bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Structure of Topologically Associating Domains and gene regulation embedded in the 5C data revealed by a polymer model and stochastic simulations

O. Shukron [1], D. Holcman [*]

[*1] Ecole Normale Supérieure, 46 rue d'Ulm 75005 Paris, France and Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom. Corresponding author email:david.holcman@ens.fr

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Chromatin organization is probed by chromosomal capture data, from which the encounter probability (EP) between genomic sites is represented in a large matrix. However, this matrix is obtained by averaging the EP over cell population, where diagonal blocks called TADs, contains hidden information about sub-chromatin organization. Our aim here is to elucidate the relationship between TADs structure and gene regulation. For this end, we reconstruct the chromatin dynamics from the EP matrix using polymer model and explore the transient properties, constrained by the statistics of the data. To construct the polymer, we use the EP decay in two steps: first, to account for TADs, we introduce random connectors inside a restricted region defining the TADs. Second, we account for long-range frequent specific genomic interactions in the polymer architecture. Finally, stochastic simulations show that only a small number of randomly placed connectors are required to reproduce the EP of TADs, and allow us to compute the mean first time and the conditional encounter probability of three key genomic sites to meet. These encounter times reveal how chromatin can self-regulate. The present polymer construction is generic and can be used to study steady-state and transient properties of chromatin constrained on 5C data.

1 Introduction

The chromatin molecule is organized in heterogenous sub-regions of various sizes, as recently revealed by Chromosome Capture (5C) data [1, 2]. This multi-scale organization is due to short and long-range genomic interactions between DNA segments, collected over large cell population. It was recently shown [3, 4] that the mammalian chromatin, at a resolution of 3kB, contains an organization at 1Mbps scale, where some structures are enriched in intra-connectivity, reflecting an increased encounter probability between these segments. These encounter properties are represented in a two-dimensional encounter frequency (EF) matrix, containing diagonal blocks called Topologically Associating Domains (TADs) [3, 5]. These blocks associated with gene regulation [3], DNA replication units [6], or DNA entanglement and cross-linking by gluing molecules such as cohesin, CTCF, and condensin [3].

Looping between chromatin sites are precisely the events sampled by chromosome capture data (3C, 4C, 5C, HiC) [3, 7], and single cell HiC confirms that configurations can vary between cell types and phases [8]. In that context, TADs represent averaged chromatin conformations, characterized by a higher mean number binding compared to non-TAD regions.

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To interpret the encounter probability (EP) in the 5C data, polymer models are utilized, such as the Rouse polymer [23] that predicted a -3/2 decay exponent for the EP with genomic distance. Polymer models have been used to described chromosomal territories [9], further suggesting that inactive genes are located inside these territories. By adding local binding sites on polymer models, TAD structures can emerge [10, 11, 12, 13], where the decay exponent of the EP can vary at different scales. 5C EPs at the X-inactivation centre was recently used to calibrate interaction potentials between beads in a freely jointed chain, allowing to assess internal 3D mean distances inside TADs [5]. By incorporating reversible binding of diffusing molecules on a self-avoiding polymer [14], some chromatin conformation have been identified. Inter chromosome distances at large scale resolution of 1Mb pairs were also found using minimization algorithms [15]. In parallel, single particle trajectories of tagged DNA locus [16, 17, 18, 19, 20, 21, 22] reveal that chromatin is constantly remodeled, characterized by mean square displacement function and its anomalous behavior.

Our aim here is to elucidate the relationship between TADs structure and gene regulation. For this end, we present a general procedure to construct a coarse-grained polymer model that accounts for the statistical properties of the chromosome capture data, such as the decay of the EP and TADs. The reconstructed polymer is then used to study transient properties such the interaction between two given sites. We start with a Rouse polymer model, which consists of beads connected by harmonic spring, and use the EP of the 5C data of mammalian X chromosomes to constrain further monomer interactions. The construction procedure is divided into two steps: first, to account for heterogeneity in the 5C data, we add connectors (cross-links) between genomic sites chosen at random, and show we can reproduce TAD block. We show that the number of random connectors that is needed to be added is uniquely determined from data. However, this step is insufficient to recover the EP decay probability. Thus, in the second step, we account for consistent long-range interaction, represented by the local maximum of the EP matrix. We validate this reconstruction by showing that the EP-matrix, constructed from simulations of the polymer following the two-step procedure, has the same decay exponent (for each monomer) as that of the empirical data. Once the polymer model is constructed, we estimate the conditional encounter probability and the associated mean first passage time between three specific genomic sites. The present polymer construction is generic and can be used to study transient properties inside or between any TADs or any 5C matrix.

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2 Materials and Methods

2.1 Construction a generalized Rouse polymer from encounter frequency map

The Rouse model describes a polymer as a collection of beads $R_n (n = 1...N)$ connected by harmonic springs and driven by Brownian motion. The energy of the polymer is given by [23]

$$\phi_{Rouse}(R_1, .., R_N) = \frac{1}{2}\sum_{j=1}^{N-1} \kappa(R_j - R_{j+1})^2, \tag{1}$$

where here we set $\kappa = \frac{3k_B T}{\gamma b^2}$ and b is the standard deviation of the distance between adjacent monomers, γ is the friction coefficient, k_B the Boltzmann coefficient and T the temperature.

To account for a sub-chromatin region, characterized by a higher encounter probability than the rest, we added connections between monomer-pairs chosen randomly (with uniform distribution) inside this subregion $\mathcal{C}_{\mathcal{N}}$ such that an additional potential

$$\phi_{Rand}(R_1, .., R_N) = \frac{1}{2}\sum_{j,k\in\mathcal{C}_{\mathcal{N}}} \kappa(R_j - R_k)^2, \tag{2}$$

is added to ϕ_{Rouse}, where $\mathcal{C}_{\mathcal{N}}$ is the ensemble of indices defining the regions. In addition, to account for consistent long-range interactions, reflected by peaks in EP matrix (Fig. 1C), we connected monomer-pairs by spring as follows: first, the positions of peaks form a subset S_{Max} of the ensemble of the off-diagonal local maxima in the EP-matrix, such that their EP is higher than a threshold T_{th}. In practice, we chose the threshold T_{th} to represent the encounter probability for the nearest neighbor monomers in $M_{i,j}$, that is

$$T_{th} = \frac{\sum_i M_{ii-1} + M_{ii+1}}{\sum_{i,j} M_{ij}}. \tag{3}$$

The spring constants $\kappa_{m,n}$ between monomer m and n in S_{Max} are determined from the empirical encounter probability $P_{m,n}$ at each peak position. For a Rouse chain the joint probability density function of beads R_m and R_n is given by [23, p.15]

$$\Phi(R_m, R_n, \Delta_{m,n}) = \left(\frac{3}{2\pi b^2 \Delta_{m,n}}\right)^{3/2} \exp\left(-\frac{3(R_m - R_n)^2)}{2b^2\Delta_{m,n}}\right) \tag{4}$$

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $\Delta_{m,n} = |m-n|$. For the nearest neighbors $\Delta_{m,n} = 1$ and the encounter probability occurs at small distances so that the exponential is almost 1.

$$P_{m,n} = \left(\frac{3}{2\pi b^2}\right)^{3/2} \approx \left(\frac{\kappa_{m,m+1}}{2\pi}\right)^{3/2}, \tag{5}$$

We approximate the chromatin as a polymer chain with a uniform variance b^2 between adjacent monomers, thus, the constant $\bar{b}$ is estimated as the mean EP over all neighboring monomers:

$$\left(\frac{3}{2\pi\bar{b}^2}\right)^{3/2} \approx \sum_i (P_{ii-1} + P_{ii+1}) = \frac{\sum_i M_{ii-1} + M_{ii+1}}{\sum_{i,j} M_{ij}} = T_{th} \tag{6}$$

To account for the long-range interactions, we applied formula 5 to estimate the effective spring constant from the empirical EP $\tilde{P}$,

$$\kappa_{m,n} = 2\pi\tilde{P}^{2/3}_{m,n}. \tag{7}$$

The energy related to peak interactions is described by

$$\phi_{Peaks}(R_1, ..., R_N) = \frac{1}{2}\sum_{n,m\in\mathcal{C}_{\mathcal{P}}} \kappa_{n,m}(R_n - R_m)^2 \tag{8}$$

where $\mathcal{C}_{\mathcal{P}}$ is the ensemble of monomer indices associated to the selected peak positions. In summary, the total energy of a polymer containing random connected and prescribed peaks, is the sum of three energies 1-2,8,

$$\Phi(R_1, .., R_N) = \phi_{Rand}(R_1, .., R_N) + \phi_{Peaks}(R_1, ..., R_N) + \phi_{Rouse}(R_1, .., R_N), \tag{9}$$

and the stochastic equation of motion for $n = 1, .., N$ is

$$\frac{dR_n}{dt} = -\nabla_{R_n}\Phi(R_1, .., R_N) + \sqrt{2D}\frac{d\omega_n}{dt} \tag{10}$$

where $D = \frac{k_B T}{\gamma}$ is the diffusion constant, γ is the friction coefficient, and ω_n are independent 3-dimensional Gaussian noise with mean 0 and standard deviation 1.

2.2 Polymer model associated 5C data

To account for the 5C-data, comprised of a subsection of the X-chromosome from female mice embryonic stem cells reported in [3], showing TAD D and

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

E as two diagonal blocks (Fig. 1A), we use the coarse-grained procedure of [5], with a polymer of length $N = 307$. Each monomer represents a genomic segment of $3kb$ and is connected to its 2 nearest neighbors by an harmonic spring (see subsection above). TAD D (resp. E) is represented by the range of beads from 1 to $N_D = 106$ (resp. 107 to 307). The number of monomer is TAD E is $N_E = 201$.

To reproduce the empirical EP extracted from data, we connected non nearest neighbor pairs of monomers chosen randomly with probability $\frac{1}{N_1}$ (resp. $\frac{1}{N_2}$)) where $N_1 = (N_D - 2)(N_D - 1)/2$ (resp. $N_2 = (N_E - 2)(N_E - 1)/2$). The number of connectors in each TAD is a fraction ξ of the total possible number of non nearest neighbor pairs, for TAD D (resp. E), we have $C_D = \xi_D \frac{N_1}{100}$ (resp. E $C_E = \xi_E \frac{N_2}{100}$) and $\xi_D, \xi_E \in [0, 100]$ will be extracted from data. Random connectors were not added between monomers belonging to different TADs. The procedure of adding random loops in a Rouse polymer is implemented using the energy of random loops, as described in the previous section. Finally, 24 connectors were added (see SI) to all polymers, corresponding to the selected peaks present in the EP matrix.

2.3 Numerical simulations of the reconstructed polymer model

Using the method described in the two previous subsections, we generated polymer realizations, each differ in the position of random connectors inside each TAD. To generated statistics for the EP, we simulate the polymer past its relaxation time τ_R. The time τ_R is determined for each realization by $\tau_R = 1/(\kappa_{min}\lambda_1)$, where D is the diffusion coefficient, κ_{min} is the minimal positive spring constant, and λ_1 is the smallest non-vanishing eigenvalue of the polymer's connectivity matrix [24], which we calculate numerically. In simulations, we divide equation 10 by $\sqrt{D}$ and the spring constants are scaled by the fiction coefficient such that$\kappa = \frac{3k_BT}{\gamma b^2}$. The encounter frequency matrix of the 307 monomers is computed at time τ_R, where two monomers are considered to have encountered if their distance is less than ϵ. The time step for all simulations is $\Delta t = 10^{-2}[sec]$.

2.3.1 Parameters

We summarize in Table 1 the values of parameters used in simulations

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Parameter	Value	Description
b	0.6 $[\mu m]$	STD of distance between adjacent monomers
D	4×10^{-2} $[\mu m^2/sec]$	Monomer diffusion coefficient [25]
ϵ	0.03 $[\mu m]$	Encounter distance
γ	3.1×10^{-5} Ns/m	Friction coefficient citeAmitai2012a
κ	3×10^{-5} N/m	Spring constant

Table 1: values of simulation parameters

3 Results

3.1 The encounter probability of coarse-grained 5C data

We first described how we constructed a polymer model from the symmetrized 5C frequency matrix M (Fig. 1A). By symmetrizing the EP matrix, we averaged-out asymmetrical fluctuations. The 5C data we used represent a sub-region of the X chromosome ($\approx 92,000$ bps), that was previously segmented into two regions called Topological Associating Domains (TADs) D and E [3]. The matrix M was further coarse-grained by beaning the encounter frequencies into 307 monomers of $3kbps$ [5], where TAD D (resp. TAD E) is represented by the first 106 monomers (resp. 107-307), as shown in Fig. 1A. We introduce a general polymer model (Fig. 1B) with arbitrary configuration, the properties of which will be extracted from M_{em}.

The encounter probabilities between monomer m and monomer n are computed from matrix M by

$$P_e(|m-n||n) = \frac{M_{n,n+|m-n|} + M_{n,n-|m-n|}}{\sum_{m=1}^{N} M_{n,m}}, \tag{11}$$

which depends on the genomic distance $|m-n|$ (Fig.1C). Although the encounter probabilities decayed with $|m-n|$ for each n, they contain peaks that reflect consistent long-range interactions between monomers. To quantify the decay of the EP, we fitted its average value $P_e(|m-n|) = \frac{1}{N}\sum_{n=1}^{N} P_e(|m-n||n)$ (black dotted line in Fig. 1C) with the function

$$\tilde{P}(|m-n|) = \frac{C}{|m-n|^{\beta}}, \tag{12}$$

where C and $\beta > 0$ are two constants. For a Rouse polymer, the EP function $\tilde{P}$ is characterized by a decay exponent $\beta = 3/2$ [23]. Fitting (12) to data,

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

revealed that $\beta = 0.77$, from which we concluded that the polymer model should be modified to account for higher compaction than allowed by a Rouse polymer [26].

To better account for the heterogeneity in the EP of each monomer, we plotted the distributions of the exponent β_n for $n = 1..307$ along the polymer (Fig. 1D blue dots). The exponents β_n were extracted by fitting the function 12 to the empirical EPs 11. The large variability in β_n, $n = 1..307$ reflects the local heterogeneity of the chromatin architecture at the current scale (a monomer represents 3kbps). The average β for TAD D and E was found to be $\beta_D = 0.81$ and $\beta_E = 0.78$, respectively. The local minima of β deviated significantly from the mean values (Fig.1D red squares), which can be associated with chromatin features (Fig. 1E): such values can arise due to specific long-range interactions, or boundary between chromatin subdomains. Point m_{100} (monomers 102-107 in Fig. 1D red) is indeed located at the boundary between TAD D and E, while m_{24} and m_{162} are characterized by strong long-range interactions. To conclude, the distribution of β values extracted from the EP is heterogeneous, which can disclose chromatin subregions and long-range strong interactions. We shall account in the next two sections for these characteristic features and include in our polymer model both random, and persistent long-range connections between monomer.

3.2 Encounter probability of random loop polymer model

To determine the level of connectivity that should be added to a generalized Rouse polymer in order to reproduce the EP-decay with a prescribed exponent β, we first studied the case of one TAD-like region in a 307 monomer chain. We added connections between random non nearest neighbor monomer-pairs in the subregion 103-203 (Fig. 2A). The number of connectors, or the connectivity percentage ξ (number of connected monomer-pairs, see Materials and Methods), was increased in the range $0 - 2\%$. By adding connectors, the EP between distant monomers has increased, as represented in the EP-matrix (Fig. 2B). In contrast, outside the region 103-203 the EPs were similar to the case $\xi = 0$ (linear chain), showing that the connected region did not affect the EP in the non-connected one. At this stage, we have shown that adding random connectors allows recovering some statistical characteristics of a TAD region.

To find the optimal number of connectors necessary to recover a given TAD, we set out to elucidate the relationship between the connectivity percentage ξ and the decay exponent β. We have simulated an ensemble of polymer to their relaxation time (see Materials and Methods), and used the equi-

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

librium configuration to estimate the EP of each monomer for $\xi \in [0, 2]\%$. We calculated β by fitting the function 12 to simulation data: the values of β_n for $n \in [103, 203]$ decreased with ξ. Indeed, for $\xi \approx 0.2\%$, the coefficients β_n decreased below the Rouse exponent (equal to $\beta_{Rouse} = 1.5$), indicating compact polymer configurations. For $\xi = 2\%$, the mean exponent β_n and $n = 103 - 203$ was $\bar{\beta} = 0.47$, with a minimal value 0.42 obtained for the boundary monomers 103 and 203 (Fig. 2C). These results confirm that the polymer condenses in smaller regions, measured by the mean square radius of gyration R_g [23], decaying from $R_g^2 = 11.9$ for $\xi = 0$ to $R_g^2 = 11$ at $\xi = 2\%$ (Fig. 2D). Values of β outside the TAD region (monomers 1 to 102 and 204-307) were mostly unchanged, fluctuating around $\beta = 1.5$, confirming that statistical Rouse properties are not affected when connectors are added to the middle region (103-203). Finally, the average value of β_n (computed over $n = 103 - 203$) versus ξ is shown in Fig. 2D and, as we shall see, can serve to extract the connectivity percentage ξ from the empirical data.

To reproduce the two TAD E and D regions of the X-chromosome, we started with a polymer of 307 monomers and added random connectors (green arrows) between monomers 1-106 and between monomers 107-307, as described in Fig. 2E upper panel). This partition follows the empirical TAD segmentation [3, 5]. Three polymer realizations for $\xi = 0, 0.2, 1\%$ are shown in Fig. 2E bottom panel, showing polymer condensation into two regions. The encounter frequency matrix show that two TAD-like regions, named TAD1 and TAD2 (Fig. 2F), emerge as ξ increases from 0 to 2%. To extract the exponent β (Fig. 2G, colored curves) we fitted the function 12 to the EP-matrix for each monomer inside TAD1 and 2. In both cases, the exponent β decreased below $\beta_{Rouse} = 1.5$ ($\xi = 0\%$ blue curve) and for $\xi = 0.2\%$, 9 and 39 random connectors were added for TAD1 and TAD2, respectively. The boundary between TADs is characterized by an abrupt decay of the beta value. The β exponent (averaged over each TAD), plotted with respect to the connectivity ξ (Fig. 2H), was used to determine the number of connectors necessary to reconstruct the empirical data. Indeed, we extracted from the EP matrix (Fig. 1A) that $\beta_D = 0.81, \beta_E = 0.78$ the associated connectivity percentages $\xi_D = 0.12, \xi_E = 0.23$, respectively (Fig. 2H). To conclude, we estimated the number of random connectors needed to be added on a Rouse polymer such that the β exponent of the reconstructed and empirical EP-matrix are identical.

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.3 Incorporating long-range empirical interactions in the polymer model

Another key feature present in the 5C EF-matrix (Fig. 1A) is the ensemble of consistent long-range interactions between monomers (Fig. 1C). To account for these interactions, we connected monomers corresponding to off-diagonal local maxima of the EF matrix, which exceed a given threshold (see Materials and Methods). We found 24 long-range connections: 7 (resp. 13) inside TAD D (resp. E) and 4 across the two (see SI for the list of monomers pairs) as shown in Fig. 3A. As revealed by simulations of a Rouse polymer with added connectors, the polymer configuration are condensed, characterized by the radius of gyration about $R_g = 9.1$ (compared to 12 for the Rouse chain). Three realizations of the polymer are shown in Fig. 3A. We adjusted the spring constant $\kappa_{m,n}$ between monomer m and n corresponding to consistent long range interactions, based on the values of their EPs (see Materials and Methods). The scaled coefficient $\kappa_{m,n}$ are summarized in the SI, and are between 1.1 to 3 times higher than the spring constant assigned to connectors of the linear backbone ($\kappa = 0.97$).

To quantify the effect of adding long-range connections, we computed the exponents β_n by fitting the function 12 to the EPs from simulations, and compared it to the ones computed from the experimental data (Fig. 3B). The experimental β_n (red) were generally lower than the ones obtained from simulations (blue), indicating that the reconstructed chromatin polymer is more condensation in both TADs. Furthermore, we concluded that the addition of long-range connectors is insufficient to reproduce the statistics of the 5C data.

3.4 Combination of random loops and long-range interactions to construct a polymer model of a TAD

We previously evaluated separately the effect of random connectors and long-range interactions on the EPs. We computed the decay exponent β, that we compared with coarse-grained 5C data. We now combine these two constraints, such that specific and non-specific connectors are added to a generalized Rouse polymer (Fig. 4). The connectivity percentage matching that of the experimental data in each TAD was found to be $\xi_D = 0.23\%$ and $\xi_E = 0.12\%$. These value are summarizing the sum of the contribution from the two types of connectors (see Fig.2H). For long-range specific interactions (Fig. 3), we have previously obtained $\beta_D = 1.02, \beta_E = 0.99$, corresponding to $\xi_D = 0.12\%, \xi_E = 0.07\%$(Fig. 2H) for TAD D and E, respectively. We,

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

therefore, attributed the remaining percentages to the addition of random connectors, that is $\xi_D = 0.11, \xi_E = 0.05$. The number of added random connectors corresponding to $\xi_D = 0.11\%, \xi_E = 0.05\%$ are 6 and 10 in TAD D and E, respectively.

We started with a Rouse chain (Fig. 4A (gray)) and added connectors between monomer pairs corresponding to peaks of the EP matrix (red). Three polymer realizations, simulated with the two types of connectors, are shown in Fig. 4B, characterized by a radius of Gyration $R_g = 6.4$. We computed the EF-matrix (Fig. 4C) that showed several similarity with the experimental data (compare Fig. 1A with Fig. 4B), for which two TAD-like structures are visible. Next, we quantified the similarity between the two matrices by comparing the decay exponent β_n,$(n = 1..307)$. We used the function 12 to fit the EP of monomers $1 - 307$ after long time polymer simulations (see Materials and Methods). The fitted value for β shows an excellent agreement with the experimental β values (Fig 3C). To conclude, long deterministic and short-range stochastic interactions are sufficient to account for the decay rate of EP extracted from coarse-grained 5C data.

3.5 Simulation of transient property of TADs: conditional encounter probabilities and mean times for three genomic sites to interact

We showed previously how to construct a polymer, the statistical properties of which match the ones extracted from 5C data. However, these data cannot be used to study transient properties of the chromatin, as they represent a static genomic encounter interactions, averaged over cell realizations. We shall use now the reconstructed polymer described above, constrained by the steady-state properties of the 5C data (Fig. 5A), to evaluate the transient properties of the chromatin. We estimate the mean first encounter time and the probability that monomer 26 (position of the Linx) meet monomer 87 (Xite) before monomer 64 (Chic1). These monomers represents three key sites on the X chromosome [5, 3], located in TAD D. We show three realizations and indicate the location of the three sites inside TAD D (yellow) and E (blue), that do not mix.

We started the polymer simulation from the steady-state distribution and use 10 000 runs. As predicted by the narrow escape theory [27, 28], the encounter time between two of the three monomers is Poissonian, as confirmed by their computed distribution (Fig. 5B), for which the reciprocal of the encounter rate is the mean encounter time. We found that the EP is $P = 0.55$, while the mean times are quite comparable of the order of

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

131s (see table C in Fig. 5). Finally, to check the impact of TAD E on the mean encounter time inside TAD D, we ran another set of stochastic simulations after we removed TAD E (Fig. 5D). Surprisingly, the EP was inverted compared to the case of no deletion, while the mean encounter has increased by almost 50% to 195s (Linx to Chic1) and 205s (Linx to Xite). This result suggests that TAD E contributed in modulating the interaction probability and the mean time, and thus further indicates that the search time inside a TAD depends on neighboring chromatin configuration.

4 Discussion

We presented here a general method to construct a coarse-grained polymer model from 5C encounter probability (EP) matrix. This construction preserves some of the statistical properties of the 5C data, such as the decay rate of the EP of each monomer. The method uses Rouse polymer as a starting point, where persistent long-range connectors, corresponding to local peaks of the 5C data, and random connectors are added in each realization. Connectors are represented by springs between long-range monomer-pairs and reflect the ensemble of chromatin architectures, as seen in the 5C-matrix. Sub-regions with encounter enrichment (TADs) are accounted for by adding connectors between random monomer-pairs, the number of connectors was extracted from EP-matrix. Finally, although contact maps (5C data) represents a steady-state distribution of an ensemble of looping events, patterns such as TADs appear in a large sample of millions of nuclei. We have demonstrated that using polymer reconstruction, transient properties can be studied, such as the mean encounter time between any two sites. The present approach is quite different from reconstruction methods that consist on inferring 3D structures of a genome from 5C data from contact frequency between sequences, which can be assumed to be Poissonnian [29] or not [30, 31].

Interestingly, TADs emerge here as a consequence of adding independent connectors randomly attached to monomers inside the subpart of the polymer that defines the TAD. These connectors could represent binding proteins such as cohesin [19] in the heterogeneous chromatin population of the 5C experimental data. Previous models explore the effect of connectors between regions of the chromatin [14, 13] and examine the consequence on the EP-decay rate. Here, we use random connectors to resolve a reverse engineering problem, and to recover the degree of connectivities from the EP-decay rate (see Fig.2). To accurately account for the statistical prop-

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

erties of the EP-matrix, we used the exponent β of each monomer, and estimated the strength of interactions. For that reason, the present model extends the binders model developed in [14] that consider constant binding.

The mean looping time in free space depends only on the distance between monomers, while in confined domain this time also includes the effect of the boundary, which affects distant monomers [32, 28, 26]. It was thus surprising that a chromatin TAD subregion could influence the mean encounter between monomers located in a different region (Fig. 5). This effect is certainly due to the long-range inter-TAD interactions. Finally, the present method and algorithm can be used to reconstruct a polymer model at a given scale (number of monomers and number of bps coarse-grained in a monomer) that should be specified by the user. The polymer model following the specification described here, should reproduce the local condensation and can be used to study any transient chromatin properties involving any monomer pair.

5 Funding

This work was supported by the Marie Curie Award.

6 Acknowledgments

We thank E. Heard, A. Amitai and L. Giorgetti for many interesting discussions on this subject.

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Dekker, J. (2014) Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture. *Epigenetics & chromatin,* **7**(1), 1.

[2] Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research,* **16**(10), 1299–1309.

[3] Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., et al. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature,* **485**(7398), 381–385.

[4] 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,* **485**(7398), 376–380.

[5] Giorgetti, L., Galupa, R., Nora, E. P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014) Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell,* **157**(4), 950–963.

[6] Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., et al. (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature,* **515**(7527), 402–405.

[7] Langowski, J. and Heermann, D. W. (2007) Computational modeling of the chromatin fiber. In *Seminars in cell & developmental biology* Elsevier Vol. 18(5), pp. 659–667.

[8] Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature,* **502**(7469), 59–64.

[9] Branco, M. R. and Pombo, A. (2006) Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol,* **4**(5), e138.

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[10] Nicodemi, M. and Pombo, A. (2014) Models of chromosome structure. *Current opinion in cell biology,* **28**, 90–95.

[11] Junier, I., Martin, O., and Képès, F. (2010) Spatial and topological organization of DNA chains induced by gene co-localization. *PLoS Comput Biol,* **6(2)**(2), e1000678.

[12] Bohn, M. and Heermann, D. W. (2010) Diffusion-driven looping provides a consistent framework for chromatin organization. *PloS one,* **5**(8), e12218.

[13] Olarte-Plata, J. D., Haddad, N., Vaillant, C., and Jost, D. (2016) The folding landscape of the epigenome. *Physical Biology,* **13**(2), 026001.

[14] Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, A., and Nicodemi, M. (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences,* **109**(40), 16173–16178.

[15] Trieu, T. and Cheng, J. (2014) Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic acids research,* **42**(7), e52–e52.

[16] Dion, V. and Gasser, S. M. (2013) Chromatin movement in the maintenance of genome stability. *Cell,* **152**(6), 1355–1364.

[17] Gasser, S. M. (2016) Nuclear Architecture: Past and Future Tense. *Trends in Cell Biology,*.

[18] Seeber, A., Dion, V., and Gasser, S. M. (2013) Checkpoint kinases and the INO80 nucleosome remodeling complex enhance global chromatin mobility in response to DNA damage. *Genes & development,* **27**(18), 1999–2008.

[19] Bronstein, I., Israel, Y., Kepten, E., Mai, S., Shav-Tal, Y., Barkai, E., and Garini, Y. (2009) Transient anomalous diffusion of telomeres in the nucleus of mammalian cells. *Physical review letters,* **103**(1), 018102.

[20] Javer, A., Long, Z., Nugent, E., Grisi, M., Siriwatwetchakul, K., Dorfman, K. D., Cicuta, P., and Lagomarsino, M. C. (2013) Short-time movement of E. coli chromosomal loci depends on coordinate and subcellular localization. *Nature communications,* **4**.

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[21] Javer, A., Kuwada, N. J., Long, Z., Benza, V. G., Dorfman, K. D., Wiggins, P. A., Cicuta, P., and Lagomarsino, M. C. (2014) Persistent super-diffusive motion of Escherichia coli chromosomal loci. *Nature communications,* **5**.

[22] Amitai, A., Toulouze, M., Dubrana, K., and Holcman, D. (2015) Analysis of Single Locus Trajectories for Extracting In Vivo Chromatin Tethering Interactions. *PLoS Comput Biol,* **11**(8), e1004433.

[23] Doi, M. and Edwards, S. (1986) The Theory of Polymer Dynamics Clarendon, Oxford, .

[24] Gurtovenko, A. A. and Blumen, A. (2005) Generalized Gaussian Structures: Models for Polymer Systems with ComplexTopologies, Springer, .

[25] Amitai, A., Amoruso, C., Ziskind, A., and Holcman, D. (2012) Encounter dynamics of a small target by a polymer diffusing in a confined domain. *The Journal of chemical physics,* **137**(24), 244906.

[26] Amitai, A. and Holcman, D. (2013) Polymer model with long-range interactions: Analysis and applications to the chromatin structure. *Physical Review E,* **88**(5), 052604.

[27] Holcman, D. and Schuss, Z. (2015) Stochastic Narrow Escape in Molecular and Cellular Biology: Analysis and Applications, Springer, .

[28] Amitai, A., Kupka, I., and Holcman, D. (2012) Computation of the mean first-encounter time between the ends of a polymer chain. *Physical review letters,* **109**(10), 108302.

[29] Varoquaux, N., Ay, F., Noble, W. S., and Vert, J.-P. (2014) A statistical approach for inferring the 3D structure of the genome. *Bioinformatics,* **30**(12), i26–i33.

[30] Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A., and Noble, W. S. (2010) A three-dimensional model of the yeast genome. *Nature,* **465**(7296), 363–367.

[31] Hajjoul, H., Mathon, J., Ranchon, H., Goiffon, I., Mozziconacci, J., Albert, B., Carrivain, P., Victor, J.-M., Gadal, O., Bystricky, K., et al. (2013) High-throughput chromatin motion tracking in living yeast

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reveals the flexibility of the fiber throughout the genome. *Genome research,* **23**(11), 1829–1838.

[32] Jespersen, S., Sokolov, I., and Blumen, A. (2000) Small-world Rouse networks as models of cross-linked polymers. *The Journal of Chemical Physics,* **113**(17), 7652–7655.

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

7 Figure Captions

Figure 1: **Statistics of Conformation Capture data**. **A.** Average encounter frequency map of two 5C replica spanning ≈1Mbp genomic region containing two Topologically Associating Domain (TAD) D (monomers 1-106) and TAD E (monomers 107-307) [3], where the map was coarse-grained into 307 monomers of size 3kb [5]. **B.** Schematic representation of a polymer model with randomly connected monomers. **C.** Empirical encounter probability, P_n, for monomer n plotted with respect to the genomic distance d [monomer units], reveals long-range interactions (localized peaks). P_n are fitted with functions $Ad^{-\beta}$, where β is the decay exponent and A the normalization factor. For the mean encounter probability $\bar{P}$, we have $A = 0.08$ and $\beta = 0.77$ (thick red curve). **D.** Distribution of the β_n exponents ($n = 1..307$) (blue dots) displaying high variability: Monomers m_{24}, m_{100}, m_{162} (red square dots) with $\beta_{24} = 0.22$, $\beta_{100} = 0.46$, and $\beta_{162} = 0.33$, respectively, accounts for high peaks (first and last), while the middle one corresponds to the boundary between TADs. **E.** Encounter probability P_n for monomers n=24, 100, and 162, corresponding to local minima shown in box D.

bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2: **Statistics of simulated generalized Rouse polymer chain for various connectivity in one and two sub-regions A.** Schematic bead-spring chain connected at random positions (two-sided green arrows) between non-nearest-neighbor monomers. **B.** Encounter frequency maps of a 307 monomers chain, where connectors are added randomly between monomers 103-202 for each realization. TAD-like structure emerges as the connectivity ξ (number of connectors) increase from 0 to 2%. **C.** Distribution of β_n ($n = 1..307$) fitted by the function $Ad^{-\beta}$, where d is the distance along the chain [monomer units], to the encounter probabilities of numerical simulation. The TAD regions is characterized by a low β values. **D.** Average value of β for monomers in the interval 103-202 with respect to the connectivity percentage ξ, which is decaying due to the increase of long-range interaction between monomers of the polymer. This decay is correlated with the one of the radius of gyration (embedded sub-figure). **E.** Schematic polymer chain, where two defined regions: monomers 1-106 (TAD 1, orange circles) and monomers 107-307 (TAD 2, blue circles), are randomly connected (green arrows). No connections were added between the two TAD regions. We present in the lower panel, three snapshot realizations of a random loop chain with TAD 1 (orange) and TAD 2 (blue) and random connectors (green) for three increasing values of connectivity $\xi = 0, 0.2, 1\%$. **F.** Encounter frequency maps showing the two TAD regions for an increasing number of random connectors. **G.** Distribution of β exponent for $\xi \in [0, 2]\%$, showing the border (n=106) effect between TADs. **H.** Average β over each TAD 1 (blue) and TAD 2 (orange) for $xi \in [0, 2]$. The curves decrease until plateau at 0.42 (0.24) for TAD 1 (resp. TAD 2). We use these curves to recover the connectivity percentage ξ of the experimental TAD D, with $\beta_D = 0.74$ (resp. TAD E with $\beta_E = 0.81$) for which $\xi_D = 0.23\%$ (resp. $\xi_E = 0.12\%$.

.

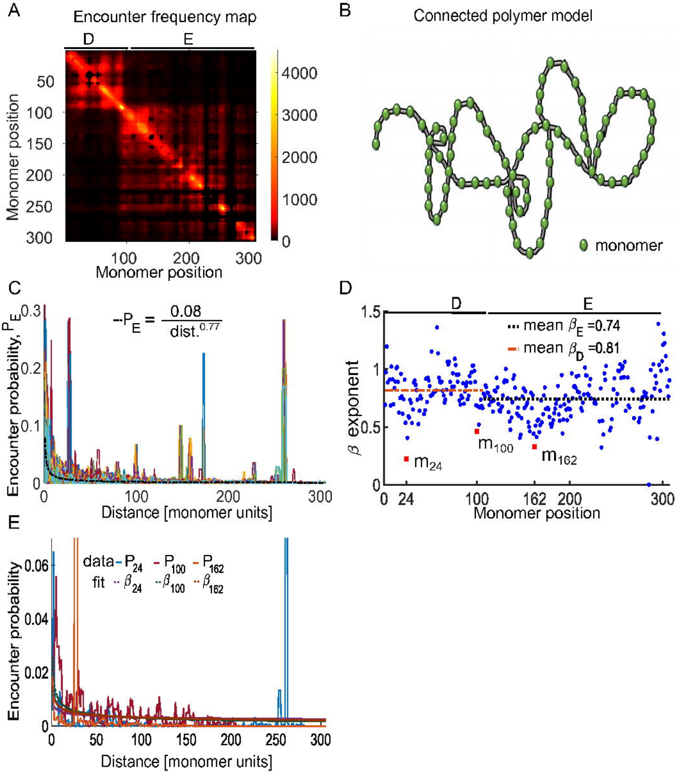
bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

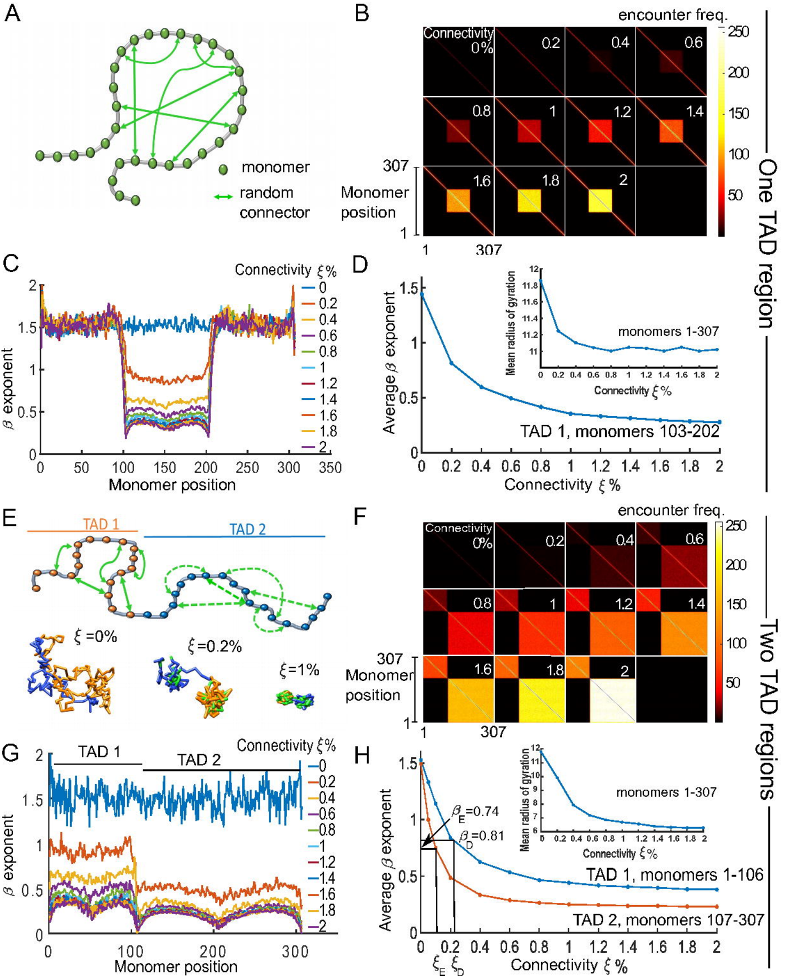
Figure 3: **Effect of persistent long-range connectors on polymer folding.** The connectors account for peaks present in the empirical 5C encounter frequency maps. **A.** Upper panel: schematic representation of the bead-spring polymer model with added fixed connectors (red) representing specific long-range monomer interactions (peaks) shown in Fig. 1C. Lower panel: three different realization of the same polymer, showing TAD D (orange), TAD E (blue), and fixed connectors (red). **C.** Simulated (blue) and experimental (red) β exponent of the fitted encounter probability. The polymer model contains only specific long-range interactions (no random connectors $\xi_D = 0$, $\xi_E = 0$). Such a model cannot reproduce the encounter frequency maps of the experimental data. The average β values for TAD D and E are $\beta_D = 1.02$ (resp. $\beta_E = 0.99$).

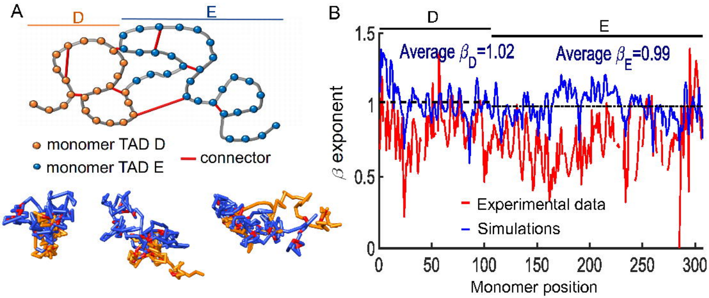
Figure 4: **Coarse-grained reconstruction of chromatin using extracted random loops and connectors corresponding to peaks of the 5C data**. **A.** Schematic polymer model, where TAD D (orange, monomers 1-106), and TAD E (blue, monomers 107-307) are recovered by random loops (green arrows) according to the connectivity ξ and persistent long-range connectors (red bars), corresponding to peaks of the 5C data. **B.** Three realizations of the polymer model. **C** Encounter frequency matrix of the simulated polymer model, showing two TADs, where the off-diagonal points correspond to fixed connectors. **D.** Comparison between β computed from experiments and simulations data, confirming that the present polymer model accounts for the statistics of encounter frequencies.

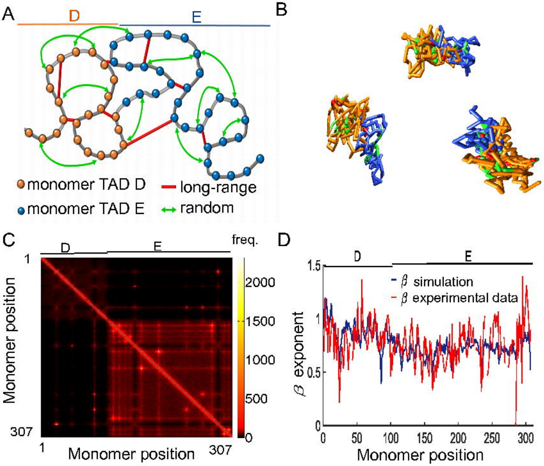
bioRxiv preprint doi: https://doi.org/10.1101/065102; this version posted July 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

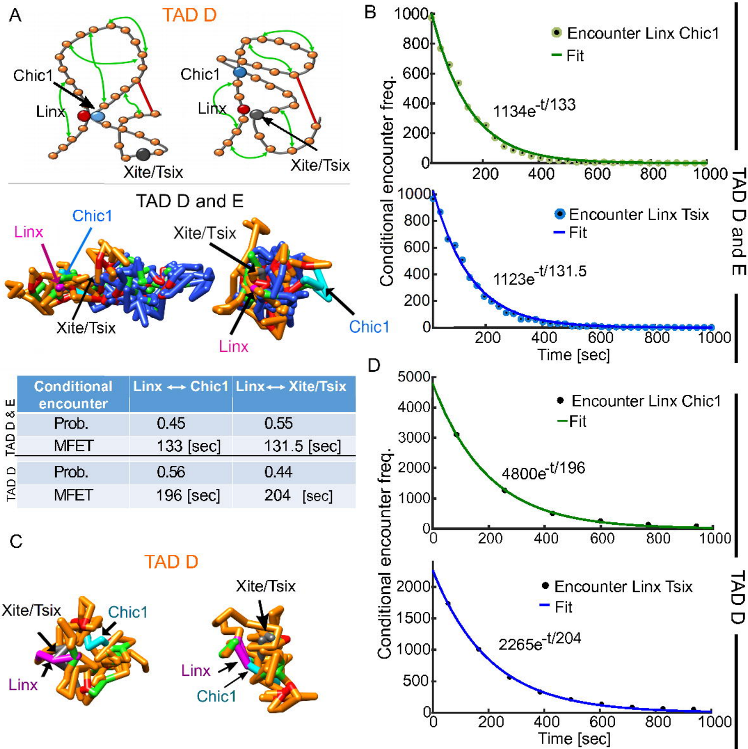
Figure 5: **Transient properties of the chromatin: Conditional mean time and probability for three sites to meet**. **A.** (upper panel) Representation of the polymer model for TAD D (orange, monomers 1-106), where loci Linx (monomer 26, red) meets Chic1 (monomer 67, cyan) and Xite/Tsix (monomer 87, gray), respectively. Random connectors (green arrows) and specific long range-connectors (red bar) are added, following the connectivity ξ recovered from data. Fixed connectors (red bars) correspond to specific peaks of the 5C data. Three realizations (bottom panel) of the polymer model containing TAD D and E, show the encounter of Linx (magenta) with Chic1 (cyan), and Xite/Tsix (gray), respectively. The color code is from the upper panel. **B.** Histogram of the conditional encounter times between Linx and Chic1 (upper panel, green),and Linx and Xite/Tsix (bottom panel, blue) with TAD D and E. **C.** Two polymer realization with a single TAD D (monomers 1-106, orange), showing the encounter between Linx (magenta) and Xite/Tsix (gray, left panel), and the encounter between Linx and Chic1 (cyan, right panel). **D.** Histogram of the conditional encounter times for a polymer with only TAD D, showing an exponential decay as in sub-figure C.


A
Encounter frequency map
D
E
Monomer position
50
100
150
200
250
300
100
200
300
Monomer position
4000
3000
2000
1000
0
B
Connected polymer model
monomer
C
$-P_E = \frac{0.08}{dist.^{0.77}}$
Encounter probability, P_E
0.3
0.2
0.1
0
0
100
200
300
Distance [monomer units]
D
D
E
mean β_E =0.74
mean β_D =0.81
β exponent
1.5
1
0.5
0
m_{24}
m_{100}
m_{162}
0
24
100
162
200
300
Monomer position
E
data
P_{24}
P_{100}
P_{162}
fit
β_{24}
β_{100}
β_{162}
Encounter probability
0.06
0.04
0.02
0
0
50
100
150
200
250
300
Distance [monomer units]


A
monomer
random connector
B
encounter freq.
Connectivity 0%
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
307
Monomer position
1
1
307
250
200
150
100
50
One TAD region
C
Connectivity ξ%
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
β exponent
Monomer position
0
50
100
150
200
250
300
350
D
Average β exponent
Connectivity ξ %
TAD 1, monomers 103-202
Mean radius of gyration
monomers 1-307
Connectivity ξ %
E
TAD 1
TAD 2
ξ =0%
ξ =0.2%
ξ =1%
F
encounter freq.
Connectivity 0%
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
307
Monomer position
1
1
307
250
200
150
100
50
Two TAD regions
G
TAD 1
TAD 2
Connectivity ξ%
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
β exponent
Monomer position
0
100
200
300
H
Average β exponent
β_E=0.74
β_D=0.81
Mean radius of gyration
monomers 1-307
Connectivity ξ %
TAD 1, monomers 1-106
TAD 2, monomers 107-307
ξ_E
ξ_D
Connectivity ξ %


A
D
E
monomer TAD D
monomer TAD E
connector
B
D
E
1.5
Average β_D=1.02
Average β_E=0.99
β exponent
1
0.5
0
Experimental data
Simulations
0
50
100
150
200
250
300
Monomer position


A
D
E
monomer TAD D
monomer TAD E
long-range
random
B
C
D
E
freq.
1
307
Monomer position
Monomer position
1
307
2000
1500
1000
500
0
D
D
E
1.5
1
0.5
0
β simulation
β experimental data
β exponent
0
50
100
150
200
250
300
Monomer position



	Conditional encounter	Linx ⟷ Chic1	Linx⟷ Xite/Tsix
TAD D & E	Prob.	0.45	0.55
	MFET	133 [sec]	131.5 [sec]
TAD D	Prob.	0.56	0.44
	MFET	196 [sec]	204 [sec]