[PD_{test}]$ and $[PD_{ref}]$ are the concentration of bound DNA for test site and the reference site, respectively, while $[D_{test}]$ and $[D_{ref}]$ are the concentration of free DNA for the test site and the reference site, respectively. The concentrations for all DNA sites are represented as computed fluorescence intensities as described above.

Sequence logos

The sequence logos were created using the EnoLOGOS program (18) by providing the binding probabilities, or computed weights representing relative free energy of binding, for each base at each position in the sites. All logos were plotted in bits by converting logarithms to base 2.

RESULTS

In vitro selection for binding sites for Zif268 finger 1

To simplify the conventional SELEX method, we utilized an affinity chromatography-SELEX procedure to separate the DNA–protein complex from free DNA. The Zif268 gene was cloned into a pGST vector so as to produce a GST-Zif268 fusion protein. No detectable DNA nonspecific binding to glutathione Sepharose resin allowed us to conduct *in vitro* selections with GST fusions. Since the tetra-nucleotides designed for *in vitro* selection of the binding sites of Zif268 finger 1 are fully random, it will provide an unbiased method for selecting the true sequence specific for Zif268 finger 1.

Figure 1A shows the initial DNA to be used in the selections, where the x's are positions that are completely randomized, giving rise to a mixture of all 256 different 4 bp long sequences to interact with finger 1 of the GST-Zif268 protein. Figure 1B shows 14 selected binding sites that were sequenced after the first cycle of selection. This first-round shows some selection, for e.g. the strong preference for the base G at the first position, but overall the first-round of selection could not sufficiently identify the preferred DNA sites for Zif268 finger 1. Figure 1C shows 22 selected binding sites obtained after the second-round of selection. Examination of the selection data showed a consensus sequence for Zif268 finger 1 of GCGT in agreement with a previous report (19). The consistency between these results shows that this simplified procedure, affinity chromatography-SELEX, works well.

Given the collection of selected binding sites in Figure 1 we can estimate their affinities for Zif268 finger 1, under the assumption of an additive model where each position can contribute independently to the total binding free energy. Figure 2A shows the number of each type of base at each specific position in the complete set of 36 binding sites shown in Figure 1B and C. We then convert the frequency alignment matrix into the weight matrix using the method of Berg and von Hippel (20):

$$W(b,i) = \ln \frac{n_{b,i}}{\max_b n_{b,i}}$$

where $W(b, i)$ is the weight for base b at position i, $n_{b,i}$ is the number of base b at position i, and $\max_b n_{b,i}$ is the number of base b that is most common at position i. This sets the value for the most frequent base to 0 and all of the others are negative values reflecting the decrease in occurrence compared to the preferred (consensus) base. According to the Berg and von Hippel theory these matrix values should be proportional to the difference in binding energy contributions of the different bases (20). These differences are the same as those obtained from the standard log-odds scoring method if we assume the background frequencies of the different bases are all equal (1). While that assumption is probably valid for the first-round of selection it won't be for the second, but we would need many more sequences from the first-round to accurately estimate those background frequencies, so we use this simplifying assumption. The resultant weight matrix and the sequence logo are represented in Figure 2B and C. The score for any particular site is the sum of matrix values for that site's sequence.

Quantitative DNA specificity of Zif268 finger 1

To test whether the matrix determined from the SELEX data is an accurate predictor of binding energies, and to obtain a better matrix if it is not, we chose 15 of those sequences to perform a

Table 1. Experimentally determined relative binding constants for 15 selected sites by QuMFRA assays using GGGT as the reference sequence

Sequences	Relative *Ka*
GCGT	39.93 (1.62)
GCGG	23.32 (2.39)
GAGG	10.36 (0.14)
GCAT	7.22 (0.32)
GTTT	5.68 (0.12)
GGTG	4.58 (0.12)
TGTG	4.01 (0.72)
GGGA	2.14 (0.19)
TGCG	1.78 (0.43)
CCGT	1.19 (0.12)
GGTA	1.17 (0.01)
GTGC	1.15 (0.09)
GGGT	1.00
TGAG	0.39 (0.05)
GATC	0.04 (0.007)

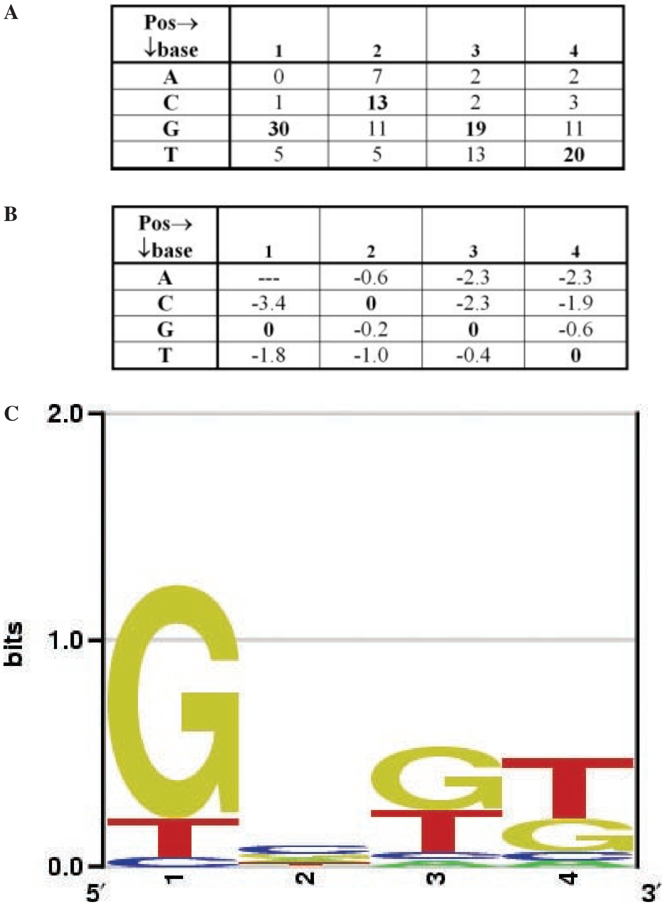
Each data were obtained from three or more independent examinations, inside of parenthesis are the SDs.

Table 2. Normalized relative affinities for the fifteen sequences whose affinities were measured and the expected relative affinities (scores) from the SELEX matrix and from the quantitative data based model

Affinity sequence	Relative K_a	ln(relative K_a)—ln(100)	SELEX score	Quantitative score
GCGT	100	0	0	0
GCGG	58	−0.5	−0.6	1.0
GAGG	26	−1.3	−1.2	−1.3
GCAT	18	−1.7	−2.3	−1.8
GTTT	14	−2.0	−1.4	−1.0
GGTG	11	−2.2	−1.2	−2.2
TGTG	10	−2.3	−3.0	−2.3
GGGA	5.4	−2.9	−2.5	−2.5
TGCG	4.5	−3.1	−2.6	−2.6
CCGT	3.0	−3.5	−3.4	−3.0
GGTA	2.9	−3.5	−2.9	−3.0
GTGC	2.9	−3.5	−1.6	−3.6
GGGT	2.5	−3.7	−0.2	−2.7
TGAG	1.0	−4.6	−4.9	−3.1
GATC	0.1	−6.9	−2.9	−5.9

The relative affinities (scores) are computed by the sum of weights for specific base across different positions for the particular DNA site.

QuMFRA assay on. These 15 sequences cover the entire space of possible sequences in the sense that every base occurs at least once at every position in the set (except that our collection is missing any occurrences of A in the first position). The quantitative binding data for that set of sequences is shown in Table 1, where the sequence GGGT was chosen as the reference. Table 2 renormalizes the same data so that the consensus sequence, GCGT, is assigned as a binding affinity of 100 and the relative affinity of every other sequence is shown compared to that. Column 3 of Table 2 is the natural logarithm of the relative affinity, compared to the consensus site, for each of the other binding sites. Column 4, the SELEX score, shows the score for each site from the weight matrix of Figure 2B. The correlation between the SELEX score and the ln(relative K_a) is 0.60. While this shows a strong relationship between the two measures, it also indicates that there is ample room for improvement in the prediction of binding affinities. Figure 3A shows the weight matrix with the best fit to the quantitative binding data using multiple linear regression (21).



A

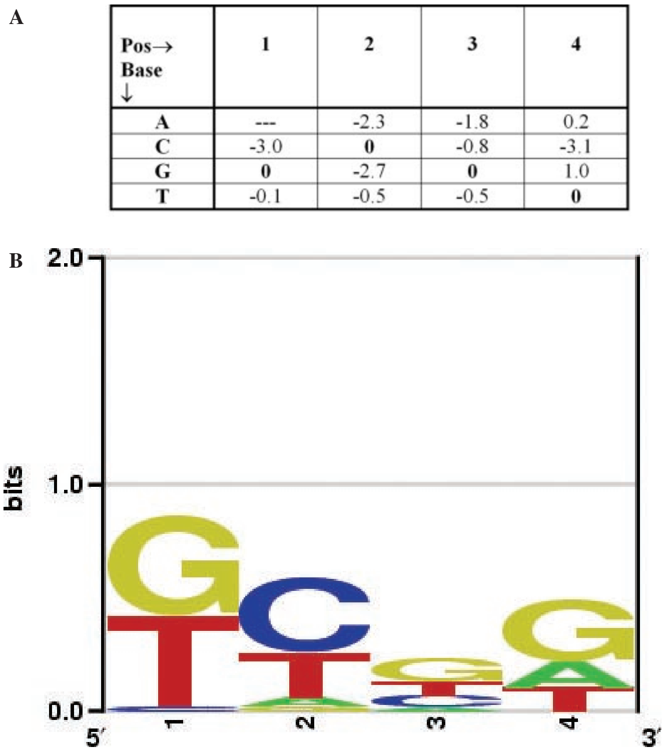
Pos→ ↓base	1	2	3	4
A	0	7	2	2
C	1	**13**	2	3
G	**30**	11	**19**	11
T	5	5	13	**20**

B

Pos→ ↓base	1	2	3	4
A	---	-0.6	-2.3	-2.3
C	-3.4	**0**	-2.3	-1.9
G	**0**	-0.2	**0**	-0.6
T	-1.8	-1.0	-0.4	**0**

Figure 2. SELEX based model for representing DNA specificity for Zif268 finger 1. (**A**) The frequency matrix for Zif268 finger 1 that was obtained from the alignment of all 36 sites shown in Figure 1. (**B**) The weight matrix for Zif268 finger 1 that was obtained from the frequency matrix. (**C**) The sequence logo for Zif268 finger 1 from the weight matrix. For the logo logarithms were converted to base 2 to display the results in bits.

Each weight (Figure 3A) represents the relative binding energy for each base at the corresponding position. The sequence logo (Figure 3B) was created using the EnoLOGOS program (18) where the inputs are the calculated energies as shown in Figure 3A. The score of each binding site by this matrix is shown in column 5 of Table 2, which has a correlation coefficient of 0.94 with the ln(relative K_a) data. This shows that one can obtain a matrix with a much better fit than using the SELEX data alone, but is not a fair comparison of the two matrices because the quantitative matrix was determined as the best fit to those data. A better comparison is to see how well the two matrices predict the binding affinities to an independent set of binding affinity measurements. Unfortunately there is a no comprehensive collection of affinity measurements for variants of the finger 1 binding site, but there are a few papers that have performed a limited number of quantitative comparisons (17,22–24). Table 3 shows the average (from those four papers) relative binding affinity to eight different variants of the consensus sequence, with changes at one or both of the two central, randomized positions. Columns 3–5 of Table 3 shows the ln(relative K_a) measurements and the scores obtained by the SELEX matrix (Figure 2B) and the



Pos→ Base ↓	1	2	3	4
A	---	-2.3	-1.8	0.2
C	-3.0	0	-0.8	-3.1
G	0	-2.7	0	1.0
T	-0.1	-0.5	-0.5	0

Figure 3. Binding model from quantitative data. (**A**) The weight matrix with the optimum parameters obtained by multiple regression on the binding affinity data. (**B**) The sequence logo for Zif268 finger 1 from that weight matrix.

Table 3. Relative affinities for nine sequences in the independent testset and the relative affinities predicted by the SELEX matrix and by the quantitative matrix

Affinity sequence	Relative K_a	ln(relative K_a)—ln(100)	SELEX score	Quantitative score
GCGx	100	0	0	0
-TG-	27	−1.3	−1.0	−0.5
-CA-	23	−1.5	−2.3	−1.8
-AG-	15	−1.9	−0.6	−2.3
-CT-	14	−2.0	−0.4	−0.5
-CC-	11	−2.2	−2.3	−0.8
-GG-	6	−2.8	−0.2	−2.7
-AA-	6	−2.8	−2.9	−4.1
-AC-	4	−3.2	−2.9	−3.1

The relative affinities (scores) are computed by the sum of weights for specific base across different positions for the particular DNA site.

Quantitative matrix (Figure 3A). The correlation with the SELEX matrix is 0.54, similar to the value obtained above. The correlation with the Quantitative matrix is 0.77, not as high as on the data for which the matrix was obtained (as expected) but significantly better than the correlation from the SELEX data alone.

DISCUSSION

Knowing the specificity of a TF is essential to locating its binding sites within the genome and mapping the regulatory network of the cell. While databases such as TRANSFAC and JASPAR have weight matrices for several TFs, they are far from complete and their accuracy in predicting new binding sites is often limited. SELEX is a versatile method for determining the specificity of any purified TF, but using the sites obtained by SELEX to model the specificity of a TF can also have limitations. Most SELEX experiments report only a few binding sites which leads to imprecise determination of specificity. SAGE-SELEX (13) makes it efficient to collect many more binding sites for the same amount of effort. But even with a large sample size obtaining an accurate model for the TF's specificity can be difficult due to the biases in the SELEX procedure and the changing input frequencies at each round.

As an alternative to doing a statistical analysis on a large sample one can determine quantitative binding affinities to a smaller sample and then obtain the model that produces the best fit to the data. dsDNA chips that contain binding sites for a TF of interest can be used to obtain quantitative binding data in a relatively high-throughput manner (6). But to use this method for a protein of unknown specificity one would have to include nearly all possible binding sites on the chip, and for proteins that bind to long sites, such as 20mers, that would be prohibitive. Once a consensus sequence for a TF is known, one can also do a thorough quantitative analysis of many similar sequences to determine the effects of different variations (14–16). Once the consensus is known and the sequence variants chosen and synthesized, this approach can also be relatively high-throughput and return very accurate binding affinity measurements. But we think the method demonstrated in this study can obtain accurate binding models for a TF of unknown specificity more efficiently than any previous method. If the specificity of the TF is initially unknown, and even the size of the binding site is not known, SELEX provides a very efficient means of obtaining a good collection of high affinity sites. One would usually start with a randomized region of 20 or more bases, since nearly all TFs bind to sites of that length or shorter. One would sequence about one hundred individual sites, probably from different rounds of SELEX. The SAGE-SELEX method can make that very efficient because about ten sites can be determined from every sequencing read (13). Standard motif alignment methods are quite good at determining the optimum alignment of SELEX data because, sites are all contained within a relatively short sequence, and from that alignment a consensus sequence is usually obvious. Now one chooses from among the sequences that were selected a subset that 'surround' the consensus sequence with variations of all bases at all positions. For a consensus binding site that is L-long, one needs at least 3L sites for quantitative analysis in order to determine all of the parameters of the weight matrix. In practice 5L–6L sites would provide more confidence in the resulting model and allow for the discovery of any important non-independent interactions in the binding sites. These sequences to be used for quantitative analysis do not need to be synthesized because they already exist in the selected pool. They just need to be 'lifted' from the sequencing vector by PCR, using the fluorescently labeled primers that are used in the QuMFRA. And they do not need to be attached to any chips but instead are used directly in the 'gel-shift' assay. Using four colors, three relative affinity measurements can be obtained for each lane of a gel, so a typical gel with 25 lanes can return up to 75 relative affinity measurements. So even for long sites of about

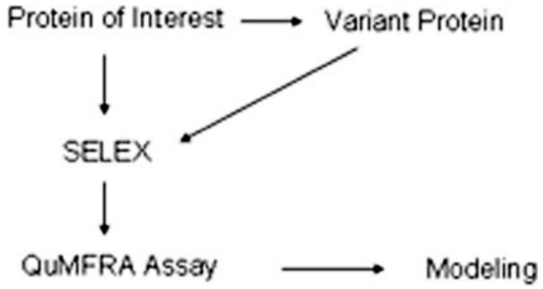


Figure 4. Strategy for obtaining quantitative data for a family of TFs. For any specific TF the SELEX data followed by the QuMFRA assay will provide the necessary information for a quantitative modeling of its specificity. Then creating variants in the TF sequence and performing the same experimental steps will allow for the development of models in which both the protein sequence and the binding site sequences are variable. Such a model constitutes a 'probabilitistic recognition code' that can be used to predict binding affinities for any combination of binding site and TF sequence from that family (25,26).

20 bases, for which one would like to make quantitative binding measurements for about 100 different sequences, and doing each measurement in duplicate, would only require about three gels.

Figure 4 outlines a procedure for applying this approach to determining the 'recognition code' for any TF family of interest (25). The vertical line is the procedure demonstrated in this study, where SELEX is applied to a specific TF to obtain a collection of binding sites, some of which are then subjected to a QuMFRA assay to determine their relative binding affinities. Modeling methods, such as the multiple regression approach used here or other methods that have been applied previously (26), can then determine the best specificity model for that TF. Then one would make variations in the TF, specifically altering amino acids that interact directly with the DNA-binding site, and repeat the procedure to obtain an optimum model for that protein. We expect that following such a procedure for any TF family could efficiently obtain sufficient data to develop a reasonably accurate recognition code model for that family.

ACKNOWLEDGEMENTS

We thank Xiaoming Zhang for help with preparation of the GST-fusion protein. This work was supported by NIH grant GM28755. Funding to pay the Open Access publication charges for this article was provided by NIH grant GM28755.

Conflict of interest statement. None declared.

REFERENCES

1. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
2. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
3. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
4. Alexander,M.K., Bourns,B.D. and Zakian,V.A. (2001) One-hybrid systems for detecting protein–DNA interactions. *Methods Mol. Biol.*, **177**, 241–259.
5. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
6. Bulyk,M.L., Huang,X., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
7. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
8. Blackwell,T.K. and Weintraub,H. (1990) Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science*, **250**, 1104–10.
9. Wright,W.E., Binder,M. and Funk,W. (1991) Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol. Cell Biol.*, **11**, 4104–4110.
10. Wolfe,S.A., Greisman,H.A., Ramm,E.I. and Pabo,C.O. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.*, **285**, 1917–1934.
11. Fields,D.S., He,Y., Al-Uzri,A.Y. and Stormo,G.D. (1997) Quantitative specificity of the Mnt repressor. *J. Mol. Biol.*, **271**, 178–194.
12. Silbaq,F.S., Ruttenberg,S.E. and Stormo,G.D. (2002) Specificity of Mnt 'Master Residue' obtained from *in vivo* and *in vitro* selections. *Nucleic Acids Res.*, **30**, 5539–5548.
13. Roulet,E., Busso,S., Camargo,A.A., Simpson,A.J., Mermod,N. and Bucher,P. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, **20**, 831–835.
14. Linnell,J., Mott,R., Field,S., Kwiatkowski,D.P., Ragoussis,J. and Udalova,I.A. (2004) Quantitative high-throughput analysis of transcription factor binding specificities. *Nucleic Acids Res.*, **32**, e44.
15. Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.
16. Man,T.K., Yang,J.S. and Stormo,G.D. (2004) Quantitative modeling of DNA–protein interactions: effects of amino acid substitutions on binding specificity of the Mnt repressor. *Nucleic Acids Res.*, **32**, 4026–4032.
17. Liu,J. and Stormo,G.D. (2005) Quantitative analysis of EGR proteins binding to DNA: assessing the additivity in both the binding site and the protein. *BMC Bioinformatics*, **6**, 176.
18. Workman,C., Yin,Y., Corcoran,D., Ideker,T., Stormo,G. and Benos,P. (2005) EnoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.*, **33**, W389–W392.
19. Swirnoff,A.H. and Milbrandt,J. (1995) DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell Biol.*, **15**, 2275–2287.
20. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.
21. Stormo,G.D., Schneider,T.D. and Gold,L. (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.*, **14**, 6661–6679.
22. Miller,J.C. and Pabo,C.O. (2001) Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *J. Mol. Biol.*, **313**, 309–315.
23. Jamieson,A.C., Wang,H. and Kim,S.-H. (1996) A zinc finger directory for high-affinity DNA recognition. *Proc. Natl Acad. Sci. USA*, **93**, 12834–12839.
24. Hamilton,T.B., Borel,F. and Romaniuk,P.J. (1998) Comparison of the DNA binding characteristics of the related zinc finger proteins WT1 and EGR1. *Biochemistry*, **37**, 2051–2058.
25. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Is there a code for protein–DNA recognition? Probab(ilistical)ly. . . *Bioessays*, **24**, 466–475.
26. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.