Co-SELECT reveals sequence non-specific contribution of DNA shape to transcription factor binding in vitro

Soumitra Pal*, Jan Hoinka and Teresa M. Przytycka*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received October 02, 2018; Revised May 31, 2019; Editorial Decision June 04, 2019; Accepted June 06, 2019

ABSTRACT

Understanding the principles of DNA binding by transcription factors (TFs) is of primary importance for studying gene regulation. Recently, several lines of evidence suggested that both DNA sequence and shape contribute to TF binding. However, the following compelling question is yet to be considered: in the absence of any sequence similarity to the binding motif, can DNA shape still increase binding probability? To address this challenge, we developed Co-SELECT, a computational approach to analyze the results of in vitro HT-SELEX experiments for TF–DNA binding. Specifically, Co-SELECT leverages the presence of motif-free sequences in late HT-SELEX rounds and their enrichment in weak binders allows Co-SELECT to detect an evidence for the role of DNA shape features in TF binding. Our approach revealed that, even in the absence of the sequence motif, TFs have propensity to bind to DNA molecules of the shape consistent with the motif specific binding. This provides the first direct evidence that shape features that accompany the preferred sequence motifs also bestow an advantage for weak, sequence non-specific binding.

INTRODUCTION

Characterizing the DNA binding specificity of transcription factors (TFs) is a fundamental step in studying gene regulation. It is well established that transcription factors typically bind to specific sequence motifs (1,2). An increasing amount of studies additionally suggest that the nucleotides adjacent to the binding motifs also contribute to the TF–target interactions (3,4). Since these residues do not make contact with the TF-binding site, it is assumed that their contribution to the binding specificity is indirect. Recently, this contribution has been proposed to be achieved by influencing the structural properties of the DNA molecule such as the minor groove width, roll, propeller twist and helical twist (5–8).

Several lines of evidence indicate that DNA structure is important for TF binding. First, there are examples of co-crystal structures that show complementarity of 3D structures of TF and DNA (6,9–11). However, the number of such co-crystal structures is relatively limited and conformational changes upon binding might also contribute to this complementarity. Next, DNA shape has been found to possess significant effects on the electrostatic potentials near the surface of the molecules, particularly in the grooves (5). Indeed, recent computational analyses revealed correlation between minor-groove width and electrostatic potential, although identity of nucleotides also plays some role in the manifestation of the potential (12). As further support, using hydroxyl radical cleavage as a measure of local DNA structure, a computational study (13) on the ENCODE regions of human genome found common DNA structural motifs apparently without any sequence consensus in a large collection of DNase I hypersensitive sites that are closely associated with the regulatory elements. In addition, a recent study that considered a family of TFs and analyzed the DNA shape of thousands of their binding sites demonstrated covariation between the protein sequence and both the sequence and shape of their DNA targets (14). Finally, several computational analyses used machine learning approaches to show that using shape features improves prediction of TF binding both in vitro (15–18) and in vivo (11,16,17). Similarly, (19) developed a method to identify shape motifs from in vivo binding data.

Despite significant effort, evidence for a role of shape for TF binding is mostly indirect. One of the principal challenges in measuring this effect directly is related to the fact that DNA shape is primarily determined by its underlying sequence. This interdependency in turn substantially aggravates the task of disambiguating the individual contribution of sequence and structure. Indeed, computational inference of DNA shape features is based on sequence infor-
mation alone and thus computationally inferred shape features could be seen as functions summarizing sequence information of the neighboring residues (20,21). Consistent with this view, machine learning models for TF-DNA binding have been reported to show a similar improvement when features based on two or three consecutive nucleotides in lieu of shape features were added (15,18,22,23). In particular, a recent study that showed sequence-to-shape conversion can be estimated nearly perfectly based on mononucleotide and dinucleotide features (22,23). In addition, some studies suggest that incorporating DNA shape features into TF-binding models provides only moderate improvements for a limited number of TFs (24). However, sequence models are agnostic with respect to the mechanisms of specificity. In particular they are oblivious to the fact that not only sequence preferences can influence the shape but also shape preferences can dictate sequence preference.

These considerations raise the question whether shape can be deconvoluted from sequence so that the contribution of shape alone could be verified. In light of the above findings, an independent shape contribution could only be assessed by testing for binding to a DNA molecule that is completely free of any similarity to the sequence binding motif but does contain the shape corresponding to the motif sequence. To obtain such an evidence, we re-examined the data obtained through a collection of HT-SELEX experiments on TF binding (18) by using a novel approach that we developed for this purpose and referred as Co-SELECT.

HT-SELEX has been extensively used to study in vitro RNA and DNA binding. This experimental technique leverages the paradigm of in vitro selection by repetitively enriching a pool of initially random RNA/DNA sequences with those that bind a target of interest (25–27). At the end of the selection process, the sequences that are bound to the target are expected to be present in multiple copies, while remaining sequences consist of weak binders and/or noise and are typically excluded from further analysis or broadly considered as ‘background’.

In the context of our study, HT-SELEX data provide several unique advantages. First, selection is performed in vitro and thus binding events are not influenced by other complex DNA transactions (such as transcription, replication etc.) that take place in vivo. For example, gene transcription is typically accompanied by negative supercoiling of the DNA molecule upstream of the transcription start site which, in turn, might critically influence DNA shape (13,28,29). Next, while the initial pool of DNA fragments is assumed to contain random DNA sequences, intermediate pools are increasingly non-random as the strongest binders are amplified and non-binders are eliminated. Weaker binders would eventually be out-competed given a sufficient number of selection rounds. These species however can survive the selection process and are expected to be enriched among the background sequences since the number of selection rounds of SELEX targeting TFs is typically low.

To leverage the properties of in vitro selection for the purpose of investigating the possibility of selection for shape features, a novel approach capable of capturing such features is needed. Here, we introduce such an approach that utilizes the concept of a shapemer. A shapemer is similar to that of a k-mer but rather than being a sequence of nucleotides, a shapemer represents a sequence of shape features at nucleotide resolution. Based on this concept, we developed Co-SELECT—a method for detecting shapemers enriched in both oligos that contain binding motifs and the background oligos that survived until the final selection round. Note that unlike previous studies that aimed at addressing the question of whether computationally derived shape features of the binding motif provide additional information relative to the sequence alone, we ask if in the absence of any sequence similarity to the binding motif, can shape alone provide a binding advantage? Our results demonstrate that this is indeed the case and additionally identify promiscuous shapemers that appear to be favored in non-specific TF binding.

**MATERIALS AND METHODS**

**Data**

In this study, we utilize the HT-SELEX datasets from (18) that are derived from the experiments in Jolma et al. (27) and complemented with additional sequencing data. To the best of our knowledge, this dataset is currently the most extensive mammalian TF-DNA binding dataset derived using HT-SELEX (18). Here we are focusing on the following three families of transcription factors: (i) basic helix-loop-helix (bHLH), (ii) E26 transformation-specific (ETS) and (iii) homeodomain. Our choice of the three families is motivated by the fact that these families are assumed to have a uniquely defined, very highly conserved ‘core-motif’. Our analysis depends on such well defined core-motifs for a clear classification of oligos into two contrasting groups: motif-free and motif-containing depending on the presence and completeness of the core-motif as described below. The default core-motifs for the three families are CACGTG, CG-GAA and TAAT, respectively (18). However, for each TF we also tested whether the selection results are consistent with the default core-motif, and if not, an alternative motif of the same length derived in (27) from the HT-SELEX results was also considered (Supplementary Section S1).

Finally, Jolma et al. (30) observed that some of their experiments show evidence of cross contamination artifacts or indication of an unsuccessful selection. Consequently, we refined our dataset by removing experiments with the above-described signatures from further analysis based on the general strategy outlined in that paper. Since our method relies on data from experiments that generated a strong core-motif, we discarded those selections that yielded ambiguous binding motifs (Supplementary Section S1). We further excluded all experiments where the first/last two nucleotides of the barcode used in the HT-SELEX experiment are overlapping with the core-motif. This ensures that in none of the analyzed experiments, an oligo is treated as motif-free while the barcode flanking the random sequence in the oligo contains a partial-motif. After filtering, out of the initial set of 212 datasets, 83 high quality datasets remained that are comprised of 9 experiments in the bHLH family, 10 experiments in the ETS family and 64 experiments in the homeodomain family (see Table S1 for more details).
The HT-SELEX procedure and classification of oligos with respect to motif presence

Our approach of analyzing the contribution of DNA shape on TF binding leverages the HT-SELEX (High-Throughput Systematic Evolution of Ligands by EXponential enrichment) protocol and resulting sequencing data. Since the basic properties of the HT-SELEX protocol are fundamental to our method, we start with a brief summary of the relevant features of HT-SELEX. HT-SELEX is an iterative procedure that starts with an initial pool of random DNA sequences (also referred to as ‘oligonucleotides’ or ‘oligos’) that are typically 20–40 nucleotides in length.

Each iteration of HT-SELEX can be viewed as a competition among the oligos for binding to the TF. Oligos that do not bind at all or bind weakly are washed out from the pool while the remaining species are amplified. A sample of the amplified pool is sequenced to allow for computational analysis while another sample is used as the input for the subsequent selection round. In this way, the proportion of high-affinity oligos in the pool increases at each iteration while non-binders and, subsequently, weaker binders are gradually eliminated/out-competed. We refer to oligos in the final round as ‘aptamers’ as they are expected to be the true binders for the target TF.

We say that an oligo is (i) ‘motif-containing’ if it contains the exact core-motif sequence, (ii) ‘motif-free’ if it does not share more than 2 nucleotides with the core-motif at any position and (iii) ‘partial-motif-containing’ otherwise. The threshold 2 was selected as the largest value such that a random oligo of length $n$ has, with probability $\geq 0.9$, at least one $k$-mer having $d$ nucleotides common with the core-motif of length $k$. For the smallest length of an oligo in our datasets ($n = 20$) and the shortest core-motif ($k = 4$), the probability of overlap on $d = 1, 2, 3$ positions are estimated as 0.99, 0.98, 0.56, respectively (Supplementary Section S2). Thus, the threshold $d = 2$ (but not $d = 3$) for the maximum number of bases matching with the core-motif ensures that a motif-free oligo can be safely assumed to carry no sequence information of the core-motif.

Importantly, analysis of the sequencing data from consecutive iterations can provide key information about the dynamics of this competition as illustrated in Figure 1. Using simulated data generated with AptaSim (31), Figure 1A shows how the proportions of five groups of oligos with varying affinity to the target change throughout the selection process. Initially, the proportion of medium and high affinity binders increases at the expense of weak binders and non-binders. In later rounds, however, the proportion of medium binders starts to decrease indicating that they are now out-competed by stronger binders. At this point, the proportion of non-binders is significantly reduced and quickly becomes negligible.

These expected trends are indeed observed in experimental data as exemplified by the selection results performed against the transcription factor myc-associated factor X (MAX) in Figure 1B. The figure demonstrates a rapid increase of the population of oligos that contain the complete core-motif, consistent with the expectation that this group contains the strongest binders. The populations of oligos that contain partial-motifs with one mismatch from the core-motif initially increase as expected from medium strength binders while the populations of oligos containing partial-motifs with 2 or 3 mismatches and the motif-free oligos are out-competed. Given four rounds of selection, even the oligos in the motif-free group that survive this competition are expected to be enriched in (presumably weak) binders.

Discretized shape strings and shapemers

We use the term ‘shapemers’ to describe constant length sequences of discretized shape values for a specific DNA feature such minor groove width (MGW), propeller twist (ProT), helical twist (HelT) and Roll. We utilized DNAShape (21), which uses a sliding-window approach to estimate DNA shape at each position of a given a input sequence. For the purpose of this study, we discretized the shape values using cutoffs based on the frequency distribution of the shape values in the oligos from the initial pools (Supplementary Section S3). Two different sets of discretization thresholds were examined to confirm that the conclusions drawn do not depend on these threshold values. We used four discretization levels for MGW, HelT and Roll denoted as S (Small), M (Medium), H (High), X (eXtra high) and three levels for ProT (S, M, H) (Figure 2 and Supplementary Section S3). In this study, we used shapemers of length 6. This length was selected as it corresponds to the length of the longest core-motif considered in this study.

We analyzed the shapemers from the two groups of oligos, i.e. motif-containing and motif-free, separately. While the shapemers from the motif-containing oligos carry the sequence information of the core-motif, the shapemers from the motif-free oligos are devoid of such information.

Indeed, we confirmed that shapemers enriched within the subpopulation of motif-free aptamers do not share sequence similarity with the core-motif (Supplementary Section S4).

Computing enrichment of shapemers

In this paper, we use fold enrichment to denote the ratio between the frequency of a particular feature in the final SELEX pool to its frequency in the initial pool. Here, we focus on the enrichment of shapemers and use a Markov model to estimate the expected presence of shapemers in the initial pool. To analyze the enrichment of shapemers in motif-specific binding, we consider only the shapemers that are contained in the interval consisting of the core-motif and a flanking nucleotide on either side (Figure 3). For motif-free binding, we do not have any prior assumption that shapemer might be involved in binding and thus we consider all shapemers contained in the aptamer. We denote the shapemers thus identified from the motif-containing and motif-free oligos as ‘core shapemers’ and ‘motif-free shapemers’, respectively.

Let $j \in \{m, n\}$ index the two sets of shapemers of type core and motif-free, respectively. We compute $R_j(x)$, the enrichment of shapemer $x$ in set $j$, as

$$R_j(x) = \frac{P_{\ell,j}(x)}{P_{0,j}(x)} \quad \text{for all} \quad j \in \{m, n\},$$

(1)
where \( P_{i,j} \) denotes the probability of a shapemer \( x \) appearing in set \( j \) and round \( i \). We compute the probabilities for the final round \( f \) as

\[
P_{f,j}(x) = \frac{H_{f,j}(x)}{\sum_x H_{f,j}(x)} \quad \text{for all} \quad j \in \{m, n\},
\]

where \( H_{f,j}(x) \) denotes the number of times shapemer \( x \) appears in set \( j \) and round \( f \).

We estimate the probabilities \( P_{0,j}(x) \) for the initial pool using a Markov model, the details of which are given in Supplementary Section S5. We discard the shapemers \( x \) with \( R_j(x) < \rho \) where \( \rho \) denotes the ‘enrichment threshold’. We experimented with three values of \( \rho = \{1.1, 1.2, 1.5\} \) and report the results for \( \rho = \{1.1, 1.2\} \) since enrichment \( \geq 1.5 \) was very rare. To reduce noise, we additionally only consider shapemers that are present in at least 1% of aptamers (Supplementary Section S6).

Statistical analysis

We assess the statistical significance of the overlap of enriched motif specific and motif-free shapemers in the final round of selection using the one-sided Fisher’s exact test with \( P \)-value cutoff 0.05. To correct for multiple testing, we used the Benjamini-Hochberg (BH) procedure (32). To compare with the control experiment (see ‘Results’ section), we use \( \text{ERCP} \) defined as the fraction of the statistically significant experimental results with \( P \)-value cutoff \( p \) to the fraction of the control results with the same \( P \)-value cutoff. To avoid division by zero, we add a pseudo-count of \( 10^{-5} \) to the fraction for control groups.

RESULTS

The Co-SELECT method for detecting shapemers present in motif-containing and motif-free aptamers

We reasoned that if the DNA shape contributes to TF binding in a sequence independent way, then after several rounds of selection we should observe enrichment of the same shape features in both motif-specific and non-motif-specific binding. With this in mind, we developed our Co-SELECT method to test for such simultaneous selection for the same shape features in two extreme subpopulations of oligos from the same HT-SELEX experiment: the first subpopulation corresponds to the set of aptamers that contain the core binding motif without any mismatches and the second subpopulation encompasses aptamers that have no similarity to the core-motif. To describe shape features we use ‘shapemers’—constant length sequences of dis-
Figure 3. Overview of the Co-SELECT method. (A) First, motif-containing (blue) and motif-free (red) aptamers are identified within the sequence pool from the final round of selection. The remaining shapemers that might contain partial-motif are disregarded. Next, the shapemers that potentially contribute to binding in the motif-containing group (core shapemers) and in the motif-free aptamers are identified. (B) Significant overlap of enriched core and motif-free shapemers suggests that the same shape feature contribute to motif specific and motif-free binding.

Co-SELECT identifies DNA shapemers that contribute to the strength of motif-specific and non-motif specific binding

We applied Co-SELECT to the HT-SELEX data that passed the quality control described in the ‘Materials and Methods’ section focusing on three large families of transcription factors: ETS, bHLH and homeodomain. These three families are structurally different and contain a strongly conserved core-motif whose length differs across the families. Overall, we observed statistically significant selection for shape for some members in each family. The results are summarized in Table 1.

Supplementary Table S2 lists, for each of the 4 shape features in each of the 83 analyzed TF experiments, the shapemers co-selected in motif-containing and motif-free pools. Many of the co-selected shapemers are TF family specific. In addition, Supplementary Figures S4–S23 show the sequence logos of the co-enriched MGW shapemers in motif-containing and motif-free pools in all selection cycles. It can be appreciated that, consistent with the selection for shape rather than for sequence, these sequence logos generally remain constant over the selection cycles.

As described in ‘Materials and Methods’ section, we opted for a highly stringent separation between motif-containing and motif-free pools in Co-SELECT analysis producing a more conservative results of Table 1. To test if a less stringent separation would increase the number of co-selected shapemers, we repeated the Co-SELECT analysis in a modified setting where the motif-containing pool include those oligos that contain any k-mer with the Hamming distance at most 1 from the core-motif. As expected, we observed that the number of statistically significant experiments increased, thus corroborating our results (see Supplementary Section S8 for details).

As mentioned in ‘Materials and Methods’ section, Co-SELECT uses Fisher’s exact test to assess the significance of the overlap between enriched core shapemers and motif-free shapemers from the same experiment. It is theoretically possible that some shapemers are generally enriched in motif-free groups independently of the targeted transcription factor. If such shapemers exists they could inflate statistical significance of the overlap. To exclude this possibility, we performed an additional control by applying Co-SELECT after swapping the subpopulations of motif-containing and motif-free aptamers between pairs of transcription factors (Figure 4). Specifically, let $TF_m$ and $TF_{nm}$ denote, respectively, the subpopulation of motif-containing and motif-free...
Figure 4. The design of control Co-SELECT experiments. Given two transcription factors (TF1, TF2) from two different families we test for the significance of the overlap of enriched core shapemers from the selection experiment for TF1 and enriched motif-free shapemers from the selection experiment for TF2. We expect that the number of co-selected shapemers in such a control experiment is smaller than in the corresponding true experiment.

Figure 5. Comparison of P-value histograms for the original experiment (dark brown) and the control experiments (light brown). $ECR_{0.05} > 1$ indicates that the number of results with $P$-value below 0.05 was higher for the original experiment than control. We note that while for MGW in homeodomain family $ECR_{0.05} < 1$, for very small $P$-values $<0.005$, $ECR_{0.005} > 1$. 
Table 1. Summary of Co-SELECT results

| Family       | #TFs | FDR cutoff | MGW 0.05 | MGW 0.10 | HeiT 0.05 | HeiT 0.10 | ProT 0.05 | ProT 0.10 | Roll 0.05 | Roll 0.10 |
|--------------|------|------------|----------|----------|-----------|-----------|-----------|-----------|-----------|-----------|
| bHLH         | 9    | 1.2        | 55.56    | 55.56    | 44.44     | 55.56     | 0.0       | 0.0       | 0.0       | 0.0       |
|              |      | 1.1        | 55.56    | 66.67    | 44.44     | 66.67     | 0.0       | 11.11     | 0.00      | 0.00      |
| ETS          | 10   | 1.2        | 20.00    | 20.00    | 0.0        | 0.0       | 0.0       | 0.0       | 0.0       | 0.0       |
|              |      | 1.1        | 30.00    | 40.00    | 0.0        | 0.0       | 0.0       | 0.0       | 0.0       | 0.0       |
| Homeodomain  | 64   | 1.2        | 1.54     | 1.54     | 13.85      | 20.00     | 0.0       | 0.0       | 0.0       | 0.0       |
|              |      | 1.1        | 4.62     | 6.15     | 38.46      | 47.69     | 0.0       | 0.0       | 4.62      | 4.62      |

For each TF family and each shape feature, we report the percentage of TF experiments with statistically significant overlap of shapemers in motif containing and motif free pools. *P*-values are BH corrected (32). We include two FDR cutoff values \{0.05, 0.10\} and two threshold values for shapemer enrichment \(\rho = \{1.2, 1.1\}\).

**Figure 6.** Differential shape preferences of two homeodomain proteins: PITX3 and HMX2. (A) Crystal structures of PITX3 bound to DNA. (B) MGW shape preference of PITX3. (C) From top to bottom: logos of motif-containing sequences (including flanking regions) in the initial pool, logos of motif-containing sequences (including flanking regions) in the initial pool that contain the most-enriched motif-free shapemer, sequence logos of the most enriched motif-free shapemer in the initial and final pools. The letters in the box below the logos show the corresponding shapemers (the difference is shown in red). The similarity between the second pair of logos different from the sequence motif logo for each protein indicates that shape rather than sequence contributed to the motif-free enrichment.

**Figure 7.** Principal Component Analysis of shapemers enriched in motif-free pool. Here each point corresponds to a TF and is color-coded based primarily on the binding domain and then on the core of the binding motif. APTX3 and HMX2 aptamers from the last selection round for a transcription factor TF. We tested for co-selection of shapemers in TF\(_1\) and TF\(_2\), where TF\(_1\) and TF\(_2\) are transcription factors from two different transcription factor families. We used all \(\binom{2}{1} = 6\) possible pairs of families and all pairs of TFs within the respective pair of families.

To compare the results of Co-SELECT from the original selection experiments to the so designed controls, we compared the *P*-value distributions for each TF family and shape (Figure 5) using the experiment to control ratio \(ECR_p\) at *P*-value \(P\) (see ‘Materials and Methods’ section). The TF families and shapes that are found to have positive \(ECR_{0.05}\) values are consistent with those which we reported to yield statistically significant *P*-values as summarized in (Table 1). These results provide evidence that the co-selection for shape is not an artifact of the experimental procedure although some promiscuous shapemers can exist that are enriched in the motif-free group across TFs and TF-families.
Co-SELECT identifies different shape preferences for two TFs with the same sequence preference

Despite the noticeably lower number of significant results for the homeodomain family, we identified several highly significant results that warranted further exploration. These include Pituitary Homeobox 3 Protein (PITX3) ($P$-value $\leq 0.00355$) and Homeobox Protein HMX2 ($P$-value $\leq 0.00051$) both recognizing TAAT as the core-motif but with different shape preferences. Next, we leveraged the availability of a crystal structure of the protein PITX3 bound to DNA in the PDB database (ID:2LKK) (Figure 6B). Previous studies demonstrated that in this crystal structure, the computationally inferred DNA shape (20) is in good agreement with the experimental data (18). In addition, machine learning based methods demonstrated that, in the case of this domain, inclusion of shape features improves the model thus suggesting that shape is important for binding. Consistent with the experimental data, in both the motif-containing and motif-free sequences, a H(XHHM)H shapemer for MGW has been selected (bracket indicates the discretized, preferred MGW of the core-motif), also corroborated by protein-binding microarray (PBM) data (see Supplementary Section S11 for details). This provides yet the most direct evidence that shape is important for binding of PITX3.

Interestingly, based on the DNA shape analysis (20) the preferred shape for the core-motifs for HMX2 is different than the preferred shape for PITX3 (Figure 6C and Supplementary Figure S4). This difference is consistent with the different preferences in the core-motif flanking residues for the two proteins. This demonstrates that PITX3 and HMX2 show consistent differential shape preferences not only for the motif-containing, but also for the motif-free oligos.

Shapemers enriched in motif-free oligos are TF-family dependent

It has been observed that TF families approximately cluster by the shape features of their binding motifs (18). Thus, we asked the question whether TF families cluster by enriched motif-free shapemers. We applied Principal Component Analysis to a binary matrix in which the feature vectors are all the shapemers $x$ that meet the following two criteria: (i) $x$ is enriched in the motif-free pool in any one of the TF experiments and (ii) in all TF experiments, no more than 30% of the occurrences of $x$ correspond to a single sequence $k$-mer. Interestingly, we also found that TFs group based on enriched motif-free shapemers, although the clusters were not perfect. (Figure 7). The bHLH family is partitioned into subgroups depending on the binding motif, including a subgroup with the CATATG motif that clustered together with homeodomain. This co-clustering is presumably due to the importance of its AT-rich center similar to the homeodomain core-motifs. This result further reinforces the notion that shape features enriched in the motif-free pool are defined by TF binding preferences.
Promiscuity of motif-free shapemers

Next, we asked whether there exist shapemers that are generally enriched in motif-free groups independently of the targeted TF. If such promiscuous shapemers exist, they could provide information about a possible mechanism for non-specific background binding to DNA. To measure promiscuity of a shapemer we compute, for each family, the fraction of TFs where the shapemer is enriched in the motif-free group. The promiscuity of the shapemer is then defined to be the minimum of the so computed fractions over all families. Figure 8 elucidates the shapemers sorted by their promiscuity measures with the identities of the most promiscuous shapemers shown in the insets.

To better understand the source of the observed promiscuity, we considered the fact that the mapping between DNA sequence k-mers and shapemers is many-to-one and that the number of sequence k-mers mapping to a given shapemer varies between shapemers. To test if the promiscuity is not simply the result of a large number of sequence k-mers mapping to a given shapemer, we performed a permutation test. Specifically, for each shapemer x we first computed the number of sequence k-mers N that map to shapemer x and then randomly selected N sequence k-mers from those that do not map to shapemer x. To compute a P-value we repeated this step 1000 times. We found that the existence of the promiscuous shapemers as shown in the inset of Figure 8 cannot be explained by a large number of sequence k-mers mapping to a given shapemer, and hence are statistically significant with a P-value <10−3.

DISCUSSION

Several previous studies provided evidence that including shape as one of the features in computational approaches that predict binding affinity given DNA sequence improves the results of such prediction (11,15–17,19). These results suggest that DNA shape is important for TF binding. However, the question if DNA shape, in the absence of any sequence similarity to the binding motif, can still increase the probability of binding was yet to be addressed.

The Co-SELECT approach is carefully designed to examine this question. We focused on TFs with very strong and well defined core-motifs allowing us to confidently identify a motif-free population of aptamers. Using two different tests, we demonstrated a robust motif-independent contribution of shape-dependent binding. It is possible that for transcription factors with less conserved binding motif, shape might have an even higher contribution to binding. For example, conservation for shape might dictate substitution rules in sequence binding motifs. It is also important to point out that our stringent criterion for calling oligos to be motif-free acutely constrains the set of aptamers that are in this category. While allowing for rigorous analysis, it could reduce its statistical power. This opens the possibility that the impact of shape for motif-free binding not only exists but is even more prominent than estimated in this work.

Interestingly, we identified promiscuous shapemers. Such shapemers might also facilitate weak, non-specific binding. Focusing on MGW, we note that the promiscuously preferred shapes are biased toward the H (high) category. Since GC rich sequences such as GC repeats and G-runs often assume DNA shape with MGW in H category, the possibility of potential artifacts of the HT-SELEX procedure related to GC-bias in amplification or sequencing cannot be completely ruled out. Yet, a recent paper demonstrated that increasing CpG density enriches for bound motifs (33). In addition, the top promiscuous sequences did not show significant enrichment of Gs (Supplementary Figure S26). Interestingly, these shape values occur frequently in the relaxed DNA as seen in the frequency distributions of shape values in Figure 2. This suggests that TF-DNA binding might have evolved to have an affinity to non-specifically bind DNA in its most abundant (favored by many sequences) shape. Such non-specific affinity to the generic DNA shape might facilitate diffusion/sliding of a TF along the DNA molecule as proposed in several theoretical models of TF-DNA binding (34–39).

DATA AVAILABILITY

The raw sequencing data from the HT-SELEX experiments are available in the ENA (http://www.ebi.ac.uk/ena) under study identifier PRJEB14744 (for the newer experiments (18) used in this analysis) and under accession numbers ERP001824 and ERP001826 (for the original experiments (27) but not used in this analysis). The source code for Co-SELECT is available on the public Github repository https://github.com/ncbi/Co-SELECT.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This research was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Heath, USA.

Conflict of interest statement. None declared.

REFERENCES

1. Stormo,G.D. (2000) DNA binding sites: representation and discovery. Bioinformatics, 16, 16–23.
2. Hippe1,H.V. and Berge.O.G. (1986) On the specificity of DNA-protein interactions. Proc. Natl. Acad. Sci. U.S.A., 83, 1608–1612.
3. Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X. et al. (2009) Diversity and complexity in DNA recognition by transcription factors. Science, 324, 1720–1723.
4. Nutiu, R., Friedman, R.C., Luo, S., Khebtukovia, L., Silva, D., Li, R., Zhang, L., Schroth, G.P. and Burge, C.B. (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat. Biotechnol., 29, 659–664.
5. Rohs, R., West, S.M., Sosinska, Y., Liu, P., Mann, R.S. and Honig, B. (2009) The role of DNA shape in protein-DNA recognition. Nature, 461, 1248–1253.
6. Stella, S., Cascio, D. and Johnson, R.C. (2010) The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes Dev., 24, 814–826.
7. Afeke, A., Schipper, J.L., Horton, J., Gordn, R. and Lukatsky, D.B. (2014) Protein-DNA binding in the absence of specific base-pair recognition. Proc. Natl. Acad. Sci. U.S.A., 111, 17140–17145.
8. Dror,I., Golan,T., Levy,C., Rohs,R. and Mandel-Gutfreund,Y. (2015) A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.*, **25**, 1268–1280.

9. Jones,S., van Heyningen,P., Berman,H.M. and Thornton,J.M. (1999) Protein-DNA interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–896.

10. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2017) Origins of specificity in Protein-DNA recognition. *Ann. Rev. Biochem.*, **79**, 233–269.

11. Mathelier,A., Xin,B., Chiu,T.-P., Yang,L., Rohs,R. and Wasserman,W.W. (2016) DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.*, **3**, 278–286.

12. Chiu,T.-P., Rao,S., Mann,R.S., Honig,B. and Rohs,R. (2017) Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-DNA binding. *Nucleic Acids Res.*, **45**, 12565–12576.

13. Garvie,C.W. and Wolberger,C. (2001) Recognition of specific DNA readout revealed by quantitative specificity models. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 10497–10500.

14. Dror,I.,Golan,T.,Levy,C.,Rohs,R.,Mandel-Gutfreund,Y.(2015) Optimization using matrix and DNA shape-based models. *Bioinformatics*, **31**, 554–563.

15. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordan,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 4654–4659.

16. Abe,N., Dror,I., Yang,L., Slattery,M., Zhou,T., Bussemaker,H.J., Rohs,R. and Mann,R.S. (2015) Deconvolving the Recognition of DNA Shape from Sequence. *Cell*, **161**, 307–318.

17. Ma,W., Yang,L., Rohs,R. and Noble,W.S. (2017) DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. *Bioinformatics*, **33**, 3003–3010.

18. Yang,L., Orenstein,Y., Jolma,A., Yin,Y., Taipale,J., Shamir,R. and Rohs,R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.*, **13**, 910.

19. Samee,M.A.H., Bruneau,B.G. and Pollard,K.S. (2019) A de novo shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst.*, **8**, 27–42.

20. Zhou,T., Yang,L., Lu,Y., Dror,I., Machado,D., Carolina,A., Ghane,T., Di Felice,R. and Rohs,R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56–W62.

21. Chiu,T.-P., Comoglio,F., Zhou,T., Yang,L., Paro,R. and Rohs,R. (2016) DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, **32**, 1211–1213.

22. Rube,H.T., Rastogi,C., Kribelbauer,J.F. and Bussemaker,H.J. (2018) A unified approach for quantifying and interpreting DNA shape readout by transcription factors. *Mol. Syst. Biol.*, **14**, e9702.

23. Le,D.D., Shimko,T.C., Aditham,A.K., Keys,A.M., Longwell,S.A., Orenstein,Y. and Fordyce,P.M. (2018) Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E3702–E3711.

24. Ruan,S. and Stormo,G.D. (2018) Comparison of discriminative motif optimization using matrix and DNA shape-based models. *BMC Bioinformatics*, **19**, 86.

25. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.

26. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.

27. Wolberger,C., Jolma,A., Yan,J., Whittington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. et al. (2013) DNA Binding specificities of human transcription factors. *Cell*, **152**, 327–339.

28. Irobalieva,R.N., Fogg,J.M., Daniel,J.C., Jr, Sutthibutpong,T., Abe,N., Chiu,T., Burger,L. and Schbeler,D. (2019) CG dinucleotides enhance promoter activity independent of DNA methylation. *CellSyst.*, **9**, 353–363.

29. Gerasimova,N.S., Pestov,N.A., Kualeva,O.I., Clark,D.J. and Studitsky,V.M. (2016) Transcription-induced DNA supercoiling: New roles of intranucleosomal DNA loops in DNA repair and transcription. *Science*, **353**, 91–95.

30. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan.J and Sillanp,M.J. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.

31. Hoinaka,J., Berezhnoy,A., Dao,P., Sauna,Z.E., Gilboa,E. and Przytycka,T.M. (2015) Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Res.*, **43**, 5699–5707.

32. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery Rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B (Methodological)*, **57**, 289–300.

33. Haril,D., Krebs,A.R., Grand,R.S., Baubec,T., Isbel,L., Wirbelauer,C., Burger,L. and Schbeler,D. (2019) CG dinucleotides enhance promoter activity independent of DNA methylation. *Genome Res.*, **29**, 554–563.

34. Winter,R.B. and Von Hippel,P.H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli lac repressor–operator interaction: equilibrium measurements. *Biochemistry*, **20**, 6948–6960.

35. Berg,O.G., Winter,R.B. and von Hippel,P.H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, **20**, 6929–6948.

36. Winter,R.B., Berg,O.G. and von Hippel,P.H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor–operator interaction: kinetic measurements and conclusions. *Biochemistry*, **20**, 6961–6977.

37. Halford,S.E. and Marko,J.F. (2004) How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.*, **32**, 3040–3052.

38. Marklund,E.G., Mahmutovic,A., Berg,O.G., Hammarn,P., Spool,D.V.D., Fanghe,D. and ElJ. (2013) Transcription-factor binding and sliding on DNA studied using micro- and macro-scopic models. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 19796–19801.

39. Mechetin,G.V. and Zharkov,D.O. (2014) Mechanisms of diffusional mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, **57**, 289–300.