



Modeling protein target search in human chromosomes

Markus Nyberg ¹, Tobias Ambjörnsson,² Per Stenberg,³ and Ludvig Lizana ^{1,*}

¹*Integrated Science Lab, Department of Physics, Umeå University, SE-901 87 Umeå, Sweden*

²*Department of Astronomy and Theoretical Physics, Lund University, SE-223 62 Lund, Sweden*

³*Department of Ecology and Environmental Science, Umeå University, SE-901 87 Umeå, Sweden*



(Received 19 September 2019; accepted 16 November 2020; published 19 January 2021)

Several processes in the cell, such as gene regulation, start when key proteins recognize and bind to short DNA sequences. However, as these sequences can be hundreds of million times shorter than the genome, they are hard to find by simple diffusion: diffusion-limited association rates may underestimate *in vitro* measurements up to several orders of magnitude. Moreover, the rates increase if the DNA is coiled rather than straight. Here we model how this works *in vivo* in mammalian cells. We use chromatin-chromatin contact data from Hi-C experiments to map the protein target-search onto a network problem. The nodes represent DNA segments and the weight of the links are proportional to measured contact probabilities. We then put forward a diffusion-reaction equation for the density of searching protein that allows us to calculate the association rates across the genome analytically. For segments where the rates are high, we find that they are enriched with active gene starts and have high RNA expression levels. This paper suggests that the DNA's 3D conformation is important for protein search times *in vivo* and offers a method to interpret protein-binding profiles in eukaryotes that cannot be explained by the DNA sequence itself.

DOI: [10.1103/PhysRevResearch.3.013055](https://doi.org/10.1103/PhysRevResearch.3.013055)

I. INTRODUCTION

Several processes in the cell nucleus start when proteins bind to specific DNA sequences. For example, transcription factors that regulate genes and the CRISPR/CAS9 complex that edits DNA [1,2]. Because target sequences are much shorter than the genome—a few base pairs compared to billions in humans—these proteins face a needle-in-a-haystack problem.

Despite the large number of potential targets, measured search times are shorter than theoretical estimates. The Lac repressor in *E. coli* needs 1–5 min to find its designated site [3] which is twice as fast as a three-dimensional (3D) diffusive search inside the bacterium's volume (≈ 2 –11 min).¹ Also, diffusion-limited association rates—Smoluchowski's rate—may underestimate *in vitro* measurements by one to two orders of magnitude [4]. These examples suggest that some proteins search by other mechanisms than simple diffusion.

One mechanism that speeds up the search is offered by the Facilitated-diffusion model [5]. In this model, the proteins alternate between 3D diffusion and 1D diffusion along the DNA. This lowers the search time because the proteins may take shortcuts through the surrounding bulk to linearly distant

DNA segments. Although criticized [6,7], the model is widely accepted after experiments in bacteria [3,8] and *in vitro* [9].

Another important aspect of target finding is rebinding. This is because proteins likely bind to a DNA segment that is close by in 3D rather than far away. Several modeling studies examined this aspect and found that search times change with DNA conformation [7,9–14]. However, because these studies treat the DNA using standard polymer models, the results cannot be generalized beyond bacteria to eukaryotes that have longer DNA with a more complex 3D structure.

The most widely used experimental method to study the 3D organization of the genome is Hi-C [15,16]. The Hi-C method cross-links close by DNA fragments inside the nucleus and gives a genome-wide map of the number of contacts between fragment pairs [Fig. 1(a)] [17]. Mammalian Hi-C maps have several interesting features where some are evolutionary conserved [18]. For example, the blocklike structure along the diagonal represents densely connected 3D domains. The locations of these domains correlate with protein binding sites, active genes, and chromatin states [19–21].

Hi-C is the state-of-the-art Chromosome Conformation Capture method that estimates the chromatin contact probabilities across the genome. However, it does not provide chromatin's 3D structure. Going from the contact map to a computer-generated 3D structure is difficult [22,23].

Because chromatin's spatial organization is so complex, there are but a few attempts to model protein search in eukaryotes. One exception [24] represents chromatin as a crumpled polymer globule. However, while it reproduces the average looping probabilities measured in human the crumpled globule lacks 3D domains.

We offer a new approach to the DNA-search problem in eukaryotes that does not rely on chromatin's explicit 3D structure. Instead, similar to Refs. [21,25,26], we incorporate the

*ludvig.lizana@umu.se

¹Diffusion-limited search time, $\tau = 4\pi aD/V$ where $D = 0.1$ – $0.5 \mu\text{m}^2/\text{s}$, $a = 5 \text{ bp}$ ($=1.3 \text{ nm}$), and $V = 1 \mu\text{m}^3$. These values give $\tau = 2.1$ – 10.7 min .

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by Bibsam.

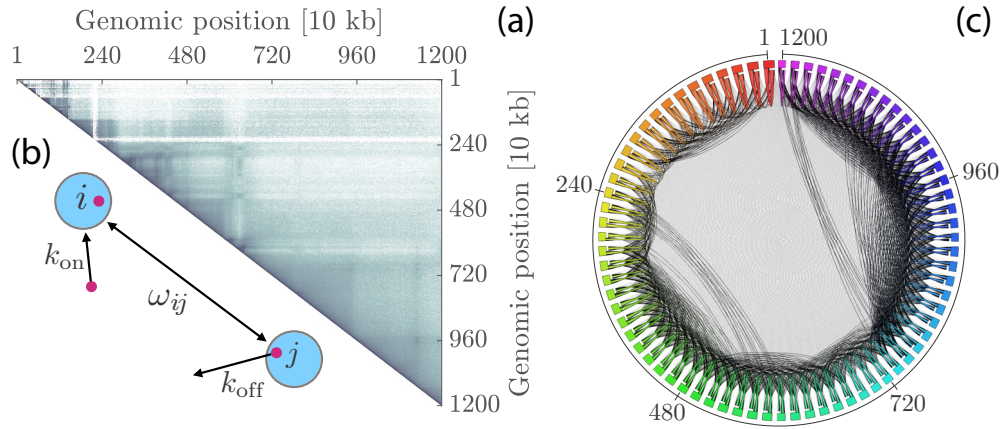


FIG. 1. Modeling protein search on DNA as search on a weighted network. (a) Hi-C map: the colors represent the number contacts (in \log_{10}) between 10 kilo-basepair (kb) DNA fragments in a part of human chromosome 21. Dark pixels indicate many contacts (dynamic range: $\sim 10^0$ – 10^5). (b) Schematic representation of the model. Key parameters: jump rates between nodes i and j (ω_{ij}), unbinding rate to the bulk (k_{off}), and rebinding rate (k_{on}). Red circles represent searching proteins. (c) Coarsened network representation of the Hi-C map in (a). Each node represents a 160 kb fragment. The link weights v_{ij} are proportional to the number of Hi-C contacts. We assume that $\omega_{ij} \propto v_{ij}$. Node numbering refers to positions along the DNA.

3D structure by representing the DNA as a network in which the nodes are DNA segments, and the link weights are the contact probabilities measured in Hi-C. Then we put forward a diffusion-reaction equation for the protein density on the network that allows us to calculate the association rate—the inverse mean-first passage time—to all nodes analytically. Correlating these rates with genetic data in humans, we find that easy-to-find loci, such as gene starts, reside in regions with active transcription.

II. THE MODEL

We model the proteins' search on chromatin as noninteracting particles that move between nodes in a weighted network that represents physically connected chromatin segments (Fig. 1). The model has three parameters. First, the jump rate ω_{ij} between nodes i and j ($i, j = 1, \dots, N$). We assume that ω_{ij} equals the probability p_{ij} to jump between segments i and j multiplied by the frequency of a successful jump (collision frequency) f_{coll} . As a proxy for p_{ij} , we use the number of Hi-C contacts v_{ij} . That is, $\omega_{ij} = f_{\text{coll}} v_{ij}$, where we treat f_{coll} as a free parameter setting the time-scale in our problem.

The second parameter is the binding rate \bar{k}_{on} to a randomly chosen node. Assuming that the protein bulk concentration c_{bulk} is constant, we may use that $k_{\text{on}} = \bar{k}_{\text{on}} c_{\text{bulk}}$; The third parameter is the unbinding rate k_{off} to the surrounding bulk. k_{on} and k_{off} have the unit: time^{-1} .

For clarity, we use population-averaged Hi-C data. As such, we lack cell-to-cell variability and some transient loops. In the model, we therefore envision the chromosomes as a rigid structures where the probability to jump from one segment to another is proportional to the number of contacts.

Based on these parameters, we formulate a diffusion-reaction equation for the protein number in node i at time t ,

$n_i(t)$:

$$\frac{dn_i(t)}{dt} = \sum_{j=1}^N \omega_{ij} n_j(t) - k_{\text{off}} n_i(t) + k_{\text{on}}. \quad (1)$$

The first term represents diffusion on the network—we put $\omega_{jj} = -\sum_{i \neq j} \omega_{ij}$ —and the two remaining terms describe the exchange with the bulk.

We let one node in the network, $i = a$, represent a target. As we focus on what happens up until it is reached, we treat the target as an absorbing node, $n_a(t) = 0$, that cannot be blocked by other searchers.

In terms of the eigenvalues λ_j and eigenvectors V_{ij} of ω_{ij} , the solution to Eq. (1) is

$$n_i(t) = \sum_{j=1}^N V_{ij} \left\{ \frac{k_{\text{on}}^j}{k_{\text{off}} - \lambda_j} [1 - e^{-(k_{\text{off}} - \lambda_j)t}] + e^{-(k_{\text{off}} - \lambda_j)t} \sum_{l \neq a} V_{jl}^{-1} n_l(0) \right\}, \quad (2)$$

where $k_{\text{on}}^i = k_{\text{on}} \sum_{j=1}^N V_{ij}^{-1}$ and $n_l(0)$ is the initial protein number concentration.

III. PROTEIN ASSOCIATION RATES

To calculate the association rate to target node a , K_a , we use that the rate is one over the mean first arrival time: $K_a = \tau_a^{-1}$. To obtain τ_a , first we calculate the number of particles that arrived to the target up to t :

$$J_a(t) = \int_0^t j_a(t') dt, \quad j_a(t) = \sum_i \omega_{ai} n_i(t) + k_{\text{on}}, \quad (3)$$

where $j_a(t)$ is the particle flux. The flux has two contributions. The first term describes particles that find the target from another DNA site, and the second term describes those that reach the target from the bulk.

Second, if N_p is the initial protein number on the network and $k_{\text{on}} = k_{\text{off}} = 0$, then $J_a(t)/N_p$ is the probability that one protein reached the target up to time t . The probability that the target has not been reached by any protein—the target’s survival probability—is therefore $S_a(t) = (1 - J_a(t)/N_p)^{N_p}$ and $\tau_a = \int_0^\infty S_a(t) dt$. Generalizing this argument for $k_{\text{on}} > 0$, the number of proteins $N_p \rightarrow \infty$ and therefore $S_a(t) \simeq \exp(-J_a(t))$ [27].

Finally, using $n_i(t)$ from Eq. (2), we get

$$K_a = \frac{1}{\int_0^\infty \exp(-J_a(t)) dt}, \quad (4)$$

$$J_a(t) = k_{\text{on}} t + \sum_{i \neq a} \omega_{ai} \int_0^t n_i(t') dt'. \quad (5)$$

A. Limiting cases for K_a

Depending on the unbinding rate k_{off} , K_a has three regimes.

(i) Small k_{off} . In this regime, most particles find the target before they unbind. This leaves the initial density approximately unchanged, $n_i(0) \simeq \rho_0(1 - \delta_{ia})$, where $\rho_0 = k_{\text{on}}/k_{\text{off}}$. Using this approximation in Eq. (5) leads to $J_a(t) \simeq \bar{J}_a t$, where $\bar{J}_a = k_{\text{on}} + \rho_0 \sum_{i \neq a} \omega_{ai} \equiv k_{\text{on}} + \rho_0 W_a$. Thus, $K_a \simeq \bar{J}_a$.

(ii) Large k_{off} . Here, the particles unbind and rebound many times before finding the target. The protein density is therefore approximately in steady-state $\bar{\rho}_a = k_{\text{on}} \times [k_{\text{off}} + W_a/(N-1)]^{-1}$ (see Appendix C 2). Using $n_i(t) \simeq \bar{\rho}_a$ and proceeding as in (i) gives $K_a \simeq k_{\text{on}} + \bar{\rho}_a W_a$.

To simplify our equations, we rescale all rate constants in terms of f_{coll} and a constant factor that ensures that the genome-wide average K_a is unity (see Appendix C). We define the average as $\langle X \rangle = (1/N_G) \sum_{i=1}^{N_G} X_i$, where $N_G \gg N$ is the number of nodes for all chromosomes. Then we denote the rescaled variables as $\hat{K}_a = K_a/\langle K_a \rangle$, $\hat{k}_{\text{on}} = k_{\text{on}}/\langle K_a \rangle$, and $\hat{k}_{\text{off}} = k_{\text{off}}/\langle K_a \rangle$. After rescaling, the regimes (i) and (ii) simplifies $\hat{k}_{\text{off}} \ll 1$ and $\hat{k}_{\text{off}} \gg 1$:

$$\hat{K}_a \simeq \hat{k}_{\text{on}} + \gamma_a V_a, \quad (6)$$

in which $V_a = \sum_{i \neq a} v_{ai}$ is the node strength, and

$$\gamma_{a1} = \frac{1 - \hat{k}_{\text{on}}}{\langle V_a \rangle}, \quad \hat{k}_{\text{off}} \ll 1, \quad (7)$$

$$\gamma_{a2} = \frac{\hat{k}_{\text{on}} \gamma_{a1}}{\hat{k}_{\text{on}} + V_a \gamma_{a1}/(N-1)}, \quad \hat{k}_{\text{off}} \gg 1. \quad (8)$$

Because Hi-C matrices are large, Eq. (6) offers a huge improvement compared to evaluating Eqs. (2), (4), and (5) directly. In Appendix D, we show that Eq. (6) holds for a broad range of \hat{k}_{off} .

(iii) Intermediate \hat{k}_{off} . When $\hat{k}_{\text{off}} \sim 1$, we cannot use Eq. (6). Instead we must evaluate Eqs. (4) and (5). In Appendix B, we also treat the case $k_{\text{on}} = k_{\text{off}} = 0$.

B. Protein association rates depend on chromatin’s 3D organization

Equation (6) suggests that the association rates change with chromatin’s 3D structure because \hat{K}_a depends on the node strength V_a . To quantify by how much, we used Hi-C data from human cell line GM12878 [28] (40 kb resolution) and

calculated \hat{K}_a ($\hat{k}_{\text{off}} \ll 1$) for chromosomes 1-21 (Fig. 2). We found that \hat{K}_a varies by several orders of magnitude relative to the genome-wide average $\langle \hat{K}_a \rangle = 1$. Most \hat{K}_a values, however, are close to the mean: $\hat{K}_a = 1 \pm 0.0027$ (95% confidence interval).

Equation (6) also suggests that chromatin’s 3D structure becomes less important as the unbinding \hat{k}_{off} is large, for example if the bulk concentration is high ($\hat{k}_{\text{on}} \propto n_{\text{bulk}}$). We see this for small V_a where $\hat{K}_a \simeq \hat{k}_{\text{on}}$ (Appendix F). We interpret this as if the particles reach the target mostly from the bulk. However, for small \hat{k}_{on} , we see that $\hat{K}_a \propto V_a$. This means that most particles find the target via jumps on the network and that the 3D structure is important.

C. Chromatin regions with high association rates are enriched with active genes

Figure 2 shows that the association rate varies across the genome. This is important for regulatory proteins, such as transcription factors, that look for promoters to control transcription. We therefore ask: are promoter regions easier to find than nonpromoter regions?

To answer this, we downloaded gene annotation data for human cells [28] to extract the gene starts. We considered all genes in the data, protein-coding and noncoding. We defined the starts by the transcription start site (TSS) that is furthest away from the gene’s end. We omitted alternative TSSs.

After we extracted the TSSs, we correlated their positions with the association rates from Fig. 2. We found that the rates grow with the number of gene starts per node (Fig. 3, pink). In the plot, the data points represent the average association rate to all nodes with the same number of gene starts, and the shaded area shows the 95% confidence interval. In other words: regions with a high density of gene starts are easy to find.

Then we asked: because these regions harbour active and inactive gene starts, are active gene-dense regions easier to find than inactive ones? To see this, we grouped the gene starts into two classes. The first group are the TSSs that reside in transcriptionally active regions. We denote these TSSs as “active.” The second group (“inactive”) consists of TSSs that are in transcriptionally inactive regions. To make the classification, we used RNA-seq data from cell line GM12878 and averaged the RNA read count in a region ± 1 kb surrounding each TSS. We define the region as active if the average read count is above one.

Based on this grouping, we found that nodes with many active TSSs have even higher association rates than if we do not separate active from inactive: gray is above pink in Fig. 3. For nodes with inactive regions, we find the inverse relationship suggesting that they are hard to find: green is below pink in Fig. 3.

Figure 3 also shows that the association rate grows slowly beyond one or two TSSs per node: adding a few extra TSSs does not make the node easier to find.

D. Chromatin regions with high RNA expression levels have high association rates

Figure 3 suggests that transcription factors quickly find highly transcribed gene starts. But how does the association

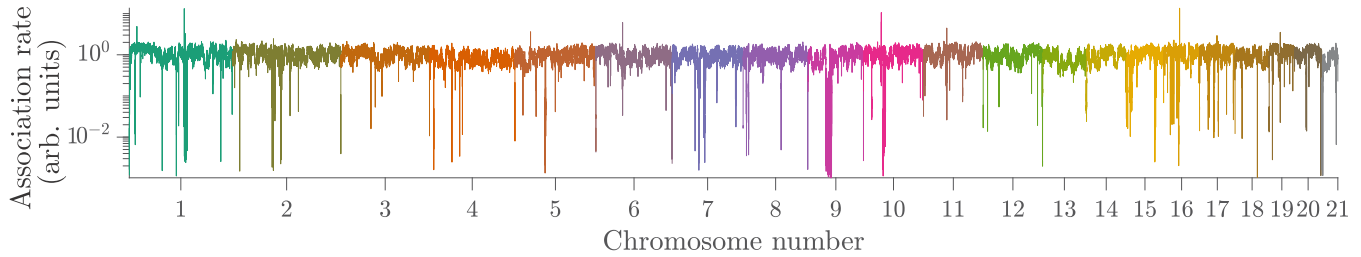


FIG. 2. Predicted genome-wide association rates \hat{K}_a for chromosomes 1-21. \hat{K}_a vary by several orders of magnitude but the 95% confidence interval is a few percent of the mean ($\langle \hat{K}_a \rangle = 1$ ($\hat{K}_a = 1.0 \pm 0.0027$)). We calculated \hat{K}_a from Eq. (6) ($\hat{k}_{\text{off}} \ll 1$) with $\hat{k}_{\text{off}} = 0.002$, $\hat{k}_{\text{on}} = 0.001$, and $\rho_0 = 0.5$. $\hat{k}_{\text{off}} \gg 1$ show similar behavior, see Appendix E

rate correlate with genome-wide expression levels (including potentially unannotated genes).

To study this, we summed the RNA expression in all nodes in the genome and ranked them based on their RNA expression level. Then we partitioned the nodes into 20 equally sized groups and calculated the association rate in each group. Shown as a violin plot [Fig. 4(a)], we find that our predicted rates vary widely but that the median (white circles) increases with high RNA expression levels (Spearman’s correlation coefficient = 0.5449 [29]). This suggests that nodes with high RNA expression levels are relatively easy to find.

To see by how much this correlation is caused by active regions harboring genes, we made two new groups: nodes with at least one active TSS and the rest—nodes with inactive or no TSSs. As before, we ranked the nodes in these large groups based on the RNA expression levels, divided them into 20 equally sized subgroups, and calculated the average association rate for each subgroup. Plotting the predicted average association rate for the two large groups versus the average

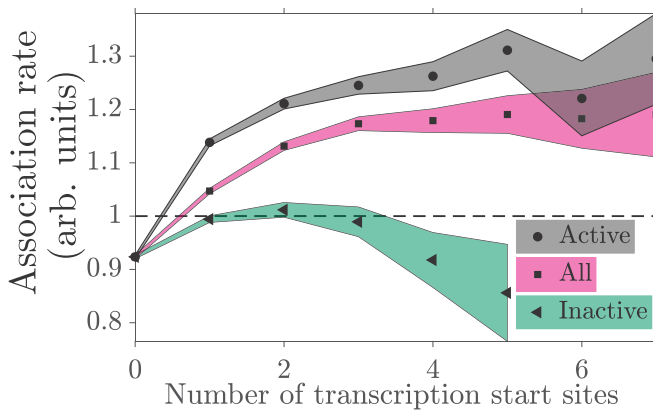


FIG. 3. Nodes with many gene starts have higher predicted association rates than nodes with few gene starts. We define the gene starts as the Transcription Start Sites (TSSs). The curves represent predicted association rates \hat{K}_a to nodes with active TSSs (gray), inactive TSSs (green), and any TSS type (pink). The active TSSs have higher \hat{K}_a than the genome-wide average (dashed), whereas nodes with inactive TSSs (green) are below (except one data point). The symbols represent the average \hat{K}_a (Eq. (6), $\hat{k}_{\text{off}} \gg 1$) and the coloured areas show the 95% confidence interval. Parameters (dimensionless, see Appendix C): $\hat{k}_{\text{off}} = 0.002$, $\hat{k}_{\text{on}} = 0.001$, and $\rho_0 = 0.5$. We omitted data points with less than 7 TSSs per node. The $\hat{k}_{\text{off}} \gg 1$ case has the same trend, see Appendix E.

RNA expression level as well as the average for all nodes [Fig. 4(a)], it is hard to discern any significant difference as all curves nearly lie on top of each other [Fig. 4(b)]. This result shows that it is not only the highly transcribed gene starts that are relatively easy to find, it is any actively transcribed region. Repeating the same analysis for another cell type (K562), we found similar results, see Appendix G.

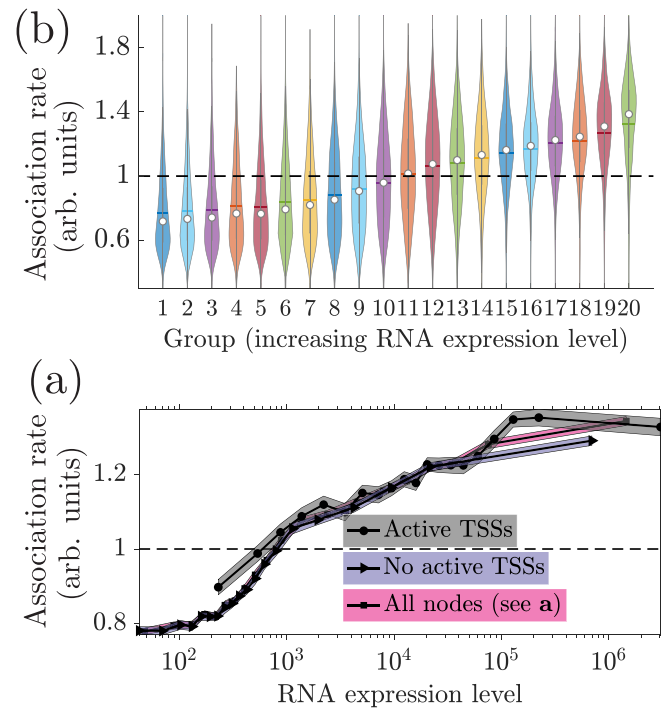


FIG. 4. Highly transcribed nodes are easy to find. (a) Genome-wide distribution of \hat{K}_a for all nodes divided into 20 groups based on their RNA expression levels. White circles: the median; horizontal bars: the mean; the dashed line; genome-wide average ($\langle \hat{K}_a \rangle = 1$). (b) Predicted \hat{K}_a as function of RNA expression level for nodes with at least one active TSS (grey) and no active TSSs (purple). Same grouping procedure as in (a). Nodes with active TSSs tend to be above the genome-wide average (18 points above $\langle \hat{K}_a \rangle$), while most nodes with no active TSSs are below (6 points above $\langle \hat{K}_a \rangle$). The shaded areas show the 95% confidence interval. Parameters: $\hat{k}_{\text{off}} = 0.002$, $\hat{k}_{\text{on}} = 0.001$, and $\rho_0 = 0.5$. We used data for cell line GM2878 but find similar results for K562, see Appendix G.

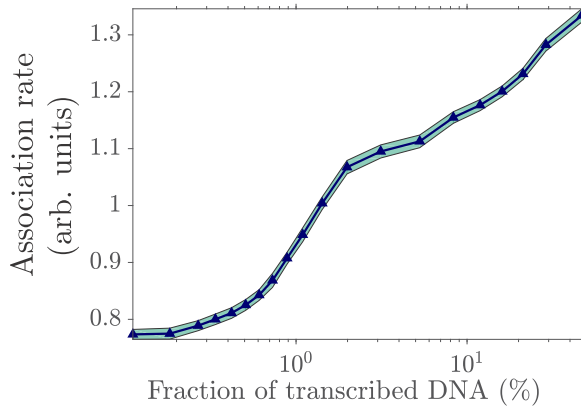


FIG. 5. Association rate increases with the fraction of transcribed DNA. This is similar to the RNA expression level which also increases with the association rate. Parameters: $\hat{k}_{\text{on}} = 0.001$ and $\hat{k}_{\text{off}} = 0.02$.

Several studies show that there is experimental noise in Hi-C data. To estimate by how much the association rates are affected by this, we removed 25% of the weakest contacts in each chromosome and recalculated \hat{K}_a . While the values differ, the qualitative results are the same, see Appendix J.

Furthermore, we also checked if the \hat{K}_a values are high because of high GC content as this is a known bias in Hi-C experiments [30]. To this end, we downloaded the reference genome sequence (hg19 [28]) and calculated the GC content (in percent) in every node. We found that it has a small positive correlation with \hat{K}_a , see Appendix L. This suggests that regions with more TSSs should have higher \hat{K}_a because gene promoters have a high GC content. However, the green area in Fig. 3 shows the opposite. It shows that \hat{K}_a declines with the number of TSSs in transcriptionally inactive regions even if the GC content grows and is comparable to the active regions (that, admittedly, have high \hat{K}_a , see grey area). We, therefore, conclude that any eventual GC bias in the Hi-C experiments is not enough to explain our results.

E. Accessible chromatin regions have high association rates

Accessible chromatin regions are, in general, associated with gene expression and regulation. We therefore asked if regions with accessible DNA have high association rates.

As a measure of DNA accessibility, we calculated the fraction of basepairs that are associated with at least one RNA transcript for every 40 kb region across the genome. Based on this fraction, we divided all nodes into 20 equally sized groups and calculated the average association rate within each one. We show the result in Fig. 5. The figure shows a clear correlation between the association rate and DNA accessibility. This correlation is marginally higher than for RNA expression: the Spearman's correlation coefficient is 0.56 compared to 0.54.

In addition to the fraction of transcribed DNA, we studied the association rates in open chromatin defined by DNase-seq data. This data set shows regions on the DNA that are cleaved by the DNase I enzyme. In Appendix K, we show that high association rates tend to have high DNase-seq signals.

Based on these findings, we conclude that accessible regions are easy to find.

IV. DISCUSSION AND SUMMARY

Protein-binding experiments show that association rates change if the DNA is short, long, straight, or coiled [4,9]. This is partly explained by the facilitated-diffusion model with simple assumptions for DNA-looping probabilities [13]. However, these assumptions are not consistent with chromatin's 3D structure in eukaryotes. To remedy this, we used Hi-C data as proxy for the 3D proximity between chromatin segments *in vivo*, and constructed a DNA-contact network. Then we formulated a diffusion-reaction equation that allowed us to calculate association rates analytically. Using human Hi-C data, we compared the predicted association rates with RNA expression data and positions of gene starts. We found that regions which are easy to find—measured by high association rates—are enriched with active genes and have high RNA expression levels.

Some of our results overlap with [31]. They found that short regions bound by Transcription factors, known as binding hotspots, coincide with chromatin loop anchors. This agrees with polymer simulations showing that particle densities are higher close to polymer loops. Just as in our work, this suggests that 3D conformation must be considered when analyzing protein-binding profiles.

We consider diffusion-limited search. However, some transcription factors, such as TetR [32], seem reaction-limited. To accommodate this case in our approach we may follow [33]: denoting the protein-DNA binding rate as k_{DNA} , and reinterpreting the on rate k_{on} as an effective on rate $k_{\text{on}}^{\text{eff}}$, we may write $1/k_{\text{on}}^{\text{eff}} = 1/k_{\text{on}} + 1/k_{\text{DNA}}$ where $K_a = k_{\text{on}}^{\text{eff}} + \gamma_a V_a$. In this work, we consider the limit $k_{\text{DNA}} \rightarrow \infty$.

We did not consider chromosome-chromosome contacts. To check this assumption, we calculated the ratio of internal versus external contacts from Hi-C data. Depending on which chromosomes we included, we found that 75%–90% of the contacts are internal. However, this number should be taken with caution because of the low signal-to-noise ratio [34].

Furthermore, we did not consider the low-copy number regime. This could be done with the chemical master equations [35]. However, as a rule of thumb, transcription factor concentrations are in the nanomolar range. This amounts to 10^0 – 10^3 proteins in bacteria and 10^3 – 10^6 in human cells. So in mammals, the many-searcher limit is reasonable.

Overall, this study provides a framework to predict protein-binding positions dictated by chromatin contact maps in the cell nucleus. As such, it opens new ways to interpret binding profiles of transcription factors that cannot be explained by the DNA sequence [1,36]. Mechanistic understanding of these profiles is essential to reach a molecular understanding of gene regulation.

ACKNOWLEDGMENTS

T.A. and L.L. acknowledge support from Swedish Research Council (Grants No. 2014-4305 and No. 2017-03848). P.S. acknowledges support from the Knut and Alice Wallenberg foundation (Grant No. 2014-0018, to EpiCoN, co-PI: P.S.). We thank Rajendra Kumar for his help with gcMapExplorer [37] to handle the Hi-C data.

APPENDIX A: PARTICLE FLUX THROUGH THE TARGET

The number of proteins that reached the target up to time t is $J_a(t)$. For nonzero k_{on} and k_{off} , it reads

$$J_a(t) = \sum_j \omega_{aj} \sum_i V_{ji} \left\{ \frac{k_{\text{on}}^i}{k_{\text{off}} - \lambda_i} \left[t - \frac{1 - e^{-t(k_{\text{off}} - \lambda_i)}}}{k_{\text{off}} - \lambda_i} \right] + \frac{1 - e^{-t(k_{\text{off}} - \lambda_i)}}{k_{\text{off}} - \lambda_i} \rho_0 \sum_{l \neq a} V_{il}^{-1} \right\} + k_{\text{on}} t, \quad (\text{A1})$$

where λ_j and V_{ij} are the eigenvalues and of eigenvectors ω_{ij} . Because $\lambda_1 = 0$ is the largest eigenvalue, we can approximate Eq. (A1) at times $t \gg k_{\text{off}}^{-1}$ with terms proportional to t

$$J_a(t) \simeq (k_{\text{on}} + T_a^{-1})t, \quad T_a^{-1} = \sum_j \omega_{aj} \bar{n}_j, \quad (\text{A2})$$

where the steady-state distribution is

$$\bar{n}_j = \sum_i \frac{V_{ji} k_{\text{on}}^i}{k_{\text{off}} - \lambda_i}. \quad (\text{A3})$$

The relation $J_a(t) \simeq (k_{\text{on}} + T_a^{-1})t$ coincides with the continuum approach in Ref. [14] for proteins that combines bulk excursions with 1D sliding (jumping to nearest neighbours in our model) and Lévy relocations with jump lengths x distributed like $\simeq |x|^{-1-\alpha}$ ($0 < \alpha < 2$). Since $\omega_{ij} \simeq |i - j|^{-1-\alpha}$ with $0 < \alpha < 1$ —on average—we see that our model is a network analog of Ref. [14].

APPENDIX B: PARTICLE FLUX THROUGH THE TARGET WITHOUT BULK EXCHANGE

Here we investigate the case when proteins do not unbind from the DNA. As $k_{\text{on}}, k_{\text{off}} \rightarrow 0$, Eq. (A1) becomes

$$J_a(t) = \sum_{k=1}^N \omega_{ak} \sum_{i=2}^N \frac{V_{ki}}{|\lambda_i|} (1 - e^{-t|\lambda_i|}) \sum_{j \neq a} \rho_0 V_{ij}^{-1} = N_p - \sum_{k=1}^N \omega_{ak} \sum_{i=2}^N \frac{V_{ki}}{|\lambda_i|} e^{-t|\lambda_i|} \sum_{j \neq a} \rho_0 V_{ij}^{-1}, \quad (\text{B1})$$

with $N_p = \rho_0(N - 1)$. For large times, we know that $J_a(t \rightarrow \infty) = N_p$ since by then all proteins have arrived to the target. This leads to the simplification in the second row. For small times $t \ll |\lambda_N|^{-1} - \lambda_N$ is the largest eigenvalue (in magnitude)—where $J_a(t) \ll N_p$, we find the same behavior as before, $J_a(t) \propto t$. This is seen by expanding Eq. (B1) around $t = 0$.

APPENDIX C: DERIVATION OF EQS. (6)–(8)
1. Fast target finding (small k_{off})

When the unbinding rate k_{off} is small compared to the association rate K_a , the number of proteins per node is close to its initial value ρ_0 by the time of the first arrival to the target, and we have the approximation

$$K_a = k_{\text{on}} + \rho_0 W_a, \quad (\text{C1})$$

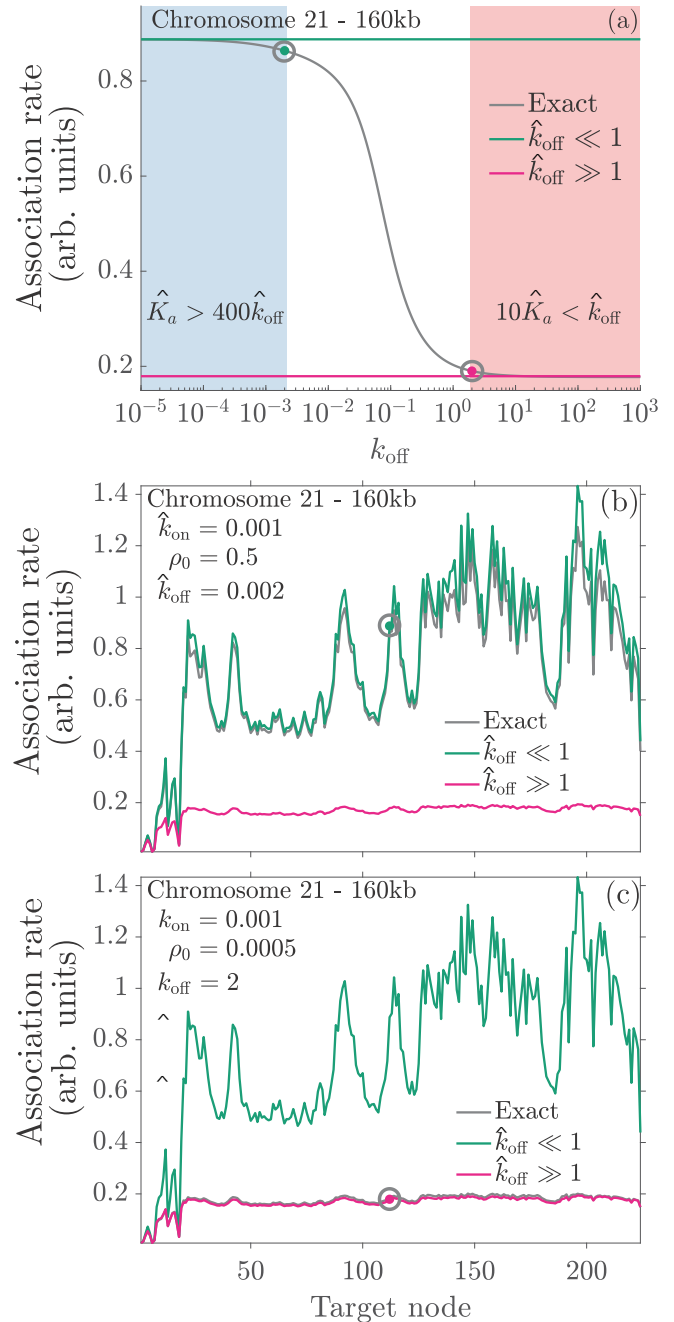


FIG. 6. Comparison of the analytical approximations to the exact, time-integrated association rate Eq. (D1), using Hi-C data from chromosome 21 at 160 kb resolution. (a) The blue area represents the fast target-search regime where $\hat{k}_{\text{off}} \ll 1$. The red area shows the opposite regime, $\hat{k}_{\text{off}} \gg 1$, where the system is close to its steady-state and target finding is slow. We put the target in the middle of the system $a = N/2$. In the two lower panels, we show the association rates' profile to all possible targets on chromosome 21 when $\hat{k}_{\text{off}} \ll 1$ (b) and $\hat{k}_{\text{off}} \gg 1$ (c). Note the green and red circled dots in (b) and (c), respectively. These correspond to the encircled parameter values in panel (a).

where $W_a = \sum_{i \neq a} \omega_{ai}$. We may find this approximation by expanding Eq. (A1) around $t = 0$ and using the inverse transformation $\sum_j V_{ij} q_j(0) = n_i(0) = \rho_0(1 - \delta_{ia})$.

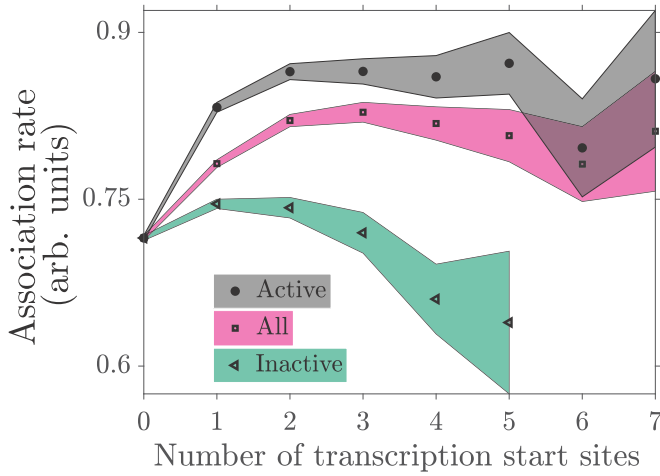


FIG. 7. Same as Fig. 3 in the manuscript but for large \hat{k}_{off} instead of small. We recover the same trend: easy-to-find regions tend to have many gene starts.

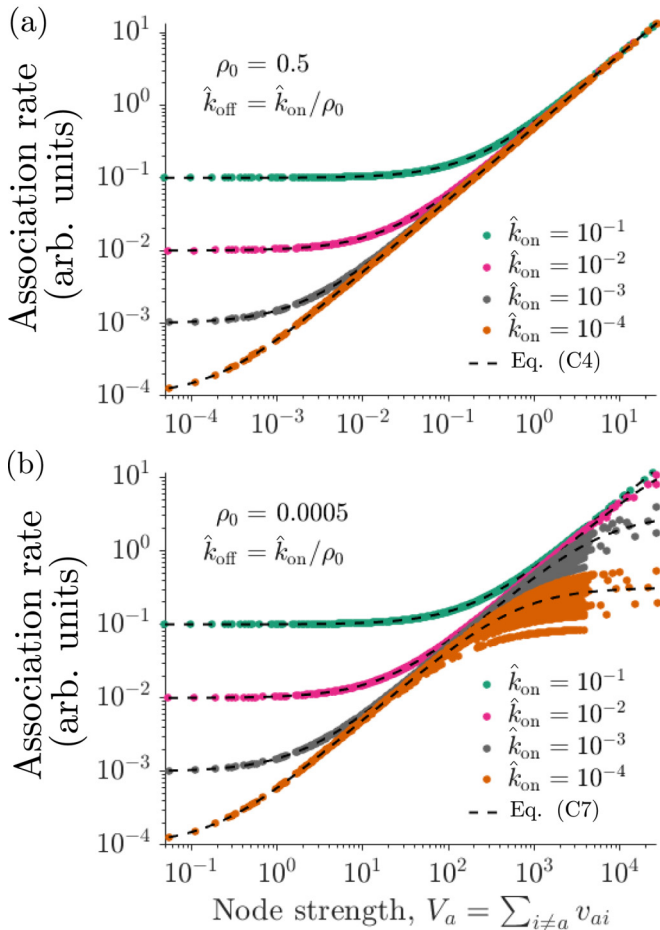


FIG. 8. The association rate vs the nodes' strength V_a (sum of all link weights). (a) All data points follow the universal law $\hat{K}_a = \hat{k}_{\text{on}} + \gamma_{a1} V_a$ [Eq. (C4)]. The density $\rho_0 = 0.5$ is kept fixed in all four cases as we increase k_{on} . (b) Association rate during steady state ($\hat{k}_{\text{off}} \gg 1$) calculated from Eq. (C8). The behavior is not universal as it depends on the number of nodes N for each chromosome contact network. The dashed line represent the analytical formulas. As N , we used the mean chromosome size.

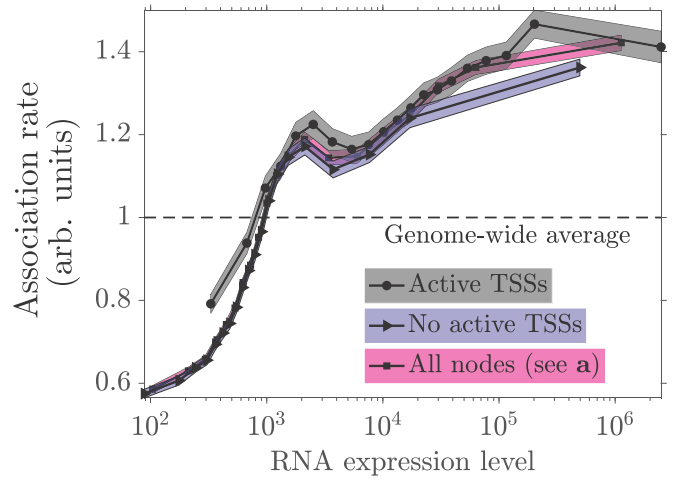
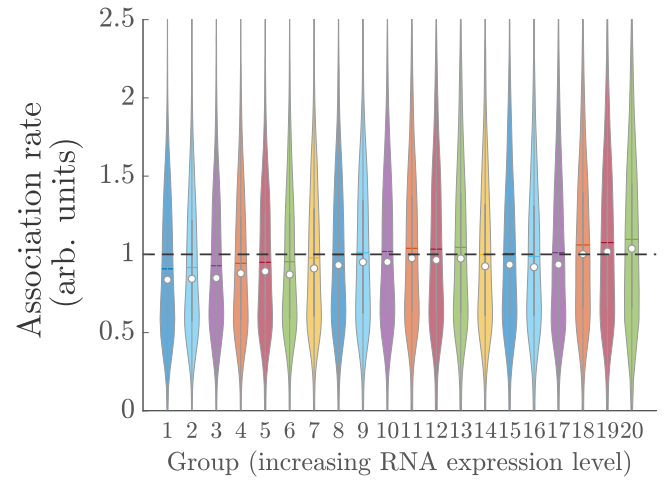


FIG. 9. Same as Fig. 4 but for cell type K562 instead of GM12878. The figure shows that highly transcribed nodes—in terms of high levels of RNA expression—are easy to find (low association rates) even if there are active genes there or not. See the caption of Fig. 4 in the manuscript for a more details on what the curves represent.

Next, we rescale Eq. (C1) so that the genome-wide averaged K_a is unity. Using that $W_a = f_{\text{coll}} V_a$, we may express f_{coll} as

$$f_{\text{coll}} = \frac{\langle K_a \rangle - k_{\text{on}}}{\rho_0 \langle V_a \rangle} \quad (\text{C2})$$

To obtain Eqs. (6) and (7) in the main text, we replace f_{coll} by Eq. (C2) and use these definitions

$$\hat{K}_a = \frac{K_a}{\langle K_a \rangle}, \quad \hat{k}_{\text{on}} = \frac{k_{\text{on}}}{\langle K_a \rangle}, \quad \gamma_{a1} = \frac{1 - \hat{k}_{\text{on}}}{\langle V_a \rangle}. \quad (\text{C3})$$

This gives

$$\hat{K}_a = \hat{k}_{\text{on}} + \gamma_{a1} V_a. \quad (\text{C4})$$

After this rescaling, the fast target-finding limit becomes $\hat{k}_{\text{off}} \ll 1$.

2. Target finding in steady state (large k_{off})

When the unbinding rate \hat{k}_{off} is large compared to the association rate \hat{K}_a , few proteins will find the target before

leaving on a bulk excursion. In this limit, the system reaches its steady state before the first arrival to the target. This leads to the approximation

$$\hat{K}_a = \hat{k}_{\text{on}} + \bar{\rho}_a W_a. \quad (\text{C5})$$

where $\bar{\rho}_a$ the number of proteins per node in steady state.

To arrive at this equation we identify in Eq. (A2) that $J(t) \simeq \hat{K}_a t$. Then we replace the \bar{n}_j by the approximate density $\bar{\rho}_a$ that we find by the following argument. In steady state, proteins bind to the DNA with rate \hat{k}_{on} . Except for the absorbing target, there are $N - 1$ nodes available to bind. Similarly, there are $\bar{\rho}_a(N - 1)$ number of proteins that unbind from the DNA with rate \hat{k}_{off} . Last, proteins are absorbed at the target with rate $T_a^{-1} = \bar{\rho}_a W_a$. These three terms sum to zero, and therefore

$$\bar{\rho}_a = \frac{\hat{k}_{\text{on}}}{\hat{k}_{\text{off}} + W_a/(N - 1)}. \quad (\text{C6})$$

We may rescale Eq. (C5) as in the previous section using $W_a = f_{\text{coll}} V_a$, f_{coll} from Eq. (C2), and the rescaled variables in Eq. (C3). Combining these equations give

$$\hat{K}_a = \hat{k}_{\text{on}} + \frac{\hat{k}_{\text{on}} \frac{\langle K_a \rangle - \hat{k}_{\text{on}}}{\rho_0 \langle V_a \rangle} V_a}{\hat{k}_{\text{off}} + \frac{V_a}{N-1} \frac{\langle K_a \rangle - \hat{k}_{\text{on}}}{\rho_0 \langle V_a \rangle}} = \hat{k}_{\text{on}} + \gamma_{a_2} V_a, \quad (\text{C7})$$

where

$$\gamma_{a_2} = \frac{\hat{k}_{\text{on}} \gamma_{a_1}}{\hat{k}_{\text{off}} + \frac{V_a}{N-1} \gamma_{a_1}}. \quad (\text{C8})$$

APPENDIX D: VALIDATION OF APPROXIMATIONS

To better understand the validity of Eqs. (C4) and (C7), we compare them to the exact association rate

$$K_a^{\text{exact}} = \left(\int_0^\infty \exp(-J_a(t)) dt \right)^{-1}. \quad (\text{D1})$$

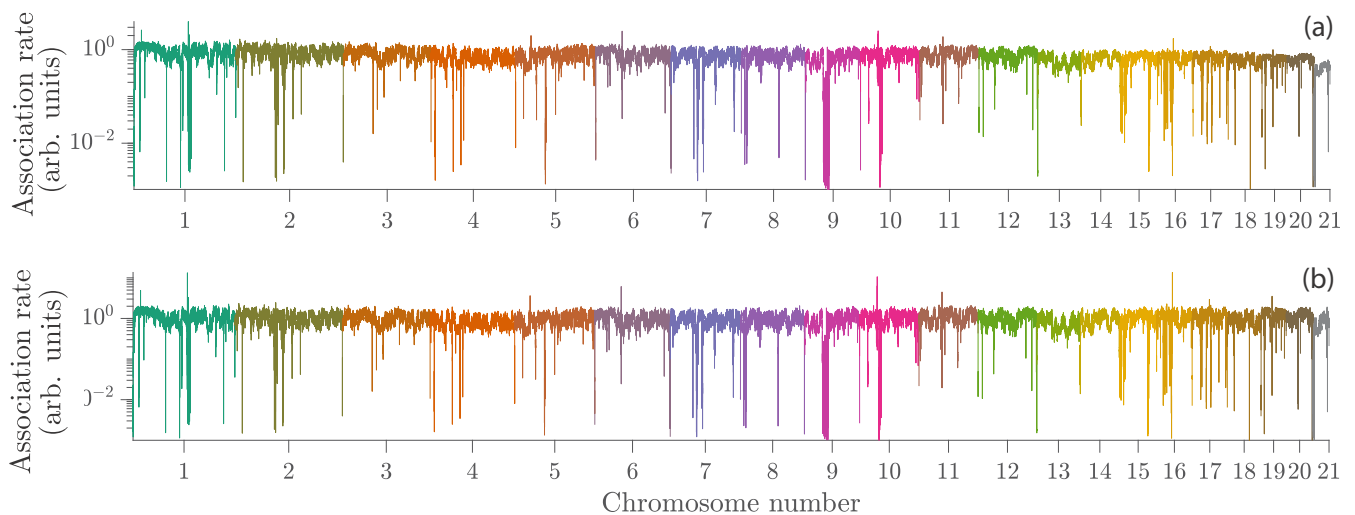


FIG. 10. Genome-wide association rates \hat{K}_a at 40 kb resolution. (a) K_a for slow unbinding rates predicted by Eq. (C7). This complements Fig. 2 showing K_a for fast unbinding rates. As in Fig. 2, most values deviate only from the average by a few percent (0.7531 ± 0.0018 , 95% confidence interval). We used these parameters: $\hat{k}_{\text{off}} = 2$, $\hat{k}_{\text{on}} = 0.001$ and $\rho_0 = 0.0005$. (b) K_a predicted by Eq. (C7) (slow unbinding rate) after we removed 25% of the weakest contacts. The genome-wide average is $\langle \hat{K}_a \rangle = 1 \pm 0.0028$ (95% confidence interval). We used same parameters as in Fig. 2: $\hat{k}_{\text{off}} = 2$, $\hat{k}_{\text{on}} = 0.001$ and $\rho_0 = 0.0005$.

Figure 6(a) shows how the association rate changes for a specific target node—we choose $a = N/2$ in human chromosome 21—as we change \hat{k}_{off} while keeping the on-rate fixed, $\hat{k}_{\text{on}} = 0.001$, and adjusting the density $\rho_0 = \hat{k}_{\text{on}}/\hat{k}_{\text{off}}$. The solid grey line shows \hat{K}_a^{exact} and the horizontal lines represent the approximations for small and large \hat{k}_{off} —Eqs. (C4) and (C7).

The blue area in Fig. 6(a) shows the large \hat{K}_a regime ($\hat{K}_a > 400\hat{k}_{\text{off}}$). Here, Eq. (C1) deviates only a few percent from \hat{K}_a^{exact} : the deviation is 2.7% ($\approx 1 - \hat{K}_a/\hat{K}_a^{\text{exact}}$) at the encircled green dot. To get this number, we used $\hat{k}_{\text{off}} = 0.002$, $\hat{k}_{\text{on}} = 0.001$ and $\rho_0 = 0.5$ —the same values that we used to create all plots in the main text.

The pink area represents the opposite limit: small \hat{K}_a ($\hat{K}_a < \hat{k}_{\text{off}}/10$). In this region, the approximation in Eq. (C7) is a good match to \hat{K}_a^{exact} . At the red dot ($\hat{k}_{\text{off}} = 2$), the relative error is 5.7%.

In the intermediate region (white area), we cannot use the simple expressions because the flux $J(t)$ has a complicated time-dependence. To get the association rate in this regime, we have to evaluate Eq. (D1) directly.

In Figs. 6(b) and 6(c), we calculate the association rate for all nodes in chromosome 21 using Eqs. (D1), (C4), and (C7) with fixed parameters (shown in the figures). The figure shows the limiting \hat{k}_{off} cases. In Fig. 6(b), the unbinding rate is small ($\hat{k}_{\text{off}} = 0.002$), and we see that the approximation (C4) match well with \hat{K}_a^{exact} whereas Eq. (C7) does not. Equation (C7) matches better in Fig. 6(c) where the unbinding rate is large ($\hat{k}_{\text{off}} = 2$).

APPENDIX E: GENOME-WIDE ASSOCIATION RATES WHEN k_{off} IS LARGE

In Fig. 2, we show the association rates when $\hat{k}_{\text{off}} \ll 1$. Here, we investigate the opposite limit by evaluating Eq. (C7) and plotting the \hat{K}_a values as a genome-wide profile

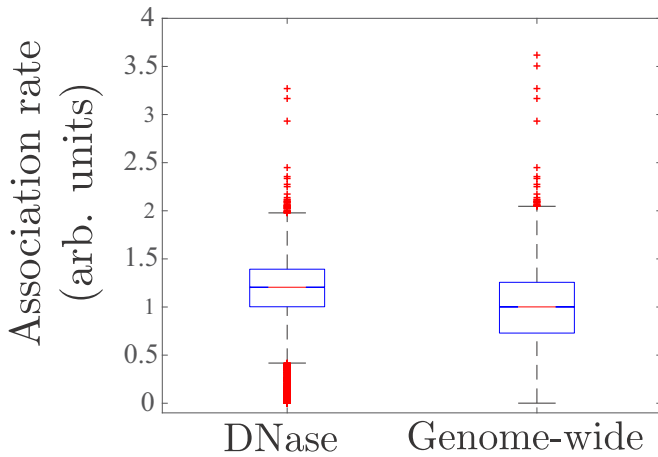


FIG. 11. (Left) A boxplot of the association rates for 40kb bins with significant DNase-seq signal (Right) Association rates for all 40-kb bins across the genome. The y axis is cut at $\hat{K}_a = 4$, which removes 13 outliers out of 66 452 data points.

[Fig. 10(a)]. We find that the curves for $\hat{k}_{off} \gg 1$ and $\hat{k}_{off} \ll 1$ (Fig. 2) are almost identical, except for an offset on the y axis.

In the large \hat{k}_{off} limit, we also show how the number of transcription start sites change with \hat{K}_a (Fig. 7). Compared to Fig. 3 in the main text for small \hat{k}_{off} , the trend is the same as: it is easy to find regions with many gene starts.

APPENDIX F: GENOME-WIDE ASSOCIATION RATE AS A FUNCTION OF NODE STRENGTH

In Fig. 8(a), we show how \hat{K}_a —calculated from Eq. (C1) ($\hat{k}_{off} \ll 1$)—varies with node strength V_a for four different values of \hat{k}_{on} with fixed $\rho_0 = 0.5$; The symbols represent values for individual nodes across the human genome. For comparison, we plot the analytical prediction Eq. (C4). We find that the search times are dominated by \hat{k}_{on} for weakly connected nodes. For strongly connected nodes, we find the universal behavior $\hat{K}_a \propto V_a$.

In Fig. 8(b), we show \hat{K}_a for all nodes in the other limit $\hat{k}_{off} \gg 1$. Here \hat{K}_a depends on the number of nodes N —via $\bar{\rho}_a$ in Eq. (C6)—and therefore we do not expect a universal large- V_a behavior.

APPENDIX G: ANALYSIS FOR CELL LINE K526

In the main text, we used data for cell line GM12878. Here we explore if some of our results are cell-type specific. To this end, we downloaded Hi-C and RNA expression data for K562 [28] and reconstructed Fig. 4, see Fig. 9. Although there are qualitative differences, the trend is the same for both cell types: association rates grow with increasing RNA expression levels.

APPENDIX H: DETERMINING TRANSCRIPTION START SITES

From ENCODE, we downloaded the gene annotations (see Ref. [38]). In the file, we searched for they keyword “gene” and extracted the coordinate for the gene start. This position is

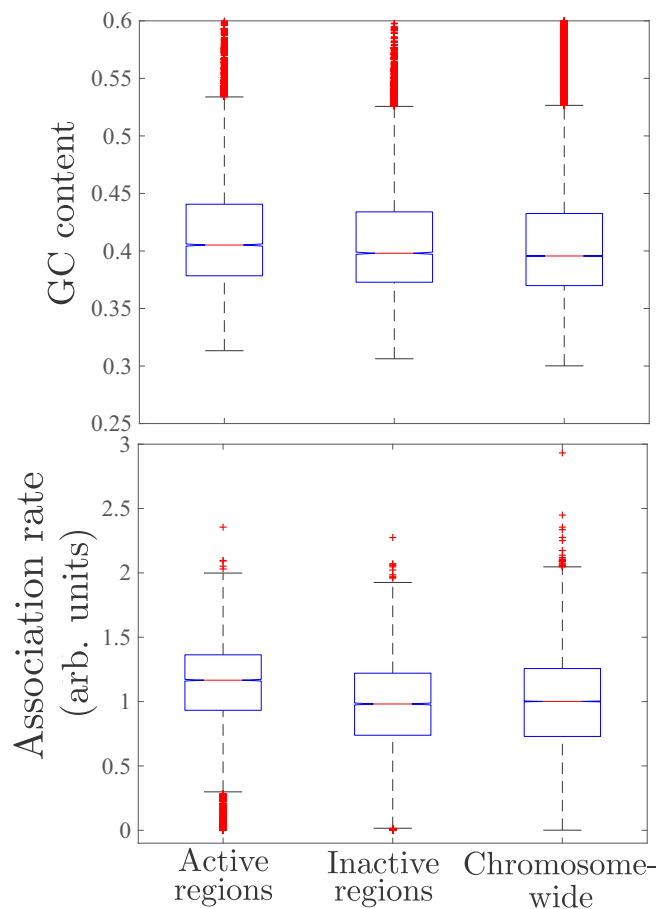


FIG. 12. (Top) Boxplot of the gc content for 40kb bins with a single transcription start site (TSS). The bins are separated in two groups—active and inactive—based on the transcriptional activity surrounding each TSS. As a reference, we show the gc content for all Hi-C bins (“genome-wide”) (Bottom) The range of association rates for the same three groups as to the left. The y axis is cut at $\hat{K}_a = 3$ which removes about 10 outliers.

the transcription start site that is farthest away from the gene’s end. In the data, there are alternative TSSs defined by lines denoted by “transcript.” We omitted those in our analysis.

APPENDIX I: DETERMINING ACTIVE VERSUS INACTIVE REGIONS

To distinguish between active and inactive regions, we use RNA expression data (ENCODE, v.19). From the RNA read counts, we calculated the average number of RNA reads per base pair, \bar{n}_{RNA} , ± 1 kb around each TSS. We defined a TSS as transcriptionally active when $\bar{n}_{RNA} \geq 1$. Given this threshold we found 32712 active and 20795 inactive TSSs.

APPENDIX J: PRUNING THE HI-C DATA

Several studies show that there is experimental noise in the Hi-C data. To see by how much this affected the values of our association rates, we removed weak contacts. Following [26], we removed 25% of the weakest contacts and then recalculated the association rates. In Fig. 10, we show the

corresponding Fig. 2 in the main text. To quantify the difference before and after pruning, we correlated the association rates in both cases with each other. We found that the correlation is as high as 0.99986 (Spearman correlation coefficient). In other words, removing a quarter of the weak contacts did not change our results.

APPENDIX K: DNA ACCESSIBILITY

We used two metrics to measure accessible DNA: the fraction of transcribed DNA (main text) and DNase-seq data. Here we describe the DNase-seq analysis.

We downloaded DNase-seq peak-data from ENCODE [28]. This data is pre-processed to keep significant DNase-seq signals. Then we selected all 40 kb regions (Hi-C bins) that have significant DNase peaks and at least one TSS. In Fig. 11, we show a boxplot of how the association rates varies among these regions compared to the genome. They are clearly higher than the genome-wide average indicating that open chromatin is easy to find.

APPENDIX L: GC BIAS

To make a fair comparison to see if our results are caused by a GC bias, we calculated the GC content and association rates in regions with the same gene density. To this end, we took all 40-kb Hi-C bins with one TSS because these are most abundant of the TSS-containing boxes. Then we downloaded the reference genome sequence (hg19) and calculated the GC

content in all Hi-C bins across the genome that had a single TSS. Then we separated these bins into two groups depending on if the TSSs reside in active or an inactive regions (based on the fraction of transcribed DNA as in Fig. 3). Then we made a box plot for the GC content for each group [Fig. 12 (top)]. Albeit statistically significant—the Wilcox sum-rank test rejects the hypothesis that the two groups come from the same distribution with $p = 10^{-12}$ —the plot shows that there is little difference in GC content between active and inactive regions. In fact, these groups do not differ much from the genome-wide GC distribution show in the rightmost box.

In the bottom panel, we show a box plot for the association rates for the same two groups. The rates deviate more than the GC content (the Wilcox sum-rank test gives $p = 10^{-250}$ for the difference between active and inactive regions). We also see that the box for the association rates for the inactive regions is slightly below the box for genome-wide rates even though the GC content is slightly higher.

We see a similar observation in Fig. 3 in the manuscript. This figure shows how \hat{K}_a changes with the number of TSSs. For TSSs in active regions, \hat{K}_a declines. For TSSs in inactive regions, it grows. However, the GC content these groups is not that different (indicated by Fig. 12). And furthermore, as the number of TSS increases, so will the GC content. But for the TSS in inactive regions, \hat{K}_a goes down. It does not go up as it would if the GC content drives \hat{K}_a .

Altogether, these findings speak against a strong GC bias even though it is hard to rule out with certainty.

-
- [1] P. J. Farnham, Insights from genomic profiling of transcription factors, *Nat. Rev. Genet.* **10**, 605 (2009).
- [2] V. Globyte, S. H. Lee, T. Bae, J.-S. Kim, and C. Joo, Crispr/cas9 searches for a protospacer adjacent motif by lateral diffusion, *EMBO J.* **38**, e99466 (2019).
- [3] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf, The lac repressor displays facilitated diffusion in living cells, *Science* **336**, 1595 (2012).
- [4] A. D. Riggs, S. Bourgeois, and M. Cohn, The lac repressor-operator interaction: Iii. kinetic studies, *J. Mol. Biol.* **53**, 401 (1970).
- [5] P. H. von Hippel and O. G. Berg, Facilitated target location in biological systems, *J. Biol. Chem.* **264**, 675 (1989).
- [6] A. B. Kolomeisky, Physics of protein–dna interactions: mechanisms of facilitated target search, *Phys. Chem. Chem. Phys.* **13**, 2088 (2011).
- [7] L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith, and A. Kosmrlj, How a protein searches for its site on dna: the mechanism of facilitated diffusion, *J. Phys. A: Math. Theor.* **42**, 434013 (2009).
- [8] J. Elf, G.-W. Li, and X S. Xie, Probing transcription factor dynamics at the single-molecule level in a living cell, *Science* **316**, 1191 (2007).
- [9] B. van den Broek, M. A. Lomholt, S.-M. J. Kalisch, R. Metzler, and G. J. L. Wuite, How dna coiling enhances target localization by proteins, *Proc. Natl. Acad. Sci. USA* **105**, 15738 (2008).
- [10] A. Amitai, Chromatin configuration affects the dynamics and distribution of a transiently interacting protein, *Biophys. J.* **114**, 766 (2018).
- [11] M. Bauer and R. Metzler, In vivo facilitated diffusion model, *PLoS ONE* **8**, e53956 (2013).
- [12] T. Hu, A. Yu. Grosberg, and B. I. Shklovskii, How proteins search for their specific sites on dna: the role of dna conformation, *Biophys. J.* **90**, 2731 (2006).
- [13] M. A. Lomholt, B. van den Broek, S.-M. J. Kalisch, G. J. L. Wuite, and R. Metzler, Facilitated diffusion with dna coiling, *Proc. Natl. Acad. Sci. USA* **106**, 8204 (2009).
- [14] M. A. Lomholt, T. Ambjörnsson, and R. Metzler, Optimal Target Search on a Fast-Folding Polymer Chain with Volume Exchange, *Phys. Rev. Lett.* **95**, 260603 (2005).
- [15] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragozcy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science* **326**, 289 (2009).
- [16] S. Kong and Y. Zhang, Deciphering hi-c: from 3d genome to function, *Cell biology and toxicology* **35**, 15 (2019).
- [17] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander *et al.*, A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping, *Cell* **159**, 1665 (2014).
- [18] J. Krefting, M. A. Andrade-Navarro, and J. Ibn-Salem, Evolutionary stability of topologically associating domains is associated with conserved gene regulation, *BMC Biol.* **16**, 87 (2018).
- [19] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, Topological domains in

- mammalian genomes identified by analysis of chromatin interactions, *Nature (London)* **485**, 376 (2012).
- [20] D. Jost, P. Carrivain, G. Cavalli, and C. Vaillant, Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains, *Nucleic Acids Res.* **42**, 9553 (2014).
- [21] S. H. Lee, Y. Kim, S. Lee, X. Durang, P. Stenberg, J.-H. Jeon, and L. Lizana, Mapping the spectrum of 3d communities in human chromosome conformation capture data, *Sci. Rep.* **9**, 6859 (2019).
- [22] M. Di Pierro, R. R. Cheng, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture, *Proc. Natl. Acad. Sci. USA* **114**, 12126 (2017).
- [23] F. Serra, M. Di Stefano, Y. G. Spill, Y. Cuartero, M. Goodstadt, D. Baù, and M. A. Marti-Renom, Restraint-based three-dimensional modeling of genomes and genomic domains, *FEBS Lett.* **589**, 2987 (2015).
- [24] J. Smrek and A. Y. Grosberg, Facilitated diffusion of proteins through crumpled fractal dna globules, *Phys. Rev. E* **92**, 012702 (2015).
- [25] R. E. Boulos, N. Tremblay, A. Arneodo, P. Borgnat, and B. Audit, Multi-scale structural community organisation of the human genome, *BMC Bioinf.* **18**, 209 (2017).
- [26] S. Sarnataro, A. M. Chiariello, A. Esposito, A. Prisco, and M. Nicodemi, Structure of the human chromosome interaction network, *PLoS ONE* **12**, e0188201 (2017).
- [27] I. M. Sokolov, R. Metzler, K. Pant, and M. C. Williams, First passage time of N excluded-volume particles on a line, *Phys. Rev. E* **72**, 041102 (2005).
- [28] Reference genes (hg19) were downloaded from ensembl database (www.ensembl.org), gene expression data (rna-seq) and hi-c data for gm12878 cells from gene expression omnibus (gsm840138 and gse63525, respectively). the used dnase peak data from encode: GM12878 DNaseI HS Peaks from ENCODE/Duke.
- [29] C. Spearman, The proof and measurement of association between two things, *Am. J. Psychology* **15**, 72 (1904).
- [30] E. Yaffe and A. Tanay, Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture, *Nat. Genet.* **43**, 1059 (2011).
- [31] R. Cortini and G. J. Filion, Theoretical principles of transcription factor traffic on folded chromatin, *Nat. Commun.* **9**, 1740 (2018).
- [32] D. Normanno, L. Boudarene, C. Dugast-Darzacq, J. Chen, C. Richter, F. Proux, O. Bénichou, R. Voituriez, X. Darzacq, and M. Dahan, Probing the target search of dna-binding proteins in mammalian cells using tetr as model searcher, *Nat. Commun.* **6**, 7357 (2015).
- [33] O. G. Berg, A. Mahmutovic, E. Marklund, and J. Elf, The helical structure of dna facilitates binding, *J. Phys. A: Math. Theor.* **49**, 364002 (2016).
- [34] S. Kaufmann, C. Fuchs, M. Gonik, E. E. Khrameeva, A. A. Mironov, and D. Frishman, Inter-chromosomal contact networks provide insights into mammalian chromatin organization, *PLoS ONE* **10**, e0126125 (2015).
- [35] N. G. van Kampen, A power series expansion of the master equation, *Can. J. Phys.* **39**, 551 (1961).
- [36] D. Schmidt, M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown, A. Marshall, C. Kutter, S. Watt, C. P. Martinez-Jimenez, S. Mackay *et al.*, Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding, *Science* **328**, 1036 (2010).
- [37] R. Kumar, H. Sobhy, P. Stenberg, and L. Lizana, Genome contact map explorer: a platform for the comparison, interactive visualization and analysis of genome contact maps, *Nucleic Acids Res.* **45**, e152 (2017).
- [38] <https://www.encodeproject.org/files/gencode.v19.annotation/>.