Modelling chromosome-wide target search

Lucas Hedström and Ludvig Lizana

Integrated Science Lab, Department of Physics, Umeå University, SE-901 87 Umeå, Sweden

* Author to whom any correspondence should be addressed.

E-mail: ludvig.lizana@umu.se

Keywords: DNA target-search, diffusion on networks, chromosome 3D folding, Hi-C data, stochastic simulations, gene regulation

Abstract

The most common gene regulation mechanism is when a transcription factor (TF) protein binds to a regulatory sequence to increase or decrease RNA transcription. However, TFs face two main challenges when searching for these sequences. First, the sequences are vanishingly short relative to the genome length. Second, there are many nearly identical sequences scattered across the genome, causing proteins to suspend the search. But as pointed out in a computational study of LacI regulation in Escherichia coli, such almost-targets may lower search times if considering DNA looping. In this paper, we explore if this also occurs over chromosome-wide distances. To this end, we developed a cross-scale computational framework that combines established facilitated-diffusion models for basepair-level search and a network model capturing chromosome-wide leaps. To make our model realistic, we used Hi-C data sets as a proxy for 3D proximity between long-ranged DNA segments and binding profiles for more than 100 TFs. Using our cross-scale model, we found that median search times to individual targets critically depend on a network metric combining node strength (sum of link weights) and local dissociation rates. Also, by randomizing these rates, we found that some actual 3D target configurations stand out as considerably faster or slower than their random counterparts. This finding hints that chromosomes’ 3D structure funnels essential TFs to relevant DNA regions.

1. Introduction

When responding to changes in the immediate environment or coordinating embryo development, cells regulate internal protein production. While there are many regulation layers, the most common approach is when DNA-binding proteins, like transcription factors (TFs), attach to target sequences to start or stop DNA transcription. However, because these sequences are so short compared to the DNA's length, finding them represents a needle-in-a-haystack problem. In humans, the targets are hundreds of millions of times shorter than the DNA itself. Despite this problem, protein production works seamlessly, and measured search times are relatively short—about an order of magnitude faster than random search [1].

This observed fast search is further perplexing because many almost-targets are scattered across the genome [2]. Most such targets reside out-of-interaction range from the actual target gene. At first glance, such dispersion should slow the search since DNA-binding proteins get held up at the wrong places. Indeed, this happens in a simple diffusion model where a searcher gets caught in trap after trap. However, if the searcher may intermittently relocate over large distances, e.g., via DNA looping, such traps can instead facilitate the search. This seems to occur for the LacI repressor in E. coli. As shown in a modeling study [3], LacI’s main target has two flanking auxiliary binding sites that help lower the average search time if LacI first binds one of them and then loops into the primary target.

In this paper, we built a mathematical model to explore whether long-distant auxiliary binding sites could facilitate target search via chromosome-wide 3D interactions. Our analysis focuses on human DNA, where there are potentially thousands of such sites. While most of them are far from the primary target (if

---

1 Typical Transcription factor sequences are 10–50 bp long. The human genome has $7 \cdot 10^9$ bps.
counting the number of base pairs), some are close in 3D due to chromatin folding that may help to channel searching proteins to their designated DNA regions. This mechanism is arguably hard to measure experimentally, but previous work hint that folded chromatin directs protein traffic. For example, using simulations, one study argued that diffusing TFs tend to end up near chromatin loop anchors [4]. This finding is consistent with so-called Highly Occupied Targets that enrich in regions engaged in long-distance DNA interactions [5]. Another computational study addressed target search in networks of polymer loops and found that search times depend critically on the spatial target location relative to the loop anchors [6].

We extend these ideas to a chromosome-wide setting. To achieve this, we take advantage of empirical data sets from Hi-C experiments, providing a realistic description of the large-scale chromosome 3D organization [7], and datasets containing binding profiles of more than 100 TFs [8, 9]. We also developed a diffusion model mixing protein search on short and large scales. On short scales, we use a biophysical model of a two-state searcher [10], resting on established Facilitated diffusion models [11–14]. On large scales, we use a network model that includes chromosome-wide re-location [15]. Using simulations, we find that the spatial configurations and binding strengths of the auxiliary binding sites strongly affect first-passage times.

2. Methods

We develop a cross-scale model for site-specific DNA search where we represent a searching TF protein as a point particle. Our model connects sequence-sensitive search on length scales shorter than a few kilo basepairs (kbp) to long mega-basepair (Mbp) leaps across the entire chromosome. To achieve this, we merge established biophysical models for facilitated diffusion, and coarse-grained chromosome-wide search [10, 15].

2.1. Sequence-sensitive search

To model search over kbp distances, we use a one-dimensional sliding model designed for TFs looking for 10–20 bps long sequence motifs [10]. In this model, the TF switches stochastically between two modes, denoted search and recognition. In search mode, the TF diffuses along the DNA, while weakly sensing the sequence. But sometimes, the TF occasionally enters recognition mode and stops moving. The more the local sequence resembles the target motif, the higher the probability of switching to recognition mode and the longer it stays bound.

We express the critical rates—transport, binding, and unbinding—in terms of a one-dimensional energy landscape, $E_{TF}(x)$ ($x = 0, 1, 2, \ldots$). This landscape represents how well the local sequence $x$ matches the consensus motif; see figure 1 for an example, and appendix A for how we calculate $E_{TF}(x)$. Below is a brief account of all model’s rates.

2.1.1. Diffusion

We assume that left and right diffusion rates, $k_l(x)$ and $k_r(x)$, depend on the energy difference between neighbouring sites, $x$ and $x \pm 1$:

\[
\begin{align*}
  k_r(x) &= k_d e^{\rho (E_{TF}(x) - E_{TF}(x+1))}, \\
  k_l(x) &= k_d e^{\rho (E_{TF}(x) - E_{TF}(x-1))},
\end{align*}
\]

where $k_d T = 1$ and $k_d$ is the natural diffusion frequency. The ruggedness factor $\rho$ controls the sequence sensitivity in search mode. As pointed out in [13], the search becomes too slow compared to experiments if the sensitivity is too high. In fact, $\rho_{\text{max}} = 0.3$ is the upper limit allowing for realistic diffusion constants [10].

Like in [10], we set $\rho = 0.2$ and $k_d = 10^7$ to ensure that the average diffusion constant over a long DNA stretch is realistic, i.e. $D = 10^7 \text{bp}^2 \text{s}^{-1} \approx 0.1 \mu \text{m}^2 \text{s}^{-1}$ [16]. The factor 1/2 in the exponent guarantees detailed balance: $k_l(x)/k_l(x+1) = e^{\rho (E_{TF}(x+1) - E_{TF}(x))}$.

2.1.2. Binding and unbinding

Next, we consider switching between search and recognition mode using two rates: $k_b(x)$ and $k_u(x)$. $k_b(x)$ is the rate to enter recognition mode (binding rate) and $k_u(x)$ denotes the rate to return to the search mode (unbinding rate). Following [10], we assume that these rates depend on the energy differences and obey detailed balance:

\[
\begin{align*}
  k_b(x) &= \gamma e^{\rho (E_{TF}(x) - E_{TF}(x))} - \Delta G, \\
  k_u(x) &= \gamma e^{\rho (E_{TF}(x) - E_{TF}(x)) + \Delta G},
\end{align*}
\]

(2)
Figure 1. Schematic showing sequence-specific and chromosome-wide search. (a) The TF (orange circle) performs sequence-sensitive search along each DNA segment (or Hi-C bin), where it may slide or switch stochastically between search and recognition modes. (b) Each box (containing a sequence of length $L_{TF}$) is associated with an energy level reflecting the similarity between the local sequence and TF target motif. We use standard methods (see section 2.1) to map this similarity onto an energy landscape. In search mode, the TF is less sensitive to the sequence than in recognition mode (blue versus orange lines). (c) Crumpled DNA polymer. Colours indicate DNA segments representing Hi-C bins. (d) Effective network of 3D interactions between DNA segments. Thin to thick edges represent low to high contact frequencies. (e) We store these frequencies as elements in an adjacency matrix used in the simulations.

Here, $\gamma = 10^7 \text{s}^{-1}$ (representing a fast base transition rate [13, 17]), and $\Delta G$ is the free energy change associated with switching between search and recognition. As in [10], we calibrate this parameter by letting the energy difference between the bound and unbound state be zero for the weakest consensus target along the whole chromosome, $\Delta G = -(1 - \rho_{max}) \max E_{TF}(x \in \text{consensus targets})$ (see section 2.2.2 for how we define consensus targets).

Finally, the TF may dissociate from the DNA when in search mode. We define the off-rate as:

$$k_{off} = \delta e^{\rho E_{TF}(x)}, \quad (3)$$

where $\delta = 10^3 \text{s}^{-1}$. In our simulations, this parameter choice corresponds to the sliding length $\approx 90$ bp, which agrees with studies. ([1, 18] suggests that the sliding length is $\lesssim 100$ bp.)

2.1.3. Inter-segmental transfer
In the cross-scale model, we include loop-assisted jumps between distant DNA sites. And similar to other studies [3, 14], we assume that this process depends critically on the free energy cost of forming a loop. However, we differ from these works in how we calculate this cost. If assuming that DNA forms loops like a polymer, say a Gaussian or worm-like chain, the loop-size distribution decays as a power law, and the free energy becomes $\Delta G_l = -a \ln l$, where $l$ is the loop size and $a$ is the looping exponent ($a = 3/2$ for a Gaussian polymer). However, as we know from Hi-C data analysis [19, 20], the contact probabilities over chromosome-wide scales do not decay as simple power-laws. Instead, Hi-C experiments show that the pairwise contact counts are more complex. Therefore, we treat DNA looping separately in the chromosome-wide search where we take advantage of Hi-C data sets. In short, whenever there is an unbinding event (with the rate $k_{off}$) the particle translocates to another site on the chromosome with a probability proportional to the number of measured chromosome contacts. This includes rebinding to the same segment, mimicking typical intersegmental transfer but to a random location. We refer to section 2.3 for details.

---

We acknowledge that the genome-averaged contact probability in human decays as $l^{-1.08}$ [25] or $l^{-0.75}$ within TADs [26]. However, individual pairwise contacts show significant deviations from these average relationships.
2.2. Defining consensus targets and auxiliary binding sites
In this paper, there is a critical difference between consensus and auxiliary binding sites. While consensus targets are experimentally-derived TF binding positions, auxiliary binding sites come from motif-matching algorithms that scan the genome sequence for potential binding sites. Interestingly, these two data sets do not always overlap, highlighting that some binding sites lack sequence similarity with the consensus motif [21].

2.2.1. Consensus targets
We downloaded the consensus site positions for each TF from the JASPAR database [8, 22]. JASPAR is a library of manually curated TF binding profiles from experiments (stored as position frequency matrix). We annotate the unique TFs by their respective position frequency matrix (PFM) matrix number [23].

2.2.2. Auxiliary binding sites
To computationally determine the auxiliary binding sites, we use the Motif Occurrence Detection Suite (MOODS) [9] that calculates the binding score \( W^{TF}(x) \) at any genomic position \( x \). The binding score is a sum over the position weight matrix (PWM) [24] across positions in a sliding window \( L^{TF} \) with the same length as the potential target:

\[
W^{TF}(x) = \sum_{i=x}^{x+L^{TF}} \log \frac{p^{TF}(a_i, i)}{\pi(a_i)}. \tag{4}
\]

where \( p^{TF}(a_i, i) \) is the PWM (equation (A1)), and \( a_i \) denotes the base pair at position \( i \) (A, C, T, or G). Also, MOODS separates significant from insignificant auxiliary binding sites using a threshold \( T \), where \( W^{TF}(x) > T \). This threshold is associated with the probability \( p_m \) that a sequence snippet is randomly generated from a background basepair frequency \( \pi(a) \). We used \( p_m = 10^{-4} \) and \( \pi(a_i) = 1/4 \) (default settings).

2.3. Chromosome-wide search
In the sequence-sensitive search, the protein diffuses along the DNA until it finds a consensus site or dissociates. This section describes how we model rebinding to a potentially distant DNA segment after dissociation.

Rebinding depends critically on the spatial arrangement of DNA segments. While not a part of the classical facilitated diffusion model, several theoretical and experimental studies showed that search times change with correlated rebinding and the type of DNA conformations [3, 12, 15, 27–32]. Here, we build one of these approaches that circumvent knowing the chromatin’s explicit large-scale 3D structure [15]. Instead, it treats DNA’s 3D interactions as a contact network and infers contact probabilities between DNA segments from experimental Chromosome Conformation Capture data (we use Hi-C data from [7] and LiftOver to UCSC hg38 from UCSC hg19 [33]). This network approach embraces the full complexity of chromosomes’ semi-hierarchical organization that cannot be described using Gaussian or Fractal globule polymers with simple power-law decaying loops-sizes.

2.3.1. Network model
We consider chromatin’s 3D interactions as a weighted network where the nodes and links represent DNA segments and 3D contacts, respectively. The links’ weights are proportional to the segment-segment contact probabilities derived from Hi-C data. To model the chromosome-wide protein search, we consider a point particle jumping randomly across the network (see figure 1). These jumps are associated with rates \( \omega_{ij} \) \( (i, j = 1, \ldots, N) \) where \( N \) is the number of nodes (typically \( N \sim 10^4 - 10^5 \)). We calculate \( \omega_{ij} \) as the probability \( q_{ij} \) to jump between segments \( i \) and \( j \) multiplied by the frequency of a successful jump \( f_{coll} \) (collision frequency’). As a proxy for \( q_{ij} \), we use the number of Hi-C contacts \( v_{ij} \) (Knight-Ruiz (KR)-normalized), leading to [15]:

\[
\omega_{ij} = f_{coll} v_{ij}, \quad \omega_{ij} = \omega_{ji}. \tag{5}
\]

For clarity, we use population-averaged Hi-C data to calculate \( \omega_{ij} \). Therefore, the rates do not include cell-to-cell variability and potentially transient loops. Instead, the network model treats the chromosomes as a rigid structure where the probability of jumping from one segment to another is proportional to the contact frequency.

2.3.2. Hi-C data treatment
To extract the pair-wise DNA contacts, we used Hi-C data from human cell line GM12878 [7], downloaded from the GEO database [34]. Before calculating the rates \( \omega_{ij} \), we normalized the contact counts \( v_{ij} \) using the
Once assigned a Hi-C bin, the TF diffuses from a random locus on the corresponding 5 kbp segment. Here it performs a sequence-specific search as it diffuses through the energy landscape $E^T(x)$. If encountering deep valleys, the TF may fall into recognition mode and stay trapped for some time ($\propto 1/k_u(x)$, equation (2)).

In the chromosome-wide search, we did not consider chromosome-chromosome contacts. Admittedly, this is an approximation as there are interactions with other chromosomes, albeit much less frequent. To check this assumption, we calculated the ratio of internal versus external contacts from Hi-C data [15]. Depending on the chromosome, about 75%–90% of the contacts are internal. However, because of the low signal-to-noise ratio [38], this number should be taken cautiously.

Furthermore, we used 5 kb Hi-C data. 5 kb represents a threshold separating sequence-sensitive and chromosome-wide searches in our cross-scale model. Admittedly, this threshold is arbitrary, and there are two main reasons for this choice. First, 5 kb agrees with length scales used in other facilitated-diffusion studies that typically vary between 1–10 kb (e.g. [3, 10]). Second, on length scales much larger than 5 kb, say 50–100 kb, the looping probabilities no longer follow simple power-law relationships. Instead, on these scales, chromatin interactions are more complex. Balancing these arguments, we picked 5 kb as a conservative choice. But to be clear, this choice does not reflect any inherent limitation in our approach. Our cross-scale framework can handle genome-wide Hi-C matrices at any resolution that may include intra- and inter-chromosomal contacts.

### 2.4. Cross-scale computational approach

We combined sequence-sensitive and chromosome-wide searches in a cross-scale computational framework to simulate TF search over large distances. This framework builds on the Gillespie algorithm [39] to calculate transition probabilities and time step lengths. Below are the main steps:

1. We start with a TF in a randomly selected Hi-C bin (node). We exclude isolated bins (where the row sum of the Hi-C matrix is non-zero), such as loci on the centromere and the telomeres.
2. Once assigned a Hi-C bin, the TF diffuses from a random locus on the corresponding 5 kbp segment. Here it performs a sequence-specific search as it diffuses through the energy landscape $E^T(x)$. If encountering deep valleys, the TF may fall into recognition mode and stay trapped for some time ($\propto 1/k_u(x)$, equation (2)).
3. The TF may also dissociate from DNA (with the rate $\sim k_{off}$). If so, it will instantaneously relocate to another Hi-C bin, say from bin $i$ to $j$, with the relocation probability $\omega_{ij}$ (see equation (5)). This intermittent transition includes rebinding to the same Hi-C bin, but at a random position, $x'$. Modelling rebinding in this way approximately captures facilitated diffusion, albeit with a uniform jumping probability rather than a power-law.
4. We repeat these steps until the TF reaches a Hi-C bin that contains at least one consensus target site (i.e., consensus bins act as absorbing targets). After storing the search time, we initiate a new run where the TF starts from another random Hi-C bin.

To speed up the simulations, we made considerable data pre-processing. Foremost, before the simulations, we parsed the chromosome’s sequence and calculated all TF-specific rates—$k_L$, $k_R$, $k_u$, $k_b$, and $k_{off}$ (see equations (2) and (3))—and stored them in a separate file that we load once the simulation starts. By calculating the rates in this way, we significantly reduced the computational time. Particularly, this sped up the switch between chromosome-wide and sequence-specific search.

### 3. Results & discussion

#### 3.1. Search times to consensus target bins vary among TFs

To study search-time variations among TFs, we downloaded 108 TF binding profiles from [22], annotated by their matrix IDs [23]. For each TF, we curated lists of consensus targets and auxiliary binding sites and assigned these to their designated 5 kb Hi-C bins (nodes in the DNA contact network). We noted that most bins contain zero or one consensus target, but some have more than ten. Next, we simulated the search for each TF type separately using our cross-scale model (see section 2.4). Across $>10^6$ simulation runs with different starting conditions, we recorded the first-passage times to any bin containing at least one consensus target.

Because the genomic locations and the number of consensus targets differ significantly, we expect to observe a wide search-time variability between the TFs. Indeed, plotting the distribution of the $10^6$ search

---

3 Close to 80% of the genes are active, which is significantly more that the other chromosomes where about 50% are active [15].

4 We implemented intermittent power-law jumps within the 5 kb region too, but we did not notice any meaningful differences when studying large-distance search times.
Figure 2. Target bin finding times for 10 TFs searching for any consensus target bins on chromosome 19 (TF labels: MA1644.1, MA0502.2, etc.). (a) Letter-value plots showing the distribution of first-passage times to any of the consensus sites. For each TF, we simulated \( \sim 10^6 \) search times from random starting points. When plotted in log scale, we note that the search times vary by several orders of magnitude (within and between TFs). (b) Letter-value plots of search times with normalized energy landscapes (see appendix B). Comparing panels (a) and (b), we see that the variability between TFs becomes smaller. Yet, there are still significant differences. (c) Fraction of target bins containing at least one consensus target versus the median first-passage time (normalized energy landscape). Here we show simulation results from 60 TFs, where marks indicate the TFs in panel (b). The green line shows the fit \( \phi^{-0.97} \) (note the logged axes). This panel indicates that the median search times depend mainly on \( \phi \) after we normalize the energy landscape.

For 10 TF types (figure 2(a)), we see that some TFs are faster than others, where the medians (horizontal lines) may differ by almost two orders of magnitude. To better understand what causes these variations, we analyzed how search times change with two metrics: effective diffusion constants and the number of consensus target bins.

As explained in Methods, we envision the TF diffusing through an energy landscape when sliding along DNA. This landscape has valleys and hills reflecting how close the local sequence reassembles the consensus target motif. If there are many local minima, we expect the TF will get stuck repeatedly, thus leading to a lower diffusion constant. For example, the effective diffusion constant associated with a random one-dimensional landscape is approximately \( D = D_0 (1 + \sigma_E^2/2)^{1/2} e^{-7\sigma_E^4/4} \), where \( \sigma_E^2 \) is the landscape’s variance, and \( D_0 \) is the diffusion constant when it is flat [27]. In other words, rougher energy landscapes lead to lower diffusion constants. Therefore, to correct for different \( \sigma_E^2 \) when comparing TF search times, we renormalized the landscapes’ variance to unity (the mean was already zero) and made new simulations (see appendix B for details). We found that the variations are much smaller after re-normalization—compare panels (a) and (b) in figure 2. However, the median search time still differs by a factor of 10 between the fastest and slowest TF type.

Besides varying effective diffusion constants, the number of consensus targets and auxiliary binding sites also change. For example, comparing two TFs having short or long target motifs, say 10 and 20 bp, it is clear that the one looking for the 10 bp motif encounters many more almost-targets in a long random basepair sequence than the one searching for a 20 bp target. Therefore, if tracking search times to any consensus target bin, we expect that TFs having high target counts will have short average search times.

But there is yet another factor worth considering. While the total target number is important, it is also critical to how far the targets are from each other along the DNA. For example, if several targets lie within the TF’s sliding length, i.e., distance covered before dissociating, it will likely find any of them if in proximity. Therefore, dense target clusters may act as effective targets (see theoretical analysis in appendix D). Furthermore, how these effective targets distribute in the DNA contact network also matters. For example, if spread out evenly, we expect that the average search time is shorter than if constrained to a few Hi-C bins. To study such effects, we plotted the fraction \( \phi \) of Hi-C bins having at least one consensus target against the median-first passage time \( \tau \) to any one of these target bins for 40 TFs (figure 2(c)); we used the normalized energy landscape as in panel (b). We find a convincing inverse relationship, where the search time decays as \( \phi^{-0.97} \) (linear fit, solid line).

To better understand this relationship, we considered a simple search model: a fully connected weighted network (even if dense, Hi-C is not fully connected). If starting at some node \( i \), the probability that a searcher finds the target in one step \( P_i(1) \) equals the probability that the next node \( j \) is a target (\( \phi \)) times the probability

\[ P_i(1) = \phi \]
of translocating from \(i\) to \(j\) (\(\omega_{ij}\)). Since there are many target nodes in the network, we sum over all \(j\),

\[
P_i(1) = \sum_j \omega_{ij} \phi = \phi,
\]

where we used that all weights sum to unity \(\sum_j \omega_{ij} = 1\) (they are KR-normalized). Generalizing equation (6) to \(n > 1\) steps yields:

\[
P_i(n) = (1 - \phi)^{n-1} \phi.
\]

Since the search trajectories are similar in this toy model, averaging over initial starting points \(i\) gives the geometric distribution \(\langle P_i(n) \rangle = P(n) = (1 - \phi)^{n-1} \phi\). This distribution has the median \(\tilde{n}\) [40]:

\[
\tilde{n} = \left\lceil \frac{-1}{\log_2(1 - \phi)} \right\rceil,
\]

in which \(\log_2\) is the base-2 logarithm and \(\lceil \cdot \rceil\) is the round-up function. Since \(\phi < 10^{-1}\) for most TFs, (see figure 2(c)), we expand the enumerator for small \(\phi\), yielding:

\[
\log_2(1 - \phi) = -\frac{\phi}{\log(2)} - O(\phi^2) \approx -\frac{\phi}{\log(2)},
\]

and thus:

\[
\tilde{n} \approx \left\lceil \frac{\log(2)}{\phi} \right\rceil \sim \frac{1}{\phi}.
\]

In the normalized systems, the effect of auxiliary binding sites is not negligible, but considerably weaker than the non-normalized systems. For multiple trajectories and long search paths \((n \gg 1)\), the search time is proportional to the number of jumps, giving us:

\[
\tilde{\tau} \sim \tilde{n} \sim \phi^{-1},
\]

which is consistent with the data in figure 2.

To conclude, when considering search times to any one of the consensus target bins, we can explain most of the variation with effective diffusion constants and by how much the effective consensus targets are distributed over the chromosome. But even if these averages have moderate variability (after normalization), search times for single target bins may differ substantially. This is the next section’s topic.

### 3.2. Search time variability between individual consensus target bins

The previous section showed that search times vary if considering all consensus bins as valid targets (figure 2). But in actual gene regulation, some targets are likely more important than others. Some are also easier to find depending on their chromosomal positions. To study search time variability among individual consensus targets, we shift to a microscopic view and analyze the TF’s search to specific bins. Just as the Lac repressor gets help from surrounding auxiliary binding sites to quickly find its main operator site, as discussed in the Introduction, we are interested in what extent consensus targets occupy easy-to-access locations on the chromosome and if surrounding auxiliary binding sites arrange in favourable network positions that help TFs reach specific DNA regions having strong targets or high target densities.

To quantify this search-time variability, we calculated the median search times \(\tau\) to consensus target bins for six TFs (figure 3). Each figure panel portrays sorted search times, one TF per panel, where the solid lines represent the median \(\tau\), and the shaded areas show the 95% confidence interval (of \(\sim 10^3\) search paths). The panels show that \(\tau\), in some cases, varies by a factor of 2 to 5, which is slightly lower than the variation among the TFs themselves (figure 2).

We also note that the \(\tau\) curve associated with two TFs has an unusual step-like shape relative to the others (MA0829.2 and MA1728.1). This is because the target distribution (both auxiliary and consensus) is skewed between the chromosome arms. If starting on one side, TFs have difficulty crossing over to the other because the centromere acts as a barrier, leading to shorter search times on one side than the other. This observation manifests even on the macroscopic TF level, where we observe a broad \(\tau\) distribution (e.g., MA0829.2 in figure 2).
Figure 3. Sorted median search times $\tau$ to consensus target bins for six TFs. Each panel shows the median first-passage time $\tau$ (solid) with the corresponding 95% CI (shaded area). Each data point is calculated from $\geq 1000$ realisations from random starting positions on the Hi-C network. While some TFs show larger variation than others, overall, we note that $\tau$ varies significantly between consensus sites. Some TFs even have step-like curves (MA0829.2 and MA1728.1). This shape reflects that targets are unevenly distributed between the chromosome arms.

3.3. Node flux explains the median first-passage time to most target bins

The last section showed that the median search times to individual consensus target bins vary significantly (figure 3). To better understand why some target configurations have shorter search times than others, we analyzed the consensus-auxiliary binding site network. Particularly, we studied the link weights associated with nodes containing at least one consensus target. As we elaborated in section 2.3, we define the weight as the number of Hi-C contacts $v_{ij}$ (equation (5)).

First, we analyzed the node strength, the sum of all link weights to node $i$, $\sum_j v_{ij}$. However, since the consensus bins act as absorbing targets, we exclude consensus-consensus connections and define the (partial) node strength as:

$$S_i = \sum_{q \neq \text{consensus target bins}} v_{iq},$$

(12)

This metric reflects how tightly the rest of the network connects to consensus target bin $i$, thereby helping direct searching proteins to designated chromatin regions (a large fraction of nodes contains at least one auxiliary binding site).

To better see how the node strengths change with $\tau$, we renormalized the $S_i$ values associated with all consensus target bins and TFs and superimposed all data points in one plot (figure 4(a)). To renormalize the data, we transformed the node strengths and median first-passage times so that $\bar{S}_i$ and $\bar{\tau}$ belong to a standard normal distribution $[N(0, 1)]$ after noticing that they are approximately distributed as a bell curve (appendix F). In figure 4(a), we show all data as a scatter plot, where each point represents a consensus target bin, and the symbols and colors indicate different TFs. Because the data forms a nearly vertical cloud, although with a minor negative tilt, we conclude that there is no significant relationship between median search times and node strength (the spearman correlation coefficient is $-0.26$).

To find another network metric that better explains $\bar{\tau}$, we considered the node flux $F_i$. This quantity considers the time spent in a bin before leaving in addition to the link weight. That is:

$$F_i = \sum_{q \neq \text{consensus target bins}} \frac{v_{iq}}{\tau_{iq}},$$

(13)
where $\tau_q$ is the median time until the TF escapes node $q$, which we denote persistence time. This means that each term $v_{iq}/\tau_q$ represents the effective rate of going from node $q$ to consensus node $i$; the sum is the total rate.

To estimate $\tau_q$, we simulated sequence-specific search until unbinding (by $k_{off}$) in individual Hi-C bins, for all TFs (see appendix E). Our simulations show that there is an exponential dependence between $\tau_q$ and the number of effective targets per bin, where the exact functional form is TF-dependent (see appendix D). So when calculating the node flux in equation (13), we estimated $\tau_q$ from relationships like figure 9 and used $v_{iq}$ from the Hi-C data.

Next, we plotted the normalized node fluxes $\bar{F}_i$ against the simulated median search time to consensus target bins $\bar{\tau}$. Compared to the node strengths $\bar{S}_i$, we find a much stronger correlation: the Spearman coefficient is $-0.72$. Such a large correlation indicates that high node fluxes are reliable predictors for fast search times. Like $\bar{S}_i$ and $\bar{\tau}$, we normalized the node fluxes so that $\bar{F}_i \sim N(0, 1)$ (see appendix F).

Figure 4 illustrates the relationships between $\bar{\tau}$ and $\bar{S}_i$, and $\bar{\tau}$ and $\bar{F}_i$, respectively, for chromosome 19. But what about other chromosomes having a different 3D organization? Do similar relationships exist them? To investigate this question, we made simulations for chromosomes 10 and 20 and reproduced figure 4 in figure 12. While the Spearman coefficients differed slightly from chromosome 19, the correlations remain.

### 3.4. Optimal Network configurations for target-search

We found that the median search time depends on the node flux. Now we ask: could we achieve an even lower search time than in the original configuration if we rearranged the node flux weights? If so, how much is the original (‘wild type’) deviating compared to other random counterparts?

While it is straightforward to scramble the escape rates in equation (13) to calculate $\bar{F}_i$s, we must do new simulations to get $\bar{\tau}$ in systems with reshuffled persistence times. To this end, we randomized DNA sequences between the Hi-C bins while keeping the consensus node fixed. We call these randomized configurations ‘remapped’, as we assign DNA sequences to Hi-C bins that differ from the standard genome sequence (e.g., hg 19 or hg 28). This is similar to TrIP–Chip, where experimenters insert DNA snippets at random DNA loci to study functional response to genomic positions [41].

After randomizing the bin sequences, we simulated TF searches, calculated median first-passage times, and compared these medians to the original arrangement. Using this remapping scheme, we found that the original configuration may differ as much as one order of magnitude relative to the randomized cases. Sometimes, the original configuration stands out as much slower or faster than the remapped systems. We portray the data in figure 5.

Each figure panel shows heatmaps of median search times to the consensus target bins for five TFs. Each row represents a sorted list of the median search times $\tau^{\text{remap}}$ to one consensus target bin in 20 re-mapped configurations. The colors indicate the relative difference between the median first-passage times and the original configuration $\tau^{\text{remap}}/\tau$. From the color spectrum, we note that the remaps vary as much as between different consensus target bins (figure 3) and even between different TFs (figure 2). The heatmaps show that...
Figure 5. Median first-passage times for original and remapped systems associated with five TFs (one for each heat map). Each heat map row represents one consensus target bin, where each pixel color, blue-to-red (see colour bar), indicates high to low median search time compared to the original configuration $\tau$. We sorted faster remaps ($\tau_{\text{remap}} < \tau$) so that they fall to the left of the midpoint (note the tick marks) representing the original (i.e., where $\tau_{\text{remap}} / \tau = 1$). Slower remaps $\tau_{\text{remap}} > \tau$ end up on the right side. For all TFs, we note that some consensus targets experience a substantial shift in $\tau_{\text{remap}}$ when randomizing the sequences among the bins.

Figure 6. The relative change in search times $\tau_{\text{remap}} / \tau$ and node flux $F_{i_{\text{remap}}} / F_{i}$ for remapped systems. The plot shows the relative change in median first-passage times and node flux between the remapped and non-remapped systems. Note the logged values. Each point represents a remap for a consensus target bin, i.e., 20 times as many points as in figure 4, with colours and markers separating different TFs. The solid line shows a linear fit, with the corresponding 95% CI (shaded area). As in figure 4, we only look at the top 20% of TFs.

the network positions of the consensus target bins are just as important as the TF or target type (e.g., sequence length).

We also note that some remapped configurations are associated with much shorter or longer search times than the original configuration. This is best seen by following the heatmaps’ midpoints that indicate where the ratio $\tau_{\text{remap}} / \tau = 1$. All remaps associated with longer search times extend to the right of the midpoint, and those having shorter ones are to the left. In most cases, the division is not halfway. Instead, the distribution is skewed, where the original may be the fastest among the remaps (bottom row), and in other cases it is the slowest (top rows). For example, in the second panel from the left (MA08292.2), the heatmap forms two blocks (one upper and one lower). The consensus bins in the upper block are hard to find when put in the original configuration. But after randomization, these move to more easy-to-find network locations, yielding $\tau_{\text{remap}} < \tau$. The opposite happens in the lower block, where consensus bins become harder to find after randomization.

Finally, to validate that the relationship between median search time and node flux still holds (as we uncovered in figure 4), we computed $\mathcal{F}_i$ and $\tau$ for all remapped configurations ($\sim 10^4$ data points). We also re-scaled $\mathcal{F}_i$ and $\tau$ with the original configuration’s values, $\tau_{\text{remap}} / \tau$ and $\mathcal{F}_{i_{\text{remap}}} / \mathcal{F}_i$. We plot these quantities in figure 6. As we expect, the data points (one TF per color) and the regression curve (green) show a negative correlation also for the remapped data, where the spearman correlation coefficient is slightly less negative.
than for the original configurations (−0.6 vs. −0.72). However, this correlation blends remapped data from multiple TFs, where some are more affected by remaps than others, note the blue dots compared to the other dots. Nevertheless, making separate regression lines, all show convincing negative slopes, albeit with slightly different tilts.

4. Conclusions

The most common mechanism to regulate a gene is when a protein binds to a regulatory sequence to increase or decrease RNA transcription rates. However, these sequences are short relative to the genome length, so finding them poses a demanding search problem. Another problem is that there exist many nearly identical sequences. These almost-targets lie scattered across the genome, potentially causing proteins to get stuck. But as shown in a theoretical study in *E. coli*, such auxiliary binding sites may speed up the search if considering DNA looping. In this paper, we study if this phenomenon occurs on a chromosomes-wide scale, where DNA folding may help channel DNA-binding proteins, like TFs, to designated target regions. To this end, we develop a cross-scale approach that bridges established facilitated-diffusion models for the basepair-level search and chromosome-wide models capturing leaps spanning hundreds of millions of base pairs. As input to our cross-scale model, we used Hi-C data sets as a proxy for 3D proximity between long-ranged DNA segments (Hi-C bins) and binding profiles for more than 100 TFs. In most cases, TFs have 10–1000 experimentally verified binding positions (consensus targets) and ~10^3–10^4 almost-targets (auxiliary binding sites) scattered across a chromosome. When coarse-grained into Hi-C bins, this amounts to ~0–10 targets or binding sites per bin (sometimes there are >100 auxiliary binding sites).

When studying the search to any Hi-C bin with at least one consensus target, we find that the targets’ spatial distribution critically determines the variability among the TFs’ median search time. But rather than occupying specific bins (nodes), it is more important how evenly the consensus targets spread over the DNA-contact network. Thus, considering the targets in the aggregate, the network’s connectivity and the spatial arrangement of the auxiliary binding sites seem to play a minor role in target search.

This finding contrasts search times to single targets that likely are more relevant for gene-specific transcription regulation. For individual consensus-target bins, we found that the median search time is strongly associated with the node flux. The node flux generalizes the classical node strength (sum of weights to a node) as it combines the weight (i.e., the probability that two DNA segments form a long-ranged loop) and the dissociation rate (inverse persistence time). Across most TFs, high node flux is a strong indicator for short median search times.

We also found that the relative position of the bins containing auxiliary binding sites significantly affects search times for the consensus bins. To show this, we randomized the sequences among the Hi-C bins keeping the 3D contacts intact and compared search times to the one for the actual target configuration. We noted a broad spectrum of outcomes. Sometimes, the natural arrangement was much faster than most random variations. In others, it was slower. It would be instructive to investigate the biological significance of fast versus slow configurations to better understand if the chromosomes’ 3D organization funnels critical TFs to designated DNA regions. Since funneling seem to have evolved on the base pair level [10], similar driving forces could have contributed to DNA’s 3D organization.

Furthermore, our model assumes that the searcher may recognize any part of the DNA sequence. However, this assumption potentially needs modification if considering the coverage of epigenetic domains that tesselate the chromosome. These domains are associated with varying DNA accessibility, measured by, for instance, DNase-seq or ATAC-seq, where repressed chromatin (e.g., H3K27me3) tends to be less accessible than other parts, e.g., actively transcribed DNA. One way to incorporate epigenetic domains into the model is to block or remove parts of the Hi-C map. This reduces the searchable space. Depending on the precise locations of these epigenetic regions, the reduced search space could either speed up or slow down the search. For example, if epigenetic domains cover shortcuts across the network, they could introduce bottlenecks that increase search times. But they could also lower search times if protecting traps and preventing TFs from associating with the ‘wrong’ targets. These two effects offer an interesting antagonistic relationship that would be interesting to explore further.

Another aspect worthy of exploration is hidden targets. In some genomic regions, often epigenetically repressed, DNA (or chromatin) folds so densely that DNA-binding proteins cannot access and recognize the embedded sequence. Using DNase data as a proxy for accessible DNA, one could use our cross-scale model to explore how omitting targets will affect median search times. We leave this study as an open problem for future work.

To conclude, our work provides a framework for studying chromosome-wide search times for specific DNA sequences. Apart from TFs, we hope our results and methods will help other researchers interested in
DNA-protein binding associated with other genetic processes such as gene regulation, DNA repair [42], CRISPR/cas9 gene editing [43, 44], and epigenetics.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

We acknowledge financial support from the Swedish Research Council (Grant No. 2017-03848). We also thank Moa Lundkvist (Umeå University) for helping with the figures and data handling. This research was conducted using the High-Performance Computing Center North (HPC2N), endorsed by the Swedish National Infrastructure for Computing (SNIC), and partially funded by the Swedish Research Council (Grant Agreement No. 2018-05973).

Appendix A. Calculating the energy landscape

To estimate the energy landscape that TFs experience when interacting with DNA, we employ standard methods (e.g., [10]). These methods use the PFM containing the nucleotide occurrence at each position $i$ along the TFs target motif having length $L_{TF}$. The PFM has four rows corresponding to each nucleotide type $s = A, T, C, G$, and $L_{TF}$ columns, each representing positions $i = 0, \ldots, L_{TF}$. We retrieved PFMs from the JASPAR database [22].

To get the TF binding profile $P_{TF}(s, i)$, we converted the PFM $n(s, i)$ into a TF-specific position probability matrix (PPM). For some TFs, there might be counts missing for one or more nucleotides at position $i$. This causes problems because $E \sim \log P$ gives infinite energies if $P = 0$. To avoid this problem, we use additive smoothing [45], yielding:

$$P_{TF}(s, i) = \frac{n_{TF}(s, i) + p_s \pi(a)}{\sum_{s} (n_{TF}(s, i) + p_s \pi(a))},$$

(A1)

where $p_s$ is the pseudo count (we use $p_s = 1$) and $\pi(a) = 1/4$ is a uniform nucleotide probability distribution.

Next, we relate the PPM to the binding energy using Boltzmann’s relationship [46]:

$$\epsilon_{TF}(s, i) = -\ln P_{TF}(s, i).$$

(A2)

where $k_B T = 1$. To calculate the interaction energy at position $x$, we sum over all positions $i$ in $L_{TF}$ (see figure 1):

$$E_{TF}^+(x) = \sum_{i=0}^{L_{TF}-1} \epsilon_{TF}(s, i + x),$$

(A3)

while equation (A3) calculates the forward-looking energy (hence the superscript ‘+’), there is also the reverse motif, $E_{TF}^-(x)$, having a slightly different energy. Assuming that the TF associates randomly to either one side, we pick the lowest binding energy, that is:

$$E_{TF}(x) = \min(E_{TF}^+(x), E_{TF}^-(x)) - \langle \min(E_{TF}^+(x), E_{TF}^-(x)) \rangle_x$$

(A4)

where the negation corresponds to setting the mean energy over the whole DNA sequence to zero [10]. In our cross-scale model, we could include both the or forward and reverse energy landscapes separately. However, this would not lead to significant changes.
Appendix B. Normalizing the energy landscape

Calculating the energy landscape as a sum of varying variables (see appendix A) results in a near Gaussian distribution of $E^{TF}(x)$ [13]. We therefore re-normalize the energy landscape according to:

$$E^{TF}(x) = \frac{E^{TF}(x) - \langle E^{TF}(x) \rangle}{\sqrt{\langle (E^{TF}(x))^2 \rangle - \langle E^{TF}(x) \rangle^2}},$$  \hspace{1cm} (B1)

so that $E^{TF}(x) \sim N(0, 1)$. We show energies for ten TFs in figure 7.

We also used constant $\Delta G = \langle \Delta G \rangle_{TF}$ for each normalized TF, where $\langle \cdot \rangle_{TF}$ is the ensemble average over all TFs.

Appendix C. Median-first passage on dense networks

We showed in section 3.1 that $\tau \sim \phi^{-1}$ for a fully connected network. Here we generalize the derivation to dense but not fully connected networks (like Hi-C.)

First, we denote the probability that the TF jumps to a target bin from node $i$ as:

$$P_i(1) = \sum_{j \in \Omega_t} \omega_{ij} = \phi_i,$$  \hspace{1cm} (C1)

where $\Omega_t$ is the set of all target nodes. Formulated in this way, all $\phi_i$ are different but will fluctuate around some mean $\phi$, that is:

$$\phi_i = \phi + \epsilon_i, \hspace{0.5cm} \epsilon_i \in [-\phi, 1-\phi],$$  \hspace{1cm} (C2)

where $\epsilon_i$ is a random independent variable obeying:

$$\langle \epsilon_i \rangle = 0, \hspace{0.5cm} \langle \epsilon_i \epsilon_j \rangle = \delta_{ij}$$  \hspace{1cm} (C3)

where $\delta_{ij}$ is the Kronecker delta and $\langle \cdot \rangle$ denotes ensemble average. Using these properties gives:

$$\langle \phi_i \rangle = \langle \phi \rangle + \langle \epsilon_i \rangle = \phi.$$  \hspace{1cm} (C4)

Second, we express the probability of making $n > 1$ jumps as:

$$P_i(n) = \phi_i^1 \cdot \phi_i^2 \cdot \phi_i^3 \ldots \phi_i^n.$$  \hspace{1cm} (C5)
Figure 8. Schematic of how close-by binding site reduce their effectiveness. Here we show a list of $N$ sites with auxiliary binding site marked with red dots. For each cluster of binding site, the window $L_{BS}$ stretches $l_i$ from the outermost binding site. If a TF lands within this region, it will most probably hit one of the binding sites. However, if $L_{BS} \approx l_i$, then the cluster effectively only act as one binding site.

Figure 9. The number of effective binding sites based on the number of binding sites, comparing data and theory. The two lines in orange and blue are calculated using equation (D3) using two different sliding lengths $l_i$. The green line is calculated from binding site placement identified using MOODS for 100 TFs, see section 2.2.2. We calculated the number of effective binding sites using the mean sliding length of 90 bp. The curve from the data does not follow the 90 bp curve but rather the 200 bp curve, indicating that binding sites on the chromosome are more clustered than just by random placement [10].

Using $\phi' = (1 - \phi)$ and $\phi_i' = \phi_i' - \epsilon_i$ yields:

$$P_i(n) = (\phi' - \epsilon_1) \cdot (\phi' - \epsilon_2) \cdot (\phi' - \epsilon_3) \cdots (\phi + \epsilon_n)$$

$$= ((\phi')^{n-1} + (\phi')^{n-2}(-\epsilon_1 - \epsilon_2 - \ldots) + (\phi')^{n-3}(\epsilon_1\epsilon_2 + \ldots) + \ldots) \cdot (\phi + \epsilon_n).$$

(C6)

Averaging over all positions $i$ and using equation (C3) gives:

$$\langle P_i(n) \rangle = ((\phi')^{n-1} \ldots) \langle \phi + \epsilon_n \rangle = \phi^{n-1}\phi = (1 - \phi)^{n-1}\phi.$$

(C7)

This equation is similar to equation (7) from which the rest of the argumentation leading up to $\tau \sim \phi^{-1}$ is identical.

Appendix D. Effective binding sites

Auxiliary binding sites sequester the TF stronger than other positions along the sequence. If close together, such sites may act as one if they lie within the TFs sliding length $l_i$ (see figure 8). In this section, we calculate the relationship between the number of effective binding sites ($\Sigma_{eff}$) and the number of actual binding sites ($\Sigma$) across the Hi-C bins. As in figure 8, we define regions $L_{BS}^{i}$ of length $2l_i$ centered around each binding site. To calculate the number of effective binding sites, we sum $L_{BS}^{i}$ and divide by twice the sliding length:

$$\Sigma_{eff} = \left\lceil \frac{1}{2l_i} \sum_{i} L_{BS}^{i} + \frac{1}{2} \right\rceil,$$

(D1)

where $\lfloor \cdot + 1/2 \rfloor$ denotes rounding half up ($\lfloor \cdot \rfloor$ is the round-down function). As we merge $L_{BS}^{i}$, the limit is easy to compute:

$$\lim_{\Sigma \to \infty} \Sigma_{eff} = \lim_{\Sigma \to \infty} \left\lfloor \frac{1}{2l_i} \sum_{i} L_{BS}^{i} + \frac{1}{2} \right\rfloor = \frac{N}{2l_i}.$$

(D2)
The persistence time $t_q$ for Hi-C bins versus two binding site counts (total and effective number of binding sites). The persistence time is calculated from 100 simulations per Hi-C bin. Each point is aggregated from all bins and all TFs. We see that the green line, which counts the number of binding sites in each bin, shows a weak increase in the persistence time as the count increases. However, in the orange line, which considers effective binding sites, we see a much stronger increase as the binding site count increases.

We calculate this measure analytically by splitting the region of length $N$ into equally sized boxes $N_b = N/2l_s$. Next, if randomly placing $\Sigma$ binding sites, the expected number of boxes with only one binding site is:

$$
\Sigma_{\text{eff}}(\Sigma) = \sum_{i=0}^{\Sigma-1} (-1)^i \binom{\Sigma}{i+1} \frac{1}{(N_b)}.
$$

If the size of each box is $N_b = 2l_s$, two binding sites randomly placed in the same box corresponds to two binding sites sharing one $L_{BS}$. As such, equation (D3) approximately corresponds to the number of effective binding sites as given by equation (D1). We plot equation (D3) for two different sliding lengths against the actual results from the data. We see that the real data is more clustered than expected, yielding an 'effective' sliding length longer than the actual sliding length.

**Appendix E. Residence time on 5 kb regions**

In appendix D, we quantified the number of effective binding sites in each Hi-C bin. Here we quantify the relationship between the number of effective binding sites and the persistence time $t_q$, i.e., the time TFs spend in a Hi-C bin before disassociating.

Using our simulation framework, we place a TF in a random Hi-C bin and simulate sequence-specific search until unbinding by $k_{\text{off}}$ (see section 2.1). In figure 10, we plot $t_q$ versus two measures of binding site count, noting that both metrics show an apparent increase in persistence time as the count increases. As explained in appendix D, only counting the number of binding sites gives a weaker increase. However, looking at the number of effective binding sites, we see a more substantial persistence time growth. This shows that clustered binding sites are less effective in sequestering TFs than when spread out.

**Appendix F. Distribution of node flux, node strength, and the median persistence time**

To simplify the visualization over many TFs, we renormalized $S_i, F_i$ and $\tau$. As we show in figure 11, all these variables collapse to a standard normal distribution after the transformation:

$$
(X - \langle X \rangle)/\langle X^2 \rangle + \langle X \rangle^2, \quad X = \tau, S_i, F_i.
$$
Figure 11. Density distributions of median first-passage time $\tau$, node strength $S_i$ and node flux $F_i$ after Gaussian renormalization (equation (1)). Each colour represents a different TF. As solid line, we show the standard normal distribution $N(0,1)$.

Appendix G. Node fluxes for chromosomes 10 and 20

Figure 12. Figure 4 recreated using chromosome 20 and 10. We see that even though the Spearman correlation coefficient changes between TFs and chromosomes, the correlation to node flux compared to node strength is still clear.

ORCID iDs

Lucas Hedström 🐝 https://orcid.org/0000-0002-3315-0633
Ludvig Lizana 🐝 https://orcid.org/0000-0003-3174-8145

References

[1] Hammar P, Leroy P, Mahmutovic A, Marklund E G, Berg O G and Elf J 2012 The lac repressor displays facilitated diffusion in living cells Science 336 1595–8
[2] Marklund E, Mao G, Yuan J, Zikrin S, Abdurakhmanov E, Deindl S and Elf J 2022 Sequence specificity in DNA binding is mainly governed by association Science 375 442–5
[3] Bauer M, Rasmussen E S, Lomholt M A and Metzler R 2015 Real sequence effects on the search dynamics of transcription factors on DNA Sci. Rep. 5 10072
[4] Cortini R and Filion G J 2018 Theoretical principles of transcription factor traffic on folded chromatin Nat. Commun. 9 1–10
[5] Ikon E Z, Stampfel G, Yáñez-Cuna J O, Dickson B J and Stark A 2012 Hot regions function as patterned developmental enhancers and have a distinct cis-regulatory signature Genes Dev. 26 908–13
[6] Brackley C A, Cates M E and Marenduzzo D 2012 Facilitated diffusion on mobile DNA: configurational traps and sequence heterogeneity Phys. Rev. Lett. 109 168103
[7] Rao S S P et al 2014 A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping Cell 159 1665–80
[8] Fornes O et al 2020 Jaspar 2020: update of the open-access database of transcription factor binding profiles Nucleic Acids Res. 48 1987–1992
[9] Khoronen J, Martinmäki P, Pizzi C, Rastas P and Ukkonen E 2009 Moods: fast search for position weight matrix matches in DNA sequences Bioinformatics 25 3181–2
[10] Cencini M and Pigolotti S 2017 Energetic funnel facilitates facilitated diffusion Nucleic Acids Res. 46 558–67
[11] Benichou O, Lévy C, Moreau M and Voituriez R 2011 Intermittent search strategies Rev. Mod. Phys. 83 81
[12] Lomholt M A, van den Broek B, Kalisch S-M J, Wuite G J L and Metzler R 2009 Facilitated diffusion with DNA coiling Proc. Natl Acad. Sci. 106 6204–8
[13] Mirny L, Slutsky M, Wunderlich Z, Tafvizi A, Leith J and Korsmjr A 2009 How a protein searches for its site on DNA: the mechanism of facilitated diffusion J. Phys. A: Math. Theor. 42 434013
[14] Felipe C, Shin J and Kolomeisky A B 2021 DNA looping and DNA conformational fluctuations can accelerate protein target search J. Phys. Chem. B 125 1727–34
[15] Nyberg M, Ambjörnsson T, Stenberg P and Lizana L 2021 Modeling protein target search in human chromosomes Phys. Rev. Res. 3 013055
[16] Bagchi B, Blainey P C and Sunney Xie X 2008 Diffusion constant of a nonspecifically bound protein undergoing curvilinear motion along DNA J. Phys. Chem. B 112 6282–4
[17] Murugan R 2010 Theory of site-specific DNA–protein interactions in the presence of conformational fluctuations of DNA binding domains Biophys. J. 99 353–9
[18] 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 6961–77
[19] Dixon J R, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu J S and Ren B 2012 Topological domains in mammalian genomes identified by analysis of chromatin interactions Nature 488 376–80
[20] Lee S H, Kim Y, Lee S, Durang X, Stenberg P, Joen J-H and Lizana L 2019 Mapping the spectrum of 3D communities in human chromosome conformation capture data Sci. Rep. 9 1–7
[21] Farnham P J 2009 Insights from genomic profiling of transcription factors Nat. Rev. Genet. 10 605–16
[22] JASPAR 2020: an open-access database of transcription factor binding profiles (available at: http://jaspar.genereg.net/)
[23] Jaspar documentation (available at: https://jaspar.genereg.net/docs/) (Accessed 07 November 2022)
[24] Stormo G D, Schneider T D, Gold L and Ehrenfeucht A 1982 Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in E. coli Nucleic Acids Res. 10 2997–3011
[25] Lieberman-Aiden E et al 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome Science 326 289–93
[26] Sanborn A L et al 2015 Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes Proc. Natl Acad. Sci. 112 66456–65
[27] Slutsky M and Mirny L A 2004 Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential Biophys. J. 87 4021–35
[28] van den Broek B, Lomholt M A, Kalisch S-M J, Metzler R and Wuite G J L 2008 How DNA coiling enhances target localization by proteins Proc. Natl Acad. Sci. 105 15738–42
[29] Amitai A 2018 Chromatin configuration affects the dynamics and distribution of a transiently interacting protein Biophys. J. 114 766–71
[30] Hu T, Grosberg A Y and Shklovskii B I 2006 How proteins search for their specific sites on DNA: the role of DNA conformation Biophys. J. 90 2731–44
[31] Lomholt M A, Ambjörnsson T and Metzler R 2005 Optimal target search on a fast-folding polymer chain with volume exchange Phys. Rev. Lett. 95 260603
[32] Smrek J and Grosberg A Y 2015 Facilitated diffusion of proteins through crumpled fractal DNA globules Phys. Rev. E 92 012702
[33] James Kent W, Sugnet C W, Furey T S, Roskin K M, Pringle T H, Zahler A M and Haussler A D 2002 The human genome browser Nucleic Acids Res. 30 3896–901
[34] Kudo K, Xi Y, Wang Y, Song B, Chu E, Ju J, Russo J J and Ju J 2010 Translational control analysis by translationally active RNA capture/microarray analysis (TRP-Chip) Nucleic Acids Res. 38 e104
[35] Esadze A and Stivers J T 2018 Facilitated diffusion mechanisms in DNA base excision repair and transcriptional activation Chem. Rev. 118 11298–323
[43] Globyte V, Hwan Lee S H, Bae T, Kim J-S and Joo C 2019 CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion EMBO J. 38 e99466
[44] Lu Q, Bhat D, Stepanenko D and Pigolotti S 2021 Search and localization dynamics of the CRISPR-Cas9 system Phys. Rev. Lett. 127 208102
[45] Schütze H, Manning C D and Raghavan P 2008 Introduction to Information Retrieval vol 39 (Cambridge: Cambridge University Press)
[46] Stormo G D and Fields D S 1998 Specificity, free energy and information content in protein–DNA interactions Trends Biochem. Sci. 23 109–13