Searching for sequence features that control DNA flexibility

Yaojun Zhang, 1 Aakash Basu, 2 Taekjip Ha, 2, 3, 4, 5 and William Bialek 1, 6

1 Joseph Henry Laboratories of Physics and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544
2 Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205
3 Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218
4 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205
5 Howard Hughes Medical Institute, Baltimore, MD 21205
6 Initiative for the Theoretical Sciences, The Graduate Center, City University of New York, 365 Fifth Ave, New York, NY 10016

(Dated: December 14, 2020)

Modern genomics experiments measure functional behaviors for many thousands of DNA sequences. We suggest that, especially when these sequences are chosen at random, it is natural to compute correlation functions between sequences and measured behaviors. In simple models for the dependence of DNA flexibility on sequence, for example, correlation functions can be interpreted directly as interaction parameters. Analysis of recent experiments shows that this is surprisingly effective, leading directly to extraction of distinct features for DNA flexibility and predictions that are as accurate as more complex models. This approach follows the conventional use of correlation functions in statistical physics and connects the search for relevant DNA sequence features to the search for relevant stimulus features to the analysis of sensory neurons.

In physics we often use correlation functions to characterize the behavior of a system, and many experimentally measurable quantities are related directly to these correlation functions. As examples, the diffusion constant of a particle is an integral over the correlation function of its velocity, the X-ray diffraction pattern of a material is the Fourier transform of the correlation function of density fluctuations [1], and scattering amplitudes for elementary particles are correlation functions in the underlying quantum field theory that describes their interactions [2]. It has taken longer for this language of correlation functions to permeate the analysis of living systems.

In analyzing how single neurons respond to their inputs, it is conventional to compute the correlation between the continuous inputs and the discrete sequence of action potentials or spikes at the output [3, 4]; this “triggered correlation” seems to have been inspired more by ideas of systems identification in engineering than correlation functions in physics [5]. It eventually was realized that this approach could be generalized to higher order correlations, allowing the identification of multiple relevant input features in triggering a spike [6, 7]. In these applications, it is important that the inputs can be chosen from appropriate ensembles. More recently, correlation functions have emerged as central to the analysis of collective behavior in animal groups, much in the original spirit of their use to analyze experiments in condensed matter [8]. Here we consider the use of correlation functions to analyze experiments on the mechanics of randomly chosen DNA sequences [9].

The key step in using correlation functions to analyze neural responses was to shift from measuring responses to particular, carefully chosen sensory stimuli [10] to an unbiased exploration of many more stimuli chosen randomly from some well understood distribution. As an example, if a neuron integrates for ~ 100 msec, then recording neural activity in response to one hour of continuous random inputs is equivalent to sampling ~ 3 X 10^4 different stimuli. Long before the genomic revolution brought the term into common use, this approach thus achieved “high throughput.”

To make the discussion concrete, we consider DNA sequences \{S_i^\alpha\}, where S_i^\alpha = 1 if the base at site i is of type \alpha, and S_i^\alpha = 0 otherwise. The index i = 1, 2, · · · , N, where N is the length of the sequences we are studying, and \alpha = 1, 2, 3, 4, corresponding to A, T, C, G. If we choose sequences at random from the uniform distribution, we have \langle S_i^\alpha \rangle = 1/4 and

\[
\langle S_i^\alpha S_j^\beta \rangle = \delta_{ij} \delta_\alpha^\beta (1/4) + (1 - \delta_{ij})(1/4)^2 \tag{1}
\]

which means that the connected correlations are

\[
\langle S_i^\alpha S_j^\beta \rangle_c = \langle S_i^\alpha S_j^\beta \rangle - \langle S_i^\alpha \rangle \langle S_j^\beta \rangle = (\langle S_i^\alpha - 1/4 \rangle \langle S_j^\beta - 1/4 \rangle) = \delta_{ij}(1/4) (\delta_\alpha^\beta - 1/4) \tag{2}
\]

We can go on to compute higher order correlations, which will be relevant below; details are in Appendix [A].

Recent experiments have chosen \( M = 12,472 \) random sequences from the uniform distribution and estimated the intrinsic flexibility of these sequences by measuring the probability that they close on themselves into a loop [10]. In detail, randomly chosen sequences of length \( N = 50 \) were flanked by fixed double stranded adapters and complementary overhangs, and immobilized on a bead. The looping reaction was initiated by changing solution conditions, and after a fixed time the unlooped molecules were degraded by an enzyme that only attacks free ends. The remaining population of looped molecules
was sequenced and compared with the original ensemble; cyclizability was defined as the log ratio of probabilities for finding sequences in the looped vs control ensembles. Observations on a small number of sequences show that this measure correlates very well with direct measurement of cyclizability. The measured cyclizability was defined as the log ratio of probabilities on sequence is linear, consistent with $W = 0$, the next simplest model is

$$C_0 = \langle C_0 \rangle + \frac{1}{2} \sum_{k\ell} J_{k\ell}^\alpha (S_k^\alpha - 1/4)(S_l^\alpha - 1/4).$$

As shown in Appendix A any site diagonal term $J_{ii}^{\alpha\beta}$ in the matrix $J$ can be rewritten as a weight $W_i^{\alpha\beta}$ in the linear model, so we can set these terms to zero. We also can set $\sum_\beta J_{ij}^{\alpha\beta} = 0$, since $\sum_\alpha S_i^\alpha = 1$. Now we want to compute the correlation function

$$\langle C_0 S_i^\alpha S_j^\beta \rangle_c \equiv \langle (\langle C_0 \rangle - \langle C_0 \rangle^2) (S_i^\alpha - \langle S_i^\alpha \rangle)(S_j^\beta - \langle S_j^\beta \rangle) \rangle,$$

and we find (see Appendix A for details) that

$$\langle C_0 S_i^\alpha S_j^\beta \rangle_c = \frac{1}{16} J_{ij}^{\alpha\beta}.$$ 

As in the case of the linear model, computing correlation functions over random sequences directly recovers the underlying interaction parameters.

We emphasize that this correlation function is a matrix: we can combine the indices $(i, \alpha) \rightarrow \mu$ and $(j, \beta) \rightarrow \nu$ so that $\langle C_0 S_i^\alpha S_j^\beta \rangle_c \rightarrow M_{\mu\nu}$. This construction thus is analogous to the spike–triggered covariance matrix in the analysis of neural responses. We search for further simplification by analyzing eigenvalues and eigenvectors,

$$\langle C_0 S_i^\alpha S_j^\beta \rangle_c = \sum_n \lambda_n w_i^{\alpha}(n) w_j^{\beta}(n);$$

it will be important that eigenvectors are orthonormal,

$$\sum_{i,\alpha} w_i^{\alpha}(n) w_i^{\alpha}(m) = \delta_{nm}. $$

We estimate the correlation function $\langle C_0 S_i^\alpha S_j^\beta \rangle_c$ from the data, and then diagonalize. In Fig 3 (top) we show

FIG. 1: The distribution of the intrinsic cyclizability $C_0$ across the $\sim 10^4$ random sequences in the experiment of Ref [10]. Mean and standard deviation across random halves of the data.

FIG. 2: The connected correlation function $\langle C_0 S_i^\alpha \rangle_c$. Blue points: Mean and standard deviation across random halves of the data. Red lines: ± one standard deviation across random halves of shuffled data.
eigenvalue of triggered covariance

FIG. 3: Eigenvectors of the correlation matrix $\langle C_0 S_i^\alpha S_j^\beta \rangle_c$, Eq (8). (top) Eigenvalues $\lambda_n$ (blue) compared with results from shuffled data (red); points are means and error bars are standard deviations across randomly chosen halves of the sequences. (bottom) Leading eigenvectors $w_1(n)$, for $n = 1, 2, 3$. Scale is set by normalization, Eq (9).

the spectrum of eigenvalues $\{\lambda_n\}$, in rank order, and compare with data that have been shuffled to break any correlations between sequence and flexibility. We first notice that in both the real and shuffled data there are some true zero eigenvalues. These arise because we have $\sum_i S_i^\alpha = 1$ at each site $i$, by definition. In the shuffled data we see a spreading of the eigenvalues, which arises because we are estimating the correlation function from a finite sample [13] [16]. But in the real data there are at least two “modes” that stand out from this background.

The eigenvectors $w_1(n)$ have the same structure as position weight matrices, and pick out modes of sequence variation. Figure 3 (bottom) shows the three leading modes. We note that the first two have a clear structure, while the third—with its eigenvalue less clearly distinguished from the background noise in Fig 3—seems almost random. The first two modes show an approximate ten base periodicity, consistent with the pitch of the double helix, and are close to being a quadrature pair.

We expect eigenvectors to form exact quadrature pairs if $J$ is invariant to translations along the chain,

$$J_{ij}^{\alpha\beta} = J_{j-i}^{\alpha\beta}.$$  

(10)

If we think of $J_{ij}^{\alpha\beta}$ as an interaction between the bases as positions $i$ and $j$, then translation invariance is the statement that interactions depend on separation but not on absolute position. We can impose translation invariance by estimating $J$ from the data using the correlation function in Eq (7) and then replacing each matrix element by the average of all elements with the same value of $j-i$,

$$J_{i,j>i}^{\alpha\beta} \to \frac{1}{N-j+i} \sum_{k=1}^{N-j+i} J_{k,k+j-i}^{\alpha\beta}.$$  

(11)

As detailed in Appendix B the eigenvalues of this “cleaned” matrix stand out from the shuffled background with higher signal to noise ratio, both at large positive and large negative values; the eigenvectors are more clearly periodic; and eigenvalues come in degenerate pairs. We note that by imposing translation invariance, the number of independent parameters in the $J$ matrix is reduced from $\sim 15000$ to $\sim 600$, which significantly raises the signal to noise ratio of the inferred $J$ matrix.

We can decompose the sequence variations into modes defined by the eigenvectors, forming sequence features

$$f_n = \sum_{\alpha,\beta} w_{1,n}(n) S_i^\alpha.$$  

(12)

In Fig 4 we show the dependence of the cyclizability $C_0$ on the $f_n$ at the extremes of the spectrum. To avoid overfitting we estimate $\langle C_0 S_i^\alpha S_j^\beta \rangle_c$ and hence the eigenvectors $w_1(n)$ from half of the sequences, and then probe $C_0$ vs $f_n$ in the other half of the data. The mean behavior is almost perfectly quadratic along each feature, as predicted from Eq (5), and consistent with the absence of any linear correlation between sequence and $C_0$.

These results suggest that we should take the model in Eq (5) seriously. Again we estimate $J$ from half of the data, impose translation invariance, and predict $C_0$ for the other half of the data. Predictions vs measurements are shown in Fig 5 as a joint density: results are obtained from multiple random 50/50 splits into training and testing data. The correlation between predictions and measurements is $r = 0.59 \pm 0.01$. We can also find the contributions to $r$ from individual modes

$$C_0 = \langle C_0 \rangle + 8 \sum_n \lambda_n f_n^2,$$  

(13)

when all the modes are included, Eq (13) reduces to Eq (5). Including only the first two modes results in $r = 0.54 \pm 0.01$, suggesting that these modes make the largest contribution, as expected from the eigenvalue
FIG. 4: Cyclizability as a function of sequence features, from Eq (12). Error bars are standard deviations across random splittings of the data into 50/50 training/test sets. Lines are quadratic fits, as expected from Eq (5). Each feature is measured in units of its standard deviation across the ensemble of sequences.

spectrum, but including all modes provides significantly better predictions.

Should we be satisfied with the quality of predictions in Fig 5, or are we missing something? We have generated synthetic data on the assumption that the model in Eq (5) is exact, added noise to the resulting values of $C_0$, and repeated our analysis. In this scenario our ability to recover the underlying model is limited both by the finite number of samples and by the noise level. With noise levels in the range $\delta C \sim 0.25 - 0.3$ we find the same level of correlation between predictions and measurements as in Fig 5. There is no direct estimate of the noise level for the measurements in Ref [10], but with $\delta C \sim 0.25 - 0.3$ we would see a correlation of $r \sim 0.6 - 0.7$ between repeated measurements of $C_0$. This is slightly smaller than what is found in repeated measurements of the cyclizability on the Cerevisiae Nucleosomal Library [17], and comparable to what is seen in comparing random sequences with their reverse [10]. It thus is possible that the degree of correlation that we see between theory and experiment in Fig 5 is close to the limit set by the data itself.

What are the sequence features that control DNA flexibility? Because the eigenvectors are orthonormal, increasing the projection of the sequence onto one eigenvector necessarily decreases the projection onto others. The largest values of $C_0$ thus are predicted to occur in sequences that have maximal (squared) projection onto the first two modes. Table I shows the four predicted sequences that are extremal in this way. Characteristic features include 5–6 bp tracts of AT rich segments (i.e. TTAAA, TTTAA, and TTTAAA), followed by 5–6 bp tracts of CG rich segments (i.e. GGCCC, GGGCC, and GGGCCC), periodically. This is consistent with previous findings that molecules with AT rich stretches separated by 5 bp from GC rich stretches are more loopable [18, 19].

At the opposite extreme, sequences that maximize the squared projection to the last two modes are predicted to have the smallest values of $C_0$. These sequences have shorter lengths of repeated nucleotides, and shorter periodicities for the reappearance of the same motifs.

Early work on the sequence dependence of DNA flexibility focused on the influence of dinucleotide pairs, which could be detected in smaller data sets [20, 21]. The high throughput experiments of Ref [10] made it possible to see the influence of helical periodicity, leading to models that combine local dinucleotide features across longer distances [18]. In many ways our results recapitulate those of Ref [18], although our model is simpler.

TABLE I: Predicted DNA sequences with highest and lowest intrinsic cyclizabilities.
Beyond the analysis of DNA flexibility, our results illustrate the power of correlation functions to extract meaningful information from modern high throughput data. The analysis is simpler because the experimental sequence ensembles are fully random with no intrinsic correlations, although the discussion can be generalized. It is attractive to see the problem of finding relevant features in DNA sequences as being equivalent to the problem of finding relevant features in sensory stimuli, where in both cases relevance is defined by some functional behavior of the biological system.

Acknowledgments

YZ and WB were supported in part by the National Science Foundation through the Center for the Physics of Biological Function (PHY–1734030) and Grant PHY–1607612. AB was a Simons Foundation Fellow of the Life Sciences Research Foundation, and TH is an Investigator with the Howard Hughes Medical Institute.

Appendix A: Some details

Here we give some mathematical details for the analysis of the model in Eq (5),

\[
C_0 = \langle C_0 \rangle + \frac{1}{2} \sum_{kl, \gamma \delta} J_{kl}^{\gamma \delta} (S_k^\gamma - 1/4)(S_l^\delta - 1/4). \tag{A1}
\]

The first thing we notice is that if we shift

\[
J_{kl}^{\gamma \delta} \rightarrow J_{kl}^{\gamma \delta} + u_k^\gamma b_l,
\]

then we pick up a term in Eq \((A1)\)

\[
\sim u_k^\gamma b_l \sum_\delta (S_l^\delta - 1/4) = 0.
\]

This means that, without loss of generality, we can set

\[
\sum_\alpha J_{ij}^{\alpha \beta} = \sum_\beta J_{ij}^{\alpha \beta} = 0. \tag{A2}
\]

If we think of \(J_{ij}^{\alpha \beta}\) as a \((4N) \times (4N)\) matrix, the condition in Eq \((A2)\) reduces the rank by \(N\), which makes sense since we have \(N\) constraints \(\sum_\alpha S_i^\alpha = 1\).

We look next at the contribution from a term \(J_{kk}^{\gamma \delta}\) that is diagonal in the site indices:

\[
\sum_{\gamma \delta} J_{kk}^{\gamma \delta} (S_k^\gamma - 1/4)(S_k^\delta - 1/4) = \sum_{\gamma \delta} J_{kk}^{\gamma \delta} [\delta_{\gamma \delta} S_k^\gamma - (1/4)(S_k^\gamma + S_k^\delta) + 1/16]
\]

\[
= \sum_\gamma \left[ J_{kk}^{\gamma \gamma} - (1/2) \sum_\delta J_{kk}^{\gamma \delta} \right] S_k^\gamma + (1/16) \sum_\gamma J_{kk}^{\gamma \gamma} S_k^\gamma.
\tag{A3}
\]

where in the last step we use Eq \((A2)\). Thus the only site diagonal term that can contribute also is diagonal in the base index, and this contribution collapses back to a linear model, as in Eq (3), with \(W_k^\gamma = J_{kk}^{\gamma \gamma}\). Thus we can also zero out \(J_{kk}^{\gamma \delta}\), since it is redundant.

Now we are prepared to compute the correlation function that appears in Eq \((6)\),

\[
\left\langle (C_0 - \langle C_0 \rangle)(S_i^\alpha - 1/4)(S_j^\beta - 1/4) \right\rangle = \frac{1}{2} \sum_{kl, \gamma \delta} J_{kl}^{\gamma \delta} \left\langle (S_i^\alpha - 1/4)(S_j^\beta - 1/4)(S_k^\gamma - 1/4)(S_l^\delta - 1/4) \right\rangle \tag{A4}
\]

We notice that the average is zero if all the indices \(ijkl\) are different; more precisely if \(k\) is different from all the other indices, then we get zero. There is no term \(k = 1\), so we must have \(k = i\) or \(k = j\); let’s try \(k = i\):

\[
\left\langle (S_i^\alpha - 1/4)(S_j^\beta - 1/4)(S_k^\gamma - 1/4)(S_l^\delta - 1/4) \right\rangle = \left\langle (S_j^\beta - 1/4)(S_l^\delta - 1/4)[\delta^{\alpha \gamma} S_i^\alpha - (1/4)(S_i^\alpha + S_k^\gamma)] + 1/16 \right\rangle. \tag{A5}
\]
Since $i = k \neq l$, the only remaining choice is whether $i = j$ or not. If not, then the average factors,

$$\left\langle (S_i^\alpha - 1/4)(S_j^\beta - 1/4)[\delta^{\alpha\gamma} S_k^\gamma - (1/4)(S_i^\alpha + S_j^\gamma) + 1/16] \right\rangle = \left\langle (S_i^\alpha - 1/4)(S_j^\beta - 1/4) \right\rangle [\delta^{\alpha\gamma} (1/4) - (1/16)] = [\delta_{ij} (1/16)[\delta^{\beta\delta} - (1/4)][\delta^{\alpha\gamma} - (1/4)]. \hfill (A6)$$

On the other hand, if $i = j \neq l$ we have

$$\left\langle (S_i^\alpha - 1/4)(S_j^\beta - 1/4)(S_k^\gamma - 1/4)(S_l^\delta - 1/4) \right\rangle = \left\langle (S_i^\alpha - 1/4)(S_j^\beta - 1/4)(S_l^\delta - 1/4) \right\rangle \left\langle (S_i^\alpha - 1/4) \right\rangle = 0. \hfill (A7)$$

So what we have shown that there is one term

$$\left\langle (S_i^\alpha - 1/4)(S_j^\beta - 1/4)(S_k^\gamma - 1/4)(S_l^\delta - 1/4) \right\rangle \hfill (A8)$$

The other choice was $k = j$, which we can get by swapping $(i, \alpha) \leftrightarrow (j, \beta)$. This gives

$$\left\langle (S_i^\alpha - 1/4)(S_j^\beta - 1/4)(S_k^\gamma - 1/4)(S_l^\delta - 1/4) \right\rangle \hfill (A9)$$

Putting these together we have

$$\langle C_0 S_i^\alpha J_j^\beta \rangle_c \equiv \left\langle (C_0 - \langle C_0 \rangle)(S_i^\alpha - 1/4)(S_j^\beta - 1/4) \right\rangle$$

$$= \frac{1}{32}(1 - \delta_{ij}) \sum_{kl,\gamma\delta} J_{kl}^{ij}[\delta^{\beta\delta} - (1/4)][\delta^{\alpha\gamma} - (1/4)] + \frac{1}{32}(1 - \delta_{ij}) \sum_{kl,\gamma\delta} J_{kl}^{ij}[\delta^{\alpha\delta} - (1/4)][\delta^{\beta\gamma} - (1/4)]$$

$$= \frac{1}{32}(1 - \delta_{ij}) \sum_{\gamma\delta} J_{ij}^{\gamma\delta}[\delta^{\beta\delta} - (1/4)][\delta^{\alpha\gamma} - (1/4)] + \frac{1}{32}(1 - \delta_{ij}) \sum_{\gamma\delta} J_{ij}^{\gamma\delta}[\delta^{\alpha\delta} - (1/4)][\delta^{\beta\gamma} - (1/4)]. \hfill (A10)$$

We recall that $J_{ij}^{\gamma\delta} = J_{ji}^{\delta\gamma}$, so that

$$\langle C_0 S_i^\alpha J_j^\beta \rangle_c = \frac{1}{16}(1 - \delta_{ij}) \sum_{\gamma\delta} J_{ij}^{\gamma\delta}[\delta^{\beta\delta} - (1/4)][\delta^{\alpha\gamma} - (1/4)]$$

$$= \frac{1}{16}(1 - \delta_{ij}) \left[ J_{ij}^{\gamma\delta} - (1/4) \sum_\delta J_{ij}^{\gamma\delta} - (1/4) \sum_\gamma J_{ij}^{\gamma\delta} + (1/16) \sum_\gamma \sum_\delta J_{ij}^{\gamma\delta} \right]. \hfill (A11)$$

Now we use $\sum_\beta J_{ij}^{\alpha\beta} = 0$, and our result collapses to

$$\langle C_0 S_i^\alpha J_j^\beta \rangle_c = \frac{1}{16}(1 - \delta_{ij}) J_{ij}^{\gamma\delta} = \frac{1}{16} J_{ij}^{\gamma\delta}. \hfill (A12)$$

Appendix B: Imposing translation invariance

We impose translation invariance on the matrix $J_{ij}^{\alpha\beta}$ according to Eq. (11); Fig. 6 shows the $J$ matrix before and after this treatment. As noted in the main text, translation invariance reduces the number of free parameters in $J$ from $\sim 15000$ to $\sim 600$ and thus raises the signal to noise ratio in the inferred matrix elements. The “cleaned” $J$ not only shows clear stripes near the diagonal, suggesting strong nearest neighbor interactions in determining the cyclizability, but also displays a set of stripes separated at half-helical ($\sim 5$ bp) and helical ($\sim 10$ bp) period of DNA, suggesting a role more longer ranged interactions in determining DNA flexibility.

The eigenvalues of the cleaned $\langle C_0 S_i^\alpha J_j^\beta \rangle_c (= J/16)$ matrix stand out from the shuffled background with higher signal to noise ratio, both at large positive and large negative values, and the eigenvectors are more clearly periodic. Results are shown in Fig. 4 which should be compared with Fig. 3 in the main text. We note that although the matrix $J$ is translation invariant, the eigenvectors exhibit clear boundary effects, so that modes 199 and 200 are almost localized at the ends of the sequence.

[1] PM Chaikin and TC Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge UK, 1995).
[2] ME Peskin and DV Schroeder, An Introduction to Quantum Field Theory (Perseus Books, Reading MA, 1995).
[3] E de Boer and P Kuyper, Triggered correlation. IEEE Trans Biomed Eng 15, 169–179 (1968).
[4] F Rieke, D Warland, R de Ruyter van Steveninck, and W Bialek Spikes: Exploring the Neural Code (MIT Press, Cambridge, 1997).
[5] LF Abbott and P Dayan, Theoretical Neuroscience. Computational and Mathematical Modeling of Neural Systems (MIT Press, Cambridge MA, 2001).
[6] N Wiener, Nonlinear Problems in Random Theory (MIT Press, Cambridge MA, 1958).
[7] R de Ruyter van Steveninck and W Bialek, Real–time performance of a movement sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences. Proc R. Soc London Ser. B 234, 379–414 (1988).
[8] W Bialek and R de Ruyter van Steveninck, Features and dimensions: Motion estimation in fly vision. arXiv:q–bio/0505003 (2005).
[9] A Cavagna, I Giardina, and T Grigera, The physics of flocking: Correlation as a compass from experiments to theory. Phys Repts 728, 1–62 (2018).
[10] A Basu, DG Bobrovnikov, Z Qureshi, T Kayikcioglu, TTM Ngo, A Ranjan, S Eustermann, B Cieza, MT Morgan, M Hejna, H Rube, K–P Hopner, C Wolberger, JS Song, and T Ha, Measuring DNA mechanics on the genome scale. bioRxiv 2020.08.17.255042 (2020).
[11] DH Hubel and TN Wiesel, Receptive fields, binocular interaction, and functional architecture in the cat’s visual cortex. J Physiol (Lond) 160, 106–154 (1962).
[12] OG Berg and PH von Hippel, Selection of DNA binding sites by regulatory proteins. Statistical–mechanical theory and application to operators and promoters. *J Mol Biol* 193, 723–750 (1987).

[13] GD Stormo, DNA binding sites: representation and discovery, *Bioinformatics* 16, 16–23 (2000).

[14] JB Kinney, G Tkačik and CG Callan Jr, Precise physical models of protein–DNA interaction from high-throughput data. *Proc Natl Acad Sci (USA)* 104, 501–506 (2007).

[15] M Potters and J–P Bouchaud, *A First Course in Random Matrix Theory: for Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge UK, 2020).

[16] We have verified that the maximum and minimum eigenvalues in the shuffled data vary as $1/\sqrt{M}$, where $M$ is the number of sequences in our sample, as expected from random matrix theory.

[17] Recall that $C_0$ is the *intrinsic* cyclizability, obtained by analyzing measurements at different locations of the bead attachment. What is reported in Ref [10] is the repeatability of these individual measurements.

[18] A Basu, DG Bobrovnikov, B Cieza, Z Qureshi, and T Ha, Deciphering the mechanical code of genome and epigenome. bioRxiv 2020.08.22.262352 (2020).

[19] G Rosanio, J Widom, and OC Uhlenbeck, In vitro selection of DNAs with an increased propensity to form small circles. *Biopolymers* 103, 303–320 (2015).

[20] A Sarai, J Mazur, R Nussinov, and RL Jernigan, Sequence dependence of DNA conformational flexibility. *Biochemistry* 28, 7842–7849 (1989).

[21] S Geggier and A Vologodskii, Sequence dependence of DNA bending rigidity. *Proc Natl Acad Sci (USA)* 107, 15421–15426 (2010).