Manuscript submitted to **Biophysical** *Journal*
Article

Computing 3D chromatin configurations from contact probability maps by Inverse Brownian Dynamics

K. Kumari, B. Duenweg, R. Padinhateeri, J. R. Prakash

ABSTRACT The three-dimensional organization of chromatin, on the length scale of a few genes, is crucial in determining the functional state — accessibility and the amount of gene expression — of the chromatin. Recent advances in chromosome conformation capture experiments provide partial information on the chromatin organization in a cell population, namely the contact count between any segment pairs. However, given the contact matrix, determining the complete 3D organization of the whole chromatin polymer is an inverse problem. In the present work, an Inverse Brownian Dynamics (IBD) method has been proposed to compute the optimal interaction strengths between different segments of chromatin such that the experimentally measured contact count probability constraints are satisfied. Applying this method to the α-globin gene locus in two different cell types, we predict the 3D organizations corresponding to active and repressed states of chromatin at the locus. We show that the average distance between any two segments of the region has a broad distribution and cannot be computed as a simple inverse relation based on the contact probability alone. We also address the normalization problem of the contact count matrix and argue that extra measurements of polymer properties such as radius of gyration may be required to resolve the problem.

SIGNIFICANCE Chromosome conformation capture experiments such as 5C and Hi-C provide information on the contact counts between different segments of chromatin, but not the interaction strengths that lead to these counts. Here a methodology is proposed by which this inverse problem can be solved, namely, given the contact probabilities between all segment pairs, what is the pair-wise interaction strength that leads to this value? With the knowledge of pair-wise interactions determined in this manner, it is then possible to evaluate the 3D organization of chromatin and to determine the true relationship between contact probabilities and spatial distances.

1 INTRODUCTION

Even though all the cells in multi-cellular organisms have the same DNA sequence, they function differently based on the cell type. For example, the phenotype of a skin cell is significantly different from that of a neuronal cell (1, 2). One of the important factors for this variation is hypothezised to be the three-dimensional organization of DNA inside the cell nucleus and its variability from cell-type to cell-type (3–6). While findings of the recent chromosome configuration capture experiments (3C, 4C, 5C, Hi-C) (7–10) lend credence to this hypothesis, the outcomes of these experiments are frozen snapshots of a sparse set of points along DNA that do not give a complete understanding of the 3-dimensional organization of the genome. In this paper, a methodology based on a coarse-grained polymer model for DNA is proposed, which enables the unravelling of its spatio-temporal organization that is consistent with experimentally observed contact maps.

The complex folding of meter-long DNA into micrometer-sized chromosome, with topologically associated domains (TAD), has been revealed at few kb resolution by state of the art Hi-C experiments (11–14). More insight into the role played by the 3D organization of the genome in the functioning of a cell on the length-scale of genes, is provided by 3C and 5C experiments. Essentially, all these chromatin conformation capture experiments lead to information on the count of contacts between any pair of segments along the DNA chain backbone, represented in the form of contact ("heat") maps.

Several attempts have been made to understand the 3D organization of the genome using a variety of techniques developed previously to understand the statics and dynamics of polymers (15–22). Early models revealed both the fractal nature of individual chromosome folding and the larger scale packing of all the chromosomes into TADs of various sizes (23, 24). Subsequent studies that focus on reconstructing the 3D structure from the contact maps are predominantly based on assuming that there is a direct correlation between the magnitude of the contact and the spatial distance between the relevant pairs (25–31). These investigations have led to important insights about the 3D consequences of differences present in the contact maps such as the spatial organization of ON and OFF states of certain genes. However, all these efforts have certain limitations.

As mentioned above, nearly all the computational studies convert contact counts obtained from Hi-C experiments into spatial distances, using a pre-decided formula. That means, given a contact count matrix, such methods do not predict the

distances between different chromatin segments, rather they take the distance values as inputs, based on certain assumptions. They then use conventional Monte Carlo (or equivalent) methods to find steady state configurations of the chromatin, given a distance map between different DNA segment pairs. In other words, the existing models consider this as a "forward" problem of computing equilibrium configurations of chromatin as a consequence of assuming certain spatial distance between bead-pairs. However, the problem of computing 3D configurations of chromatin polymer, given a contact map, is not a "forward" problem but rather an "inverse" problem (32). The question is, given a contact map, what are the optimal interactions between different segments of chromatin such that the experimentally seen contact map emerges. To the best of our knowledge, no study exists that solves chromatin configurations of genes, considering it as an inverse problem. Another shortcoming of the existing computational approaches is that they do not address the normalization problem. Since there is no unique way of normalizing and converting a contact count matrix into a contact probability matrix, the existing methods cannot be used to compare 3D configurations between different data sets. Moreover, all these efforts are based on Monte Carlo methods, and hence they cannot predict the dynamics of chromatin — they only obtain information on static configurations of the genome.

In summary, while current models have made important progress in constructing 3D structure from the contact maps, they suffer from one or more of the following shortcomings:

1. An *a priori* assumption regarding the probability of contact between pairs of segments and their spatial distances.

2. The introduction of harmonic springs between interacting pairs that implies an attractive force between these pairs that does not decay with distance, but rather increases.

3. The use of simulation methods that are limited to providing information on static configurations.

4. Considering the problem of computing 3D configurations as a "forward" problem, with no attempt to determine the interaction strengths between segment pairs that lead to 3D structures that are consistent with observed contact maps.

5. The failure to obtain an accurate representation of dynamic behaviour by failing to include hydrodynamic interactions between segment pairs.

In this work, a methodology is introduced that addresses all these shortcomings. Chromatin on the length-scale of a gene is represented by a coarse-grained bead spring chain polymer model, with a potential of interaction between pairs of beads that can be tuned to accommodate varying strengths of interaction. A Brownian dynamics simulation algorithm that includes hydrodynamic interactions, and an iterative scheme based on Inverse Monte Carlo is developed that enables the generation of 3D configurations that are consistent with the contact maps. This methodology is then applied to obtain the static 3D configurations from 5C contact maps of the α-globin gene locus, both in the ON and OFF states of the gene. Further, since hydrodynamic interaction are taken into account, the approach has the potential to examine the dynamic transitions between the ON and OFF states.

The outline of the paper is as follows. The key governing equations of the model and the simulation algorithm are summarised in section 2.1. In section 2.2, the inverse Brownian dynamics method is introduced in a general context. The validation of the proposed approach with the help of a prototype is presented in section 3.1. Resolution of the issue of determining the contact probabilities from contact counts is proposed in section 3.2. Results for the static 3D configurations of α-globin locus are discussed in section 3.3, while the relationship between spatial distances and contact probabilities is highlighted in section 3.4. The principal conclusions of this work are summarized in section 4.

2 MODEL AND METHODS

2.1 Polymer model

To compute the 3-dimensional organization of the genome, the chromatin is coarse-grained into a bead-spring chain of N beads connected by $N-1$ springs. The chain configuration is specified by the set of position vectors of the beads $\boldsymbol{r}_\mu (\mu = 1, 2, ..., N)$. For simulation purposes, all distances are made dimensionless by using the characteristic length scale $l_0 = \sqrt{k_\mathrm{B}T/k_s}$ arising from the ratio of thermal energy—where k_B is the Boltzmann constant, T is the temperature—and the spring constant k_s. Throughout this manuscript the asterisk superscript is used to indicate dimensionless quantities ($r^*_\mu = r_\mu/l_0$). The adjacent beads in the polymer chain are bonded via a Fraenkel spring, with a non-dimensional spring potential $U^{\mathrm{S}*}_\mu$ between bead μ and $(\mu+1)$, given by

$$U^{\mathrm{S}*}_\mu = \frac{1}{2}\left[\left(r^*_{\mu+1} - r^*_\mu\right) - r^*_0\right]^2 \tag{1}$$

where $(r^*_{\mu+1} - r^*_\mu)$ is the non-dimensional distance between bead μ and $\mu+1$, and r^*_0 is the dimensionless natural length of the Fraenkel spring. To mimic protein-mediated interactions between different parts of the chromatin polymer, it is necessary to introduce a potential energy function. Typically, this is achieved with a Lennard-Jones (LJ) potential or with harmonic spring interactions (32). However, in the present study, the following non-dimensional Soddemann-Duenweg-Kremer (SDK) (33) potential is introduced between any two non-adjacent beads μ and ν,

$$U^{\mathrm{SDK}*}_{\mu\nu} = \begin{cases} 4\left[\left(\dfrac{\sigma^*}{r^*_{\mu\nu}}\right)^{12} - \left(\dfrac{\sigma^*}{r^*_{\mu\nu}}\right)^{6} + \frac{1}{4}\right] - \epsilon_{\mu\nu}; & r^*_{\mu\nu} \le 2^{1/6}\sigma^* \\ \dfrac{1}{2}\epsilon_{\mu\nu}\left[\cos\left(\alpha\dfrac{r^*_{\mu\nu}}{\sigma^*} + \beta\right) - 1\right]; & 2^{1/6}\sigma^* \le r^*_{\mu\nu} \le r^*_{\mathrm{c}} \\ 0. & r^*_{\mu\nu} \ge r^*_{\mathrm{c}} \end{cases} \tag{2}$$

Here $r^*_{\mu\nu} = (r^*_\mu - r^*_\nu)$ is the non-dimensional distance between beads μ and ν, $\epsilon_{\mu\nu}$ is an independent parameter to control the bead-bead attractive interaction strength between beads μ and ν and $2^{1/6}\sigma^*$ represents the minima of the potential where $U^{\mathrm{SDK}*}_{\mu\nu} = \epsilon_{\mu\nu}$. The SDK potential has the following advantages compared to the LJ potential: (i) The repulsive part of the SDK potential ($r^*_{\mu\nu} \le 2^{1/6}\sigma^*$) representing steric hindrance remains unaffected by the choice of the parameter $\epsilon_{\mu\nu}$. (ii) Protein-mediated interactions in chromatin are like effective "bonds" formed and broken with a finite range of interaction. Unlike the LJ potential, the SDK potential has a finite attractive range —the SDK potential energy smoothly reaches zero at the cut off radius, r^*_{c}, whose value is set by the choice of two parameters α and β. The parameters α and β are determined by applying the two boundary conditions, namely, $U^{\mathrm{SDK}}_{\mu\nu} = 0$ at $r^*_{\mu\nu} = r^*_{\mathrm{c}}$ and $U^{\mathrm{SDK}}_{\mu\nu} = -\epsilon_{\mu\nu}$ at $r^*_{\mu\nu} = 2^{1/6}\sigma^*$. The appropriate choice of the cut-off radius r^*_{c} has been investigated extensively in a recent study (34) and it has been shown that a value of $r^*_{\mathrm{c}} = 1.82\sigma^*$ leads to an accurate prediction of the static properties of a polymer chain in poor, theta and good solvents. The same value is adopted here in the present study.

Given a set of values $\epsilon_{\mu\nu}$ and an initial configuration of the bead-spring chain, the time evolution of the configurations of the polymer chain is evaluated using Brownian Dynamics simulations (35), which is a numerical method for solving the following Euler finite difference representation of the stochastic differential equation for the bead position vectors,

$$\boldsymbol{r}^*_\mu(t^* + \Delta t^*) = \boldsymbol{r}^*_\mu(t^*) + \frac{\Delta t^*}{4}\sum_{\nu=1}^{N}\boldsymbol{D}_{\mu\nu}\cdot(\boldsymbol{F}^{s*}_\nu + \boldsymbol{F}^{\mathrm{SDK}*}_\nu) + \frac{1}{\sqrt{2}}\sum_{\nu=1}^{N}\boldsymbol{B}_{\mu\nu}\cdot\Delta\boldsymbol{W}_\nu \tag{3}$$

Here $t^* = t/\lambda_0$ is the dimensionless time, with $\lambda_0 = \zeta/4k_s$ being the characteristic time scale, and ζ is the Stokes friction coefficient of a spherical bead. $\boldsymbol{F}^{s*}_\nu$ and $\boldsymbol{F}^{\mathrm{SDK}*}_\nu$ are the non-dimensional spring and interaction forces computed from the respective potential energy functions provided in Eqs. (1) and (2). $\Delta\boldsymbol{W}_\nu$ is a non-dimensional Wiener process with mean zero and variance Δt^* and $\boldsymbol{B}_{\mu\nu}$ is a non-dimensional tensor whose presence leads to multiplicative noise (35). Its evaluation requires the decomposition of the diffusion tensor $\boldsymbol{D}_{\mu\nu}$ defined as $\boldsymbol{D}_{\mu\nu} = \delta_{\mu\nu}\boldsymbol{\delta} + \boldsymbol{\Omega}_{\mu\nu}$, where $\delta_{\mu\nu}$ is the Kronecker delta, $\boldsymbol{\delta}$ is the unit tensor, and $\boldsymbol{\Omega}_{\mu\nu} = \boldsymbol{\Omega}(\boldsymbol{r}^*_\mu - \boldsymbol{r}^*_\nu)$ is the hydrodynamic interaction tensor. Defining the matrices $\boldsymbol{\mathcal{D}}$ and $\boldsymbol{\mathcal{B}}$ as block matrices consisting of $N \times N$ blocks each having dimensions of 3×3, with the (μ, ν)-th block of $\boldsymbol{\mathcal{D}}$ containing the components of the diffusion tensor $\boldsymbol{D}_{\mu\nu}$, and the corresponding block of $\boldsymbol{\mathcal{B}}$ being equal to $\boldsymbol{B}_{\mu\nu}$, the decomposition rule for obtaining $\boldsymbol{\mathcal{B}}$ can be expressed as $\boldsymbol{\mathcal{B}}\cdot\boldsymbol{\mathcal{B}}^{\mathrm{T}} = \boldsymbol{\mathcal{D}}$. The hydrodynamic tensor $\boldsymbol{\Omega}$ is assumed to be given by Rotne-Prager-Yamakawa (RPY) tensor

$$\boldsymbol{\Omega}(\boldsymbol{r}^*) = \Omega_1\boldsymbol{\delta} + \Omega_2\frac{\boldsymbol{r}^*\boldsymbol{r}^*}{r^{*2}}, \tag{4}$$

with

$$\Omega_1 = \begin{cases} \dfrac{3\sqrt{\pi}}{4}\dfrac{h^*}{r^*}\left(1 + \dfrac{2\pi}{3}\dfrac{h^{*2}}{r^{*2}}\right) & \text{for} \quad r^* \ge 2\sqrt{\pi}h^* \\ 1 - \dfrac{9}{32}\dfrac{r^*}{h^*\sqrt{\pi}} & \text{for} \quad r^* \le 2\sqrt{\pi}h^*, \end{cases}$$

and

$$\Omega_2 = \begin{cases} \dfrac{3\sqrt{\pi}}{4}\dfrac{h^*}{r^*}\left(1 - \dfrac{2\pi}{3}\dfrac{h^{*2}}{r^{*2}}\right) & \text{for} \quad r^* \ge 2\sqrt{\pi}h^* \\ \dfrac{3}{32}\dfrac{r^*}{h^*\sqrt{\pi}} & \text{for} \quad r^* \le 2\sqrt{\pi}h^*, \end{cases}$$

Here, the hydrodynamic interaction parameter h^* is the dimensionless bead radius in the bead-spring model and is defined by $h^* = a/(\sqrt{\pi k_B T/H})$.

Table 1: Definitions of shape functions in terms of eigenvalues of the gyration tensor, **G**. Note that, $I_1 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$, and $I_2 = \lambda_1^2\lambda_2^2 + \lambda_2^2\lambda_3^2 + \lambda_3^2\lambda_1^2$, are invariants of **G**.

Shape function	Definition
Asphericity (41, 43)	$B = \langle\lambda_3^2\rangle - \frac{1}{2}\left[\langle\lambda_1^2\rangle + \langle\lambda_2^2\rangle\right]$ (7)
Acylindricity (41, 43)	$C = \langle\lambda_2^2\rangle - \langle\lambda_1^2\rangle$ (8)
Degree of prolateness (38, 42, 43)	$S = \frac{\left\langle(3\lambda_1^2 - I_1)(3\lambda_2^2 - I_1)(3\lambda_3^2 - I_1)\right\rangle}{\left\langle(I_1)^3\right\rangle}$ (9)
Relative shape anisotropy (38, 41–43)	$\kappa^2 = 1 - 3\frac{\langle I_2\rangle}{\left\langle I_1^2\right\rangle}$ (10)

Since we are interested in the 3D organization of chromatin, we use a number of different static properties to describe the shape of the equilibrium chain. The radius of gyration of the chain, $R_g \equiv \sqrt{\langle R_g^2\rangle}$, where $\langle R_g^2\rangle$ is defined by

$$\langle R_g^2\rangle = \langle\lambda_1^2\rangle + \langle\lambda_2^2\rangle + \langle\lambda_3^2\rangle \tag{5}$$

with, λ_1^2, λ_2^2, and λ_3^2 being the eigenvalues of the gyration tensor **G** (arranged in ascending order), with

$$\mathbf{G} = \frac{1}{2N_b^2}\sum_{\mu=1}^{N_b}\sum_{\nu=1}^{N_b}\mathbf{r}_{\mu\nu}\mathbf{r}_{\mu\nu} \tag{6}$$

Note that, **G**, λ_1^2, λ_2^2, and λ_3^2 are calculated for each trajectory in the simulation before the ensemble averages are evaluated. The asymmetry in equilibrium chain shape has been studied previously in terms of various functions defined in terms of the eigenvalues of the gyration tensor (36–42). Apart from λ_1^2, λ_2^2, and λ_3^2, themselves, we have examined the following *shape functions*: the asphericity (B), the acylindricity (C), the degree of prolateness (S), and the shape anisotropy (κ^2), as defined in Table 1.

The stochastic differential equation (Eq. 3) can be solved with a semi-implicit predictor-corrector algorithm developed in Prabhakar and Prakash (44), once all the parameters are specified. However, the strength of interaction $\epsilon_{\mu\nu}$ between any two beads μ and ν is unknown *a priori*. Since they control the static conformations of a chain, their values will be different depending on whether the gene is in an 'ON' or 'OFF' state. Ultimately, the contact probability between any two segments on the gene is determined by the values of $\epsilon_{\mu\nu}$ for all pairs on the gene. Chromosome conformation capture experiments such as 5C and Hi-C provide information of the contact counts between different segments of chromatin, but not the interaction strengths that lead to these counts. Here we propose a methodology by which this inverse problem can be solved, namely, given the contact probabilities between all segment pairs, what is the pair-wise interaction strength that leads to this value? Additionally, the fact that experiments only give contact counts and not probabilities needs to be dealt with. In the section below, we first describe the inverse Brownian dynamics algorithm by which the interaction strengths $\epsilon_{\mu\nu}$ can be estimated given the set of contact probabilities $p_{\mu\nu}$. The issue of converting experimental contact counts to contact probabilities is addressed in section 3.2.

2.2 Inverse Brownian dynamics (IBD)

In the present investigation, a well-established standard method is utilized to optimise the parameters of a model Hamiltonian, such that it reproduces, as closely as possible, the values of some externally given quantities (e.g. from experiment or from other simulations). In the literature, the method is typically referred to as "Inverse Monte Carlo" (45–47). It is, however, completely independent of the underlying sampling scheme, as long as the latter produces thermal averages in the canonical ensemble. We prefer to highlight the underlying BD sampling of this study and hence refer to it here as the "Inverse Brownian Dynamics" (IBD) method. The method is best explained in general terms. It is assumed that the system is described by a phase

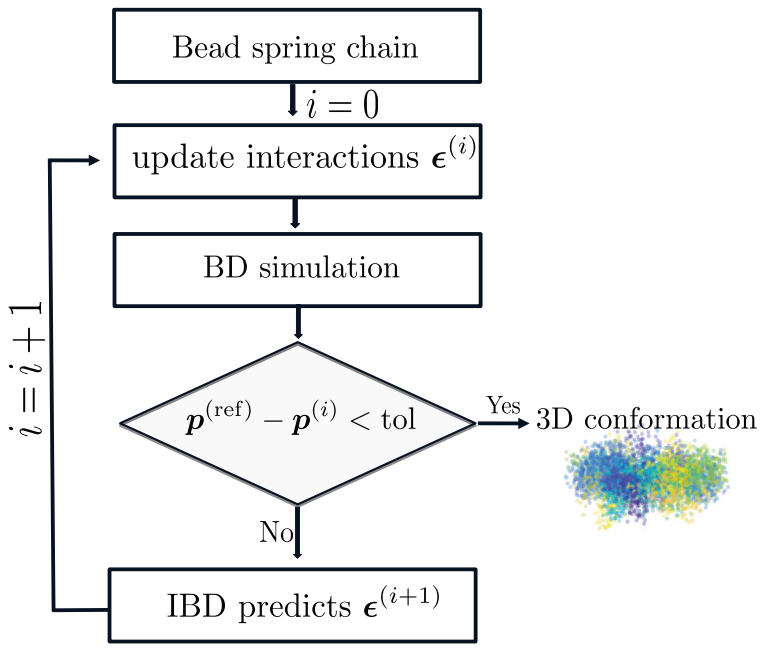


Figure 1: Flowchart for the Inverse Brownian Dynamics (IBD) method. Here $\boldsymbol{p}^{(\mathrm{ref})}$ represents the reference contact probability matrix, and $\boldsymbol{p}^{(i)}$ represents the contact probability matrix from simulations at iteration i. The interaction strength between beads μ and ν is given by $\boldsymbol{\epsilon}_{\mu\nu}$.

space variable Γ and a model Hamiltonian $\mathcal{H}(\Gamma)$. Another assumption is that the simulation produces the canonical average of some observable, given by a phase-space function $A(\Gamma)$:

$$\langle A \rangle = \frac{\int d\Gamma\, A(\Gamma)\, \exp(-\beta \mathcal{H}(\Gamma))}{\int d\Gamma \exp(-\beta (H)(\Gamma))}. \tag{11}$$

Here $\beta = 1/(k_B \mathrm{T})$. On the other hand, we have a given "target" value A_t (e.g. from experiment), which will typically differ from our simulation result. We are now interested in the dependence of the Hamiltonian on some coupling parameter J, and we wish to adjust J in order to bring $\langle A \rangle$ as closely to A_t as possible, within the limitations of the Hamiltonian as such in general, and its dependence on J in particular. In order to do this, it is desirable to obtain information on (i) in which direction J should modified, and (ii) by what amount (at least by order of magnitude). If the change of the coupling constant, ΔJ, is small, we can write down a Taylor expansion around the value $J = J_0$ where we performed the simulation:

$$\langle A \rangle (J_0 + \Delta J) = \langle A \rangle (J_0) + \chi \Delta J + O(\Delta J^2), \tag{12}$$

where the "generalized susceptibility" χ is an abbreviation for the thermodynamic derivative

$$\chi = \left. \frac{\partial \langle A \rangle}{\partial J} \right|_{J=J_0}. \tag{13}$$

The crucial point is now that χ can be directly sampled in the simulation, by making use of a standard fluctuation relation. Indeed, taking the derivative of Eq. 11 with respect to J, one finds directly

$$\chi = \beta \left[\langle AB \rangle - \langle A \rangle \langle B \rangle \right], \tag{14}$$

where B denotes another phase-space function, which is just the observable conjugate to J:

$$B(\Gamma) = -\frac{\partial \mathcal{H}(\Gamma)}{\partial J}. \tag{15}$$

In deriving Eq. 14, it is assumed that the phase-space function $A(\Gamma)$ does not depend on J, i.e. $\partial A(\Gamma)/\partial J = 0$. This is the case for most typical applications, and certainly for the present investigation.

The simplest way to do IBD, therefore, consists of (i) neglecting all nonlinear terms in Eq. 12, (ii) setting its left hand side equal to A_t , (iii) solving for ΔJ, and (iv) taking $J_0 + \Delta J$ as a new and improved coupling parameter, for which a new simulation is done, and for which the whole process is done again. To avoid overshoots, it is often advisable to not update J by the full increment ΔJ that results from solving the linear equation, but rather only by $\Delta J = \lambda \Delta J$, where λ is a damping factor with $0 < \lambda < 1$. The iteration is terminated as soon as $|\langle A \rangle| - A_t$ does not decrease any more, within some tolerance. One also has to stop as soon as χ becomes zero, within the statistical resolution of the simulation (this is, however, not a typical situation).

The method may be straightforwardly generalized to the case of several observables A_m and several coupling parameters J_n, where the number of observables and the number of couplings may be different. The Taylor expansion then reads

$$\langle A_m \rangle (\boldsymbol{J}_0 + \Delta \boldsymbol{J}) = \langle A_m \rangle (\boldsymbol{J}_0) + \sum_n \chi_{mn} \, \Delta J_n + O(\Delta \boldsymbol{J}^2), \tag{16}$$

where the matrix of susceptibilities is evaluated as a cross–correlation matrix:

$$\chi_{mn} = \beta \left[\langle A_m B_n \rangle - \langle A_m \rangle \langle B_n \rangle \right], \tag{17}$$

with

$$B_n(\Gamma) = -\frac{\partial \mathcal{H}(\Gamma)}{\partial J_n}. \tag{18}$$

Typically, the matrix χ_{mn} will not be invertible (in general, it is not even quadratic!). Therefore, one should treat the linear system of equations via a *singular-value decomposition* (SVD) and find ΔJ via the *pseudo-inverse* (PI). In practice, this means that one updates the couplings only in those directions and by those amounts where one has a clear indication from the data that one should do so, while all other components remain untouched. For details on the concepts of SVD and PI, the reader may refer to Press et al. (48) and Fill and Fishkind (49).

In the present instance, the averages $\langle A_m \rangle$ are the contact probabilities as produced by the simulations, while the target values are the corresponding experimental values (discussed in greater detail below). The corresponding phase–space functions can be written as indicator functions, which are one in case of a contact and zero otherwise. The coupling parameters that we wish to adjust are the well depths of the SDK attractive interactions, which we allow to be different for each monomer pair. The schematic representation of the IBD algorithm is provided as a Flowchart in Fig. 1. The IBD algorithm discussed here in general terms is described in more detail in the Supporting Material (SM) and applied to the specific problem considered here, along with a discussion of the appropriate SVD and PI.

3 RESULT AND DISCUSSION

3.1 Validation of the inverse Brownian dynamics method with a prototype

To validate the IBD method, a prototype of a chromatin-like polymer chain with artificially set interaction strengths ($\epsilon_{\mu\nu}$) was constructed. The data from this simulated chain was used to test the IBD algorithm, as described below. The IBD algorithm was validated for chains of length 10, 25 and 45 beads. Here we discuss the 45 bead chain case as a prototype. A few bead-pairs ($\mu\nu$) were connected arbitrarily with a prescribed value of the well-depth $\epsilon_{\mu\nu}^{(\mathrm{ref})}$ of the SDK potential. The non-zero reference interaction strengths for the connected bead-pairs $\epsilon_{\mu\nu}^{(\mathrm{ref})}$ are shown in Table 2; the remaining pairs were considered to have no attractive interaction ($\epsilon_{\mu\nu}^{(\mathrm{ref})} = 0$). The beads-spring chain was simulated until it reached equilibrium, which was quantified by computing R_g as a function of time. A stationary state was observed to be reached after eight Rouse relaxation times (50). However, equilibration was continued for fifteen Rouse relaxation times. After equilibration, an ensemble of 10^5 polymer configurations was collected from 100 independent trajectories, from each of which 10^3 samples were taken at intervals of 10^3 dimensionless time steps. From this ensemble, the contact probability $p_{\mu\nu}^{(\mathrm{ref})} = \langle \hat{p}_{\mu\nu} \rangle$ for each bead pair in the chain was computed. Here $\hat{p}_{\mu\nu}$ is an indicator function which is equal to 1 or 0 depending upon whether the μth and ν^{th} beads are within the cut-off distance of SDK potential ($r_{\mu\nu}^* \leq r_c^*$) or not ($r_{\mu\nu}^* > r_c^*$). The reference contact probabilities $p_{\mu\nu}^{(\mathrm{ref})}$, determined in this manner, are shown in Fig. 2(a). In the present instance while $p_{\mu\nu}^{(\mathrm{ref})}$ has been constructed by simulating the bead-spring chain for the given values of $\epsilon_{\mu\nu}^{(\mathrm{ref})}$, in general it refers to the experimental contact probabilities.

The IBD method was then applied to recover the reference contact probabilities $p_{\mu\nu}^{(\mathrm{ref})}$ starting with an initial guess $\epsilon_{\mu\nu}^{(0)} = 0$, i.e., all the interaction strengths are set equal to zero. As illustrated in Fig. 1, at each iteration step i, Brownian dynamics was

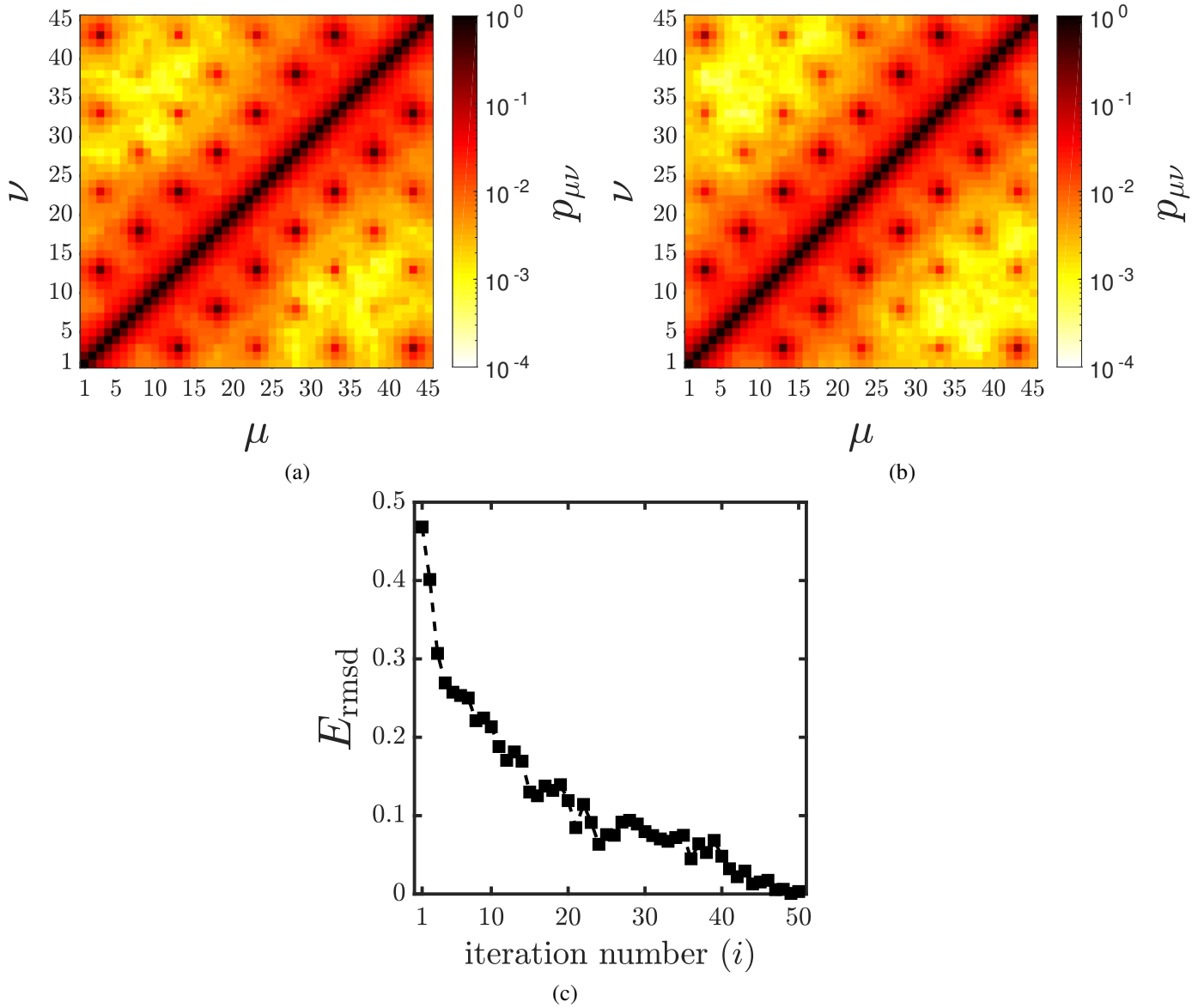


Figure 2: Validation of the IBD method with a prototype bead-spring chain with 45 beads. (a) Reference contact probability matrix. (b) Recovered contact probability matrix obtained using the IBD method (after 50 iterations, i.e., $i = 50$). The abscissa and ordinate represent the bead number along the polymer chain. The colour represents the contact probability between the beads μ and ν (see colorbar). (c) Root-mean-square deviation E_{rmsd} (Eq. 19) for each iteration.

performed for the given $\epsilon_{\mu\nu}^{(i)}$ and an ensemble of 10^5 conformations were collected. To quantify the difference between contact probabilities computed from simulation at iteration i ($p_{\mu\nu}^{(i)}$) and reference contact probabilities ($p_{\mu\nu}^{(\mathrm{ref})}$), the root mean-squared deviation $E_{\mathrm{rmsd}}^{(i)}$ was calculated

$$E_{\mathrm{rmsd}}^{(i)} = \sqrt{\frac{2}{N(N-1)} \sum_{1 \le \mu < \nu \le N} \left(p_{\mu\nu}^{(i)} - p_{\mu\nu}^{(\mathrm{ref})} \right)^2} \tag{19}$$

at each iteration. The error criteria $E_{\mathrm{rmsd}}^{(i)}$ has been used previously in Meluzzi and Arya (32), and is adopted here. At each iteration i, if the $E_{\mathrm{rmsd}}^{(i)}$ value is greater than the preset tolerance limit (tol), the interaction strength parameters $\epsilon_{\mu\nu}^{(i+1)}$ for the next iteration were calculated as given in Eq. S11 (see Supporting Material). To avoid the overshoot in interaction strength $\epsilon_{\mu\nu}^{(i+1)}$,

Table 2: Interaction strengths $\epsilon_{\mu\nu}$ and contact probabilities $p_{\mu\nu}$ for selected bead pairs (μ, ν) in a bead-spring chain with 45 beads. Values of these variables recovered using IBD are compared with those of the reference polymer chain, along with the percentage error between the reference and recovered values.

Chain length, N = 45						
bead-pair	interaction strength, $\epsilon_{\mu\nu}$			contact probability, $p_{\mu\nu}$		
	reference	recovered	% error	reference	recovered	% error
3-13	7.00	6.70	4.29	0.44	0.46	4.55
13-23	7.00	7.28	4.00	0.51	0.49	3.92
23-33	7.00	7.08	1.14	0.39	0.37	5.13
33-43	7.00	7.35	5.00	0.62	0.59	4.84
8-18	7.00	6.94	0.86	0.47	0.47	0.00
18-28	7.00	6.89	1.57	0.31	0.32	3.23
28-38	7.00	7.16	2.29	0.55	0.53	3.64
3-43	7.00	7.18	2.57	0.22	0.22	0.00

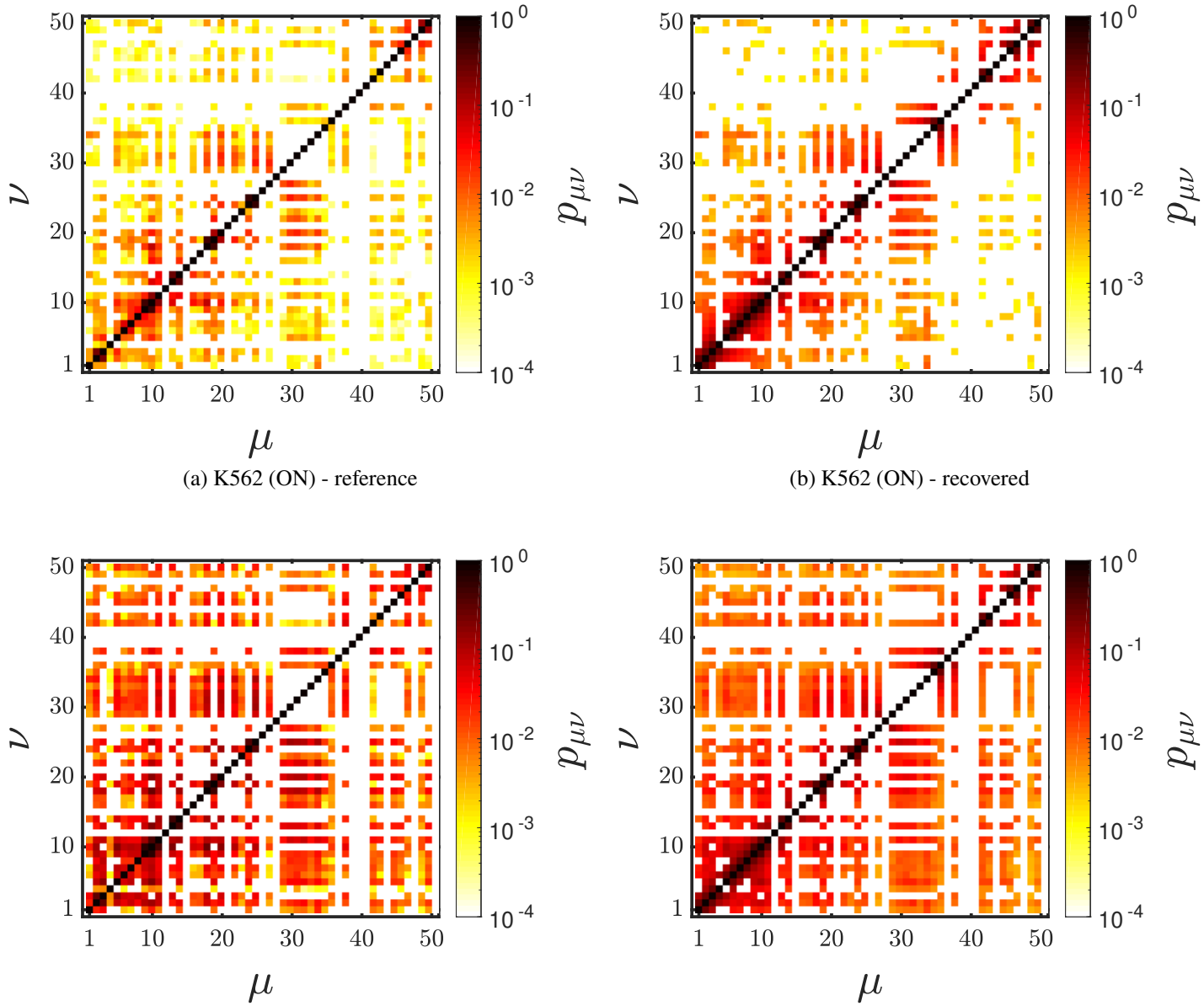
the range of $\epsilon_{\mu\nu}^{(i+1)}$ was constrained to [0, 10]. For the investigated polymer chain with 45 beads, the IBD algorithm converges ($E_{\text{rmsd}}^{(i)} < \text{tol}$) in approximately 50 iterations and $p_{\mu\nu}^{\text{ref}}$ was recovered. The recovered contact probability matrix and E_{rmsd} for each iteration are shown in Figs. 2(b) and 2(c), respectively. The recovered contact probability values along with the optimized interaction strengths $\epsilon_{\mu\nu}$ are shown in Table. 2. The error in the recovered contact probabilities and interaction strengths is less than 5%, proving the reliability of the IBD method. The largest contact probabilities are for those bead-pairs for which values of the interaction strength were chosen a priori, as given in Table 2. However, the existence of these interactions leads to the existence of contact probabilities $p_{\mu\nu}$ between all bead-pairs μ and ν. The IBD algorithm was applied to not just the specified bead-pairs but to recover all contact probabilities $p_{\mu\nu}$, for all possible pairs. The errors are given in Table 2 only for the specified values since they are the largest, the error in all the remaining contact probabilities was less than 5%. The use of singular-value decomposition and the pseudo-inverse method in order to carry out the IBD procedure are discussed in the Supporting Material. Having validated the IBD algorithm, the next section applies this technique to experimentally obtained contact probabilities of a chromatin, on the length scale of a gene.

3.2 Conversion of contact counts to contact probabilities: the normalization problem

To study the 3D organization of a gene region, the α-globin gene locus (ENCODE region ENm008) is chosen for which Bau et al. (29) have experimentally determined the contact counts using the 5C technique. This is a 500 kbp long region on human chromosome 16 containing the α-globin gene and a few other genes like LUC7L. Since 5C data does not interrogate the contact counts between all feasible 10 kbp segment pairs, many elements in the heat map have no information. This is in contrast with typical Hi-C experiments where information on all possible contact pairs are obtained. In principle, this method can be applied to Hi-C data; however, in this instance, we chose the 5C data since it has sufficiently good resolution.

For simulation purpose, the α-globin locus is coarse-grained to a bead-spring chain of 50 beads. That is, the experimental 5C data (contact count matrix of size 70×70) for the EMn008 region was converted to a contact count matrix of size 50×50. The coarse graining procedure is as follows: 500kb of the gene locus was divided into 50 beads, each comprising 10 kb equal-sized fragments. The midpoint of each restriction fragment was located and was assigned to the corresponding bead in the coarse-grained polymer. There are cases where two or more restriction fragments (each of size less than 10kbp) get mapped to the same bead. For example, consider restriction fragments r_1 and r_2 being mapped on to a single coarse-grained bead μ, and fragments r_3 and r_4 being mapped on to another bead ν. The contact counts of the coarse-grained bead-pair $C_{\mu\nu}$ can then be computed in at least three different ways. (i) Take the sum of all contact counts for the four restriction fragment combinations ($C_{\mu\nu} = C_{r_1r_3} + C_{r_1r_4} + C_{r_2r_3} + C_{r_2r_4}$) — i.e., assume that all contacts occur independently of each other, in other words not more than one of the contact pairs occurs in the same cell. This is termed here as the *independent* coarse graining procedure. (ii) Take the maximum contact count amongst all the four restriction fragment combinations ($C_{\mu\nu} = \max\{C_{r_1r_3}, C_{r_1r_4}, C_{r_2r_3}, C_{r_2r_4}\}$). This assumes that whenever the pairs having small contact counts are in contact, the pair with the largest contact count is also in contact. This is termed here as the *dependent* coarse graining procedure. These are the two extreme cases and the reality could be somewhere in between. The third option is then to choose some such intermediate value. (iii) Here, we use the approximation that the coarse grained contact count is equal to the average of the two extreme contact counts mentioned earlier, namely $C_{\mu\nu} = \frac{1}{2}[(C_{r_1r_3} + C_{r_1r_4} + C_{r_2r_3} + C_{r_2r_4}) + \max\{C_{r_1r_3}, C_{r_1r_4}, C_{r_2r_3}, C_{r_2r_4}\}]$.

The contact counts obtained from the Chromosome Conformation Capture experiments are not normalized. That is, the



(a) K562 (ON) - reference

(b) K562 (ON) - recovered

(c) GM12878 (OFF) - reference

(d) GM12878 (OFF) - recovered

Figure 3: Comparison of the reference normalized contact probabilities ((a) and (c)) with the recovered contact probabilities ((b) and (d)), obtained with the IBD method for K562 and GM12878, respectively, at $N_f = 0$.

contact count values can vary from experiment to experiment and total number of contacts are not quantified. This data cannot be compared across cell lines or across different experimental sets. The contact counts need to be converted to normalized contact probabilities ranging from 0 to 1 (contact probability = (contact count)/(total number of samples)). This would then make the data suitable for comparison with simulation. Since the total number of genome equivalent (number of cells) cannot be estimated in a chromosome conformation capture experiment, the calculation of contact probability from the contact count is highly challenging. A novel technique to normalize these counts is described here. The contact count matrix can be normalized by imposing the following constraint, namely, that the sum of times any segment pairs (μ, ν) are in contact ($C^{c}_{\mu\nu}$) and the number of times they are not in contact ($C^{nc}_{\mu\nu}$) must be equal to the total number of samples N_s. This is true for all bead-pairs i.e. $C^{c}_{\mu\nu} + C^{nc}_{\mu\nu} = N_s$, for all $\mu\ \nu$. Since only $C^{c}_{\mu\nu}$ is known, two limiting values of N_s are estimated using the following scenarios. In one scenario, it is assumed that for the segment pairs (μ, ν) which has the largest contact count in the matrix, μ and ν are always in contact in all cells. In other words $C^{nc}_{\mu\nu} = 0$; in this case N_s is simply equal to the largest element of the contact

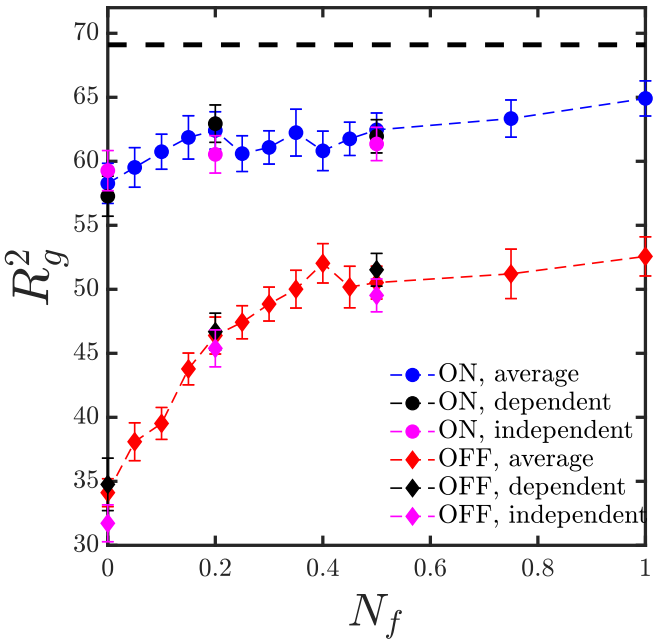


Figure 4: Spatial extension of the polymer chain, quantified by the radius of gyration, R_g^2, computed at various values of N_f, for both K562 (ON state) and GM12878 (OFF state) cell lines. All three coarse-graining techniques, i.e., dependent, independent and average, have been used. The dashed line represents the value of R_g^2 for a chain executing self-avoiding walk statistics.

count matrix. Since this is the smallest value of N_s possible, it is denoted by $(C^c_{\mu\nu})_{\max} = N_{\min}$. The other scenario estimates the sample size from the row μ for which the sum over all contact counts is the largest i.e., N_s = maximum of $(\sum_\nu C^c_{\mu\nu})$. This assumes that μ is always in contact with only one other segment in a cell and there is no situation when it is not in contact with any segment. This case is denoted as $N_{\max}$. Since the precise value of N_s is not known, N_s is varied as a parameter from $N_{\min}$ to $N_{\max}$. To systematically vary N_s, for convenience, a parameter N_f is defined,

$$N_f = \frac{N_s - N_{\min}}{N_{\max} - N_{\min}} \tag{20}$$

in the range of $[0, 1]$. Clearly, $N_f = 0$ implies $N_s = N_{\min}$, which is the lower bound for N_s and $N_f = 1$ implies $N_s = N_{\max}$, which is the upper bound. The contact probabilities at various N_f values are calculated as $p_{\mu\nu} = (C^c_{\mu\nu}/N_s)$ where $N_s = N_{\min} + N_f(N_{\max} - N_{\min})$.

For several values of N_f, the contact count matrices are normalized and IBD is carried out to obtain the optimal interaction strengths between the bead-pairs. Fig. 3(a) and 3(c) show the normalized contact probabilities at $N_f = 0$ for cell lines K562 (ON state) and GM12878 (OFF state), respectively (reference contact probabilities), when they are coarse-grained to 50 segments of length 10 kbp each, as per the procedure described above. The corresponding recovered contact probability matrices for both the cell lines from simulation are shown in Fig. 3(b) and 3(d). In both cases, the maximum percentage error between the reference and recovered contact probability of any segment pairs was found to be less than 5%.

The spatial extent of the chromatin polymer, as quantified by the square radius of gyration R_g^2, for different values of N_f is presented in Fig. 4. In the case of the cell line where the gene is ON (K562), the increase in R_g^2 for small values of N_f is relatively less prominent and becomes nearly independent of N_f as N_f approaches one. It is clear that contact probabilities decrease with increasing N_f, since N_s increases with N_f. It is consequently expected that with sufficiently large N_f, R_g^2 should approach the value for a self-avoiding walk (dashed line in Fig. 4). This can be seen to be the case when the gene is the ON (K562) state. In the cell line where the gene is OFF (GM12878), the value of R_g^2 increases relatively rapidly for small values of N_f and reaches a nearly constant value for $N_f \gtrsim 0.4$. However, the limiting value is significantly smaller than that of a self-avoiding walk. This suggests that some significant interactions are still present amongst the bead-pairs, even for N_f approaching one. The influence of the different coarse-graining procedures was examined and it was found that the value of

Table 3: Various shape property based on the eigen values of gyration tensor $\boldsymbol{G}$ are defined here for $N_f = 0$.

Shape property	K562 (ON state)	GM12878 (OFF state)
$\langle\lambda_1^2\rangle/R_g^2$	0.058	0.081
$\langle\lambda_2^2\rangle/R_g^2$	0.164	0.201
$\langle\lambda_3^2\rangle/R_g^2$	0.778	0.718
$\langle\lambda_2^2\rangle/\langle\lambda_1^2\rangle$	2.828	2.479
$\langle\lambda_3^2\rangle/\langle\lambda_1^2\rangle$	13.412	8.874
B/R_g^2	0.667	0.578
C/R_g^2	0.106	0.120
S	0.913	0.772
κ^2	0.545	0.452

R_g^2 from all the three coarse-graining procedures agreed with each other within error bars (as seen from the data at $N_f = 0, 0.2$ and 0.5, for both the cell lines). This suggests that, at least as far as R_g^2 is concerned, the choice of coarse-graining method is not vitally important.

3.3 Three-dimensional configuration of the α-globin gene locus

3.3.1 Shape functions

Since chromatin folded in 3D can have spatial organization that is beyond simple spherically symmetric packing, various non-globular 3D shape properties (as described in section 2.1) have been analysed here.

Eigenvalues of the radius of gyration tensor for polymer chains are usually reported in terms of ratios, either between individual eigenvalues, or with the mean square radius of gyration. For a chain with a spherically symmetric shape about the centre of mass, we expect $\langle\lambda_i^2\rangle/\langle r_g^2\rangle = 1/3$, for $i = 1, 2, 3$, and $\langle\lambda_i^2\rangle/\langle\lambda_j^2\rangle = 1$ for all combinations i and j. For chain shapes with tetrahedral or greater symmetry, the asphericity $B = 0$, otherwise $B > 0$. For chain shapes with cylindrical symmetry, the acylindricity $C = 0$, otherwise $C > 0$. With regard to the degree of prolateness, its sign determines whether chain shapes are preponderantly oblate ($S \in [-0.25, 0]$) or prolate ($S, \in [0, 2]$). The relative anisotropy (κ^2), on the other hand, lies between 0 (for spheres) and 1 (for rods).

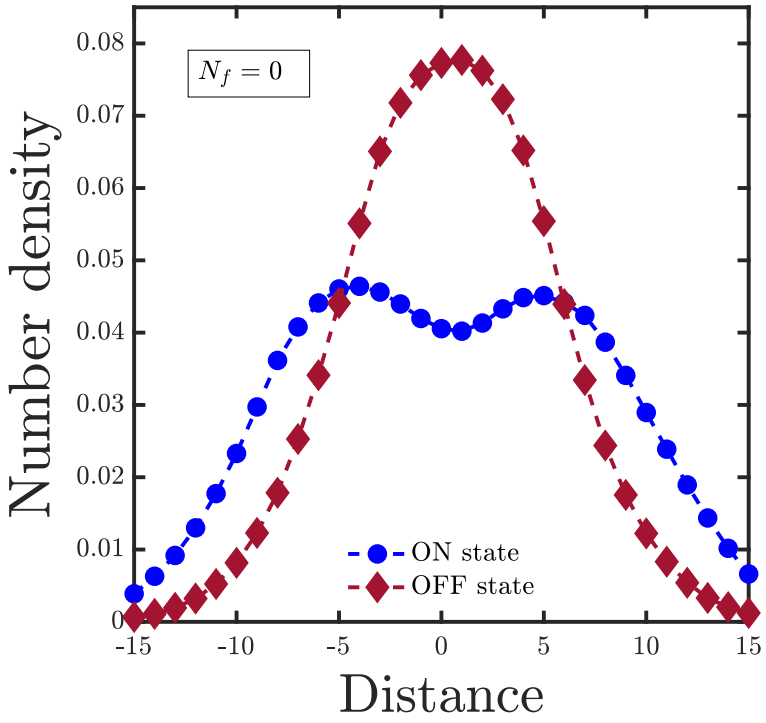
All these properties are investigated for $N_f = 0$, and compared in the ON and OFF states, as displayed in table 3. It is clear that the while the chain is highly non-spherical in both states, it appears to be slightly more spherical in the OFF than in the ON state.

3.3.2 Density profiles

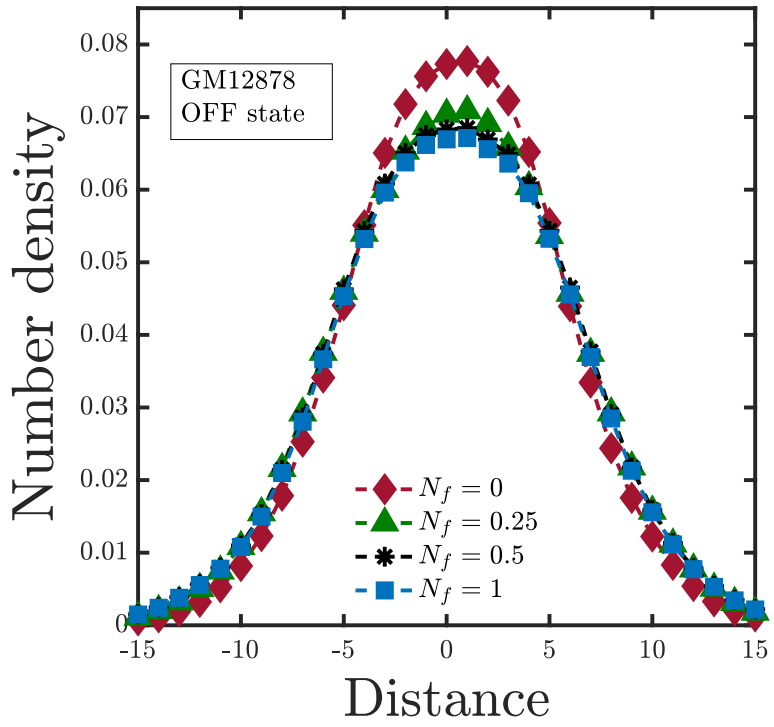
To get a different prospective on the 3D organization of the gene, the density distribution about the centre of mass was considered. In order to do this, all polymer configurations were aligned along the major axis of the radius of gyration tensor **G** and each bead position was binned and the number density of beads along the major axis was computed. As displayed in Fig. 5(a), in GM12878 (OFF state) cells, the number density shows a single peak at the center of mass position suggesting a symmetric organization around the centre of mass along the major axis. In the case of K562 (ON state) cells, the number density is seen to have a double peak, implying a bi-modal distribution of polymer beads around the centre of mass along the major axis (Fig. 5(a)), as suggested by earlier 3D models for the α-globin gene (29, 31). With an increase in N_f, a slight decrease in the number density at the core of the α-globin gene in the OFF state is observed (Fig. 5(b)), while a decrease in extent of bimodality is observed in the ON state (Fig. 5(c)). However, the differences for different N_f values are less prominent at the peripherial regions of the globule. In Fig. 5(d) the number density for different coarse-graning procedures used here is presented for the case $N_f = 0$. As can be seen, similarly to what was observed earlier for Rg^2, the different coarse-graining procedures give results that are close to each other.

3.3.3 3D confirmations

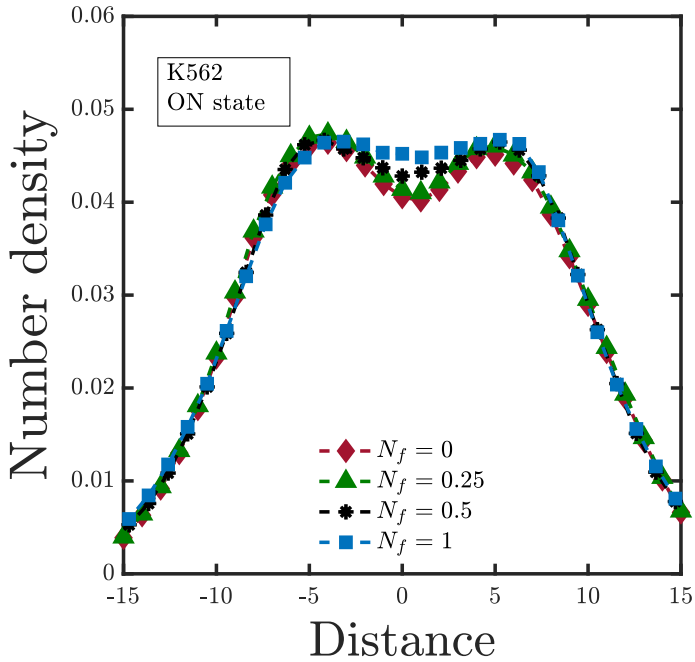
To obtain a snapshot of the 3D structure of the α-globin gene locus, different configurations from the ensemble were superimposed on top of each other, as displayed in Fig. 6, for both the cell lines at different values of N_f. Each dot represents a bead and to make them visible, they have been made transparent to some degree. Different colors in the plot represent the bead number along the contour length of the polymer chain. As indicated from the shape functions and the density profiles, the


$N_f = 0$
ON state
OFF state
Number density
Distance
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-15
-10
-5
0
5
10
15

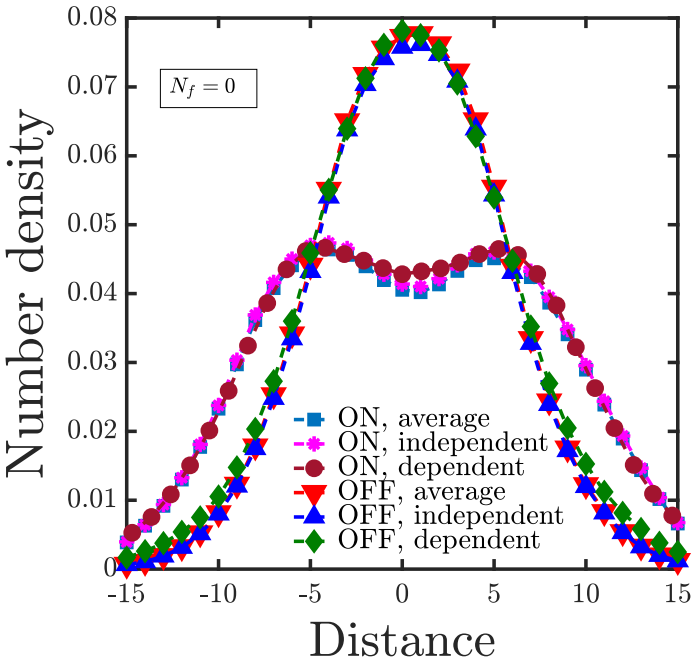
(a)


GM12878
OFF state
$N_f = 0$
$N_f = 0.25$
$N_f = 0.5$
$N_f = 1$
Number density
Distance
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-15
-10
-5
0
5
10
15

(b)



(c)



(d)

Figure 5: Comparison of the number density of beads along the major axis of the radius of gyration tensor, for: (a) ON and OFF states at $N_f = 0$, (b) the OFF state at different values of N_f, (c) the ON state for values of N_f, and (d) different coarse-graining procedures (i.e., dependent, independent and average) at $N_f = 0$. The center of mass is shifted to the origin for simplicity.

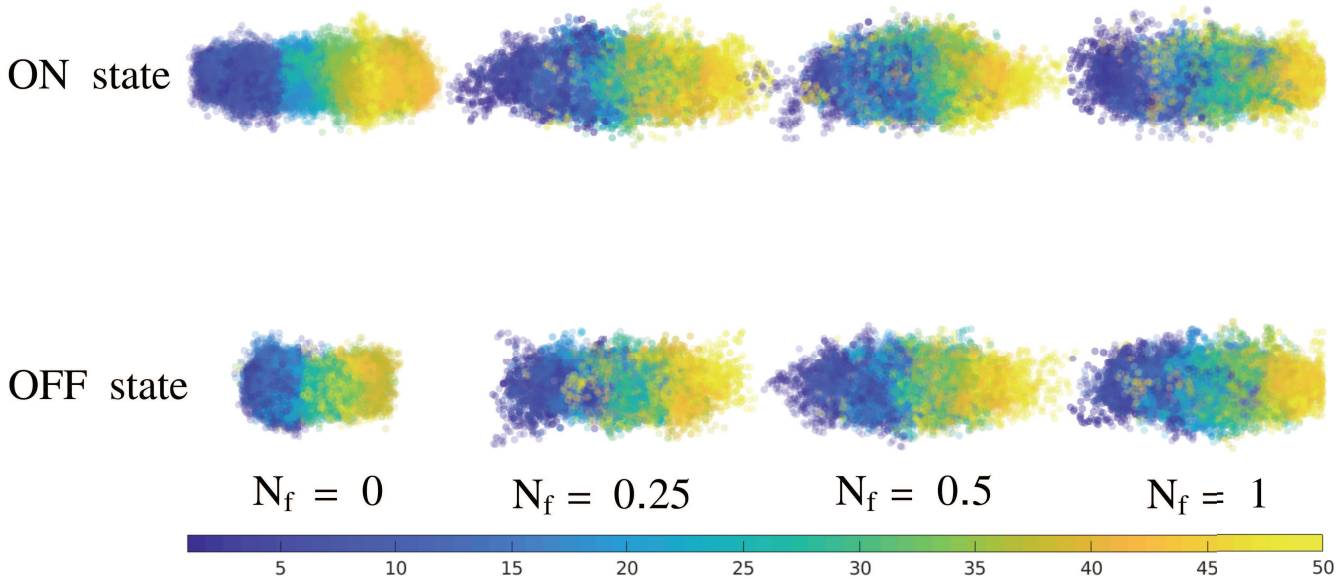


Figure 6: Snapshots of 3D configurations, obtained by aligning chains along the major axis of the radius of gyration tensor and superimposing them on top of each other with transparency. Different values of N_f are displayed for cell lines K562 and GM12878. The colour assigned to each marker (blue to yellow) represents the bead number along the contour length (bead 1 to bead 50) of the polymer chain.

snapshot shows that the structure is highly non-spherical in both cases, In particular, the K562 (ON state) cell line chromatin has a more extended configuration, with slightly higher density away from the centre of mass.

3.4 The 3D distances and contact probabilities

The 3D conformation of the α-globin gene locus has been investigated earlier (29, 30). These studies differ from the current work in some important aspects. Firstly, they assume that the contact counts between any two pairs can be converted to an equilibrium distance between those pairs through a certain pre-determined functional form. Secondly, instead of optimizing the interaction strengths to recover the contact counts, their simulations attempt to recover the equilibrium distances that have been derived from contact matrices. It is not clear in these cases whether the experimentally observed contact counts will be recovered by simulations. In this work, no assumptions have been made about the relationship between spatial distance and contact probability for any pair of beads. On the contrary in the present case, we can compute the spatial distances ($d_{\mu\nu}$) that are consistent with the contact probability matrix. Further, no configuration from the ensemble is discarded.

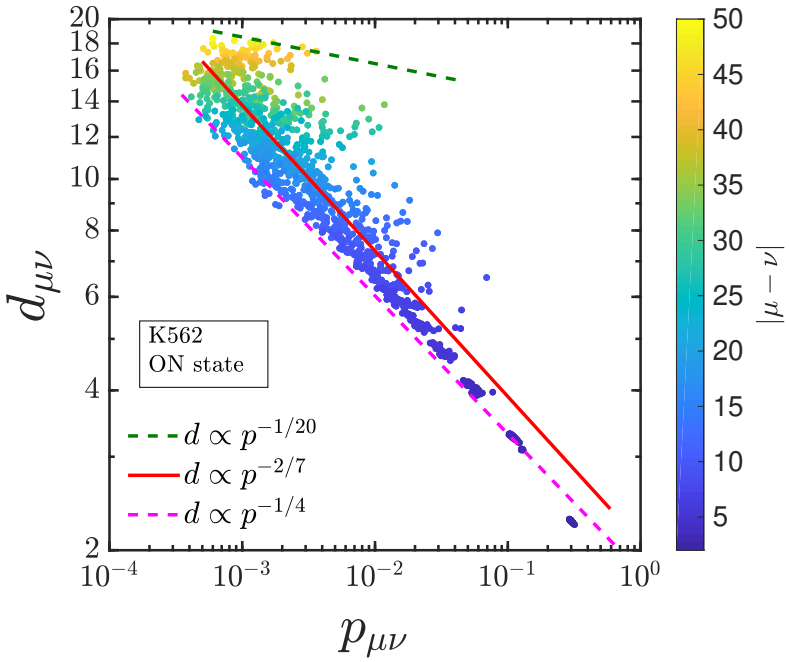
The spatial distances calculated in the current work for the contact probabilities in the ON and OFF state are shown in Fig. 7(a) for K562 (ON state) and in Fig. 7(b) for GM12878 (OFF state) cell lines. Each point in these figures represents the ensemble-averaged 3D distance between a given pair of beads (y-axis) having a contact probability as indicated in the x-axis. As is immediately apparent, a wide range of 3D distances is possible, unlike what was assumed in earlier studies. It appears that the average 3D distance is not just a function of contact probability $p_{\mu\nu}$ (where the interaction between the beads plays a role), but is also a function of the distance along the contour between the beads ($|\mu - \nu|$) – the color variation in Figs. 7(a) and (b) indicates the influence of contour length. The red line in both the figures are fitted power-laws to the data, drawn as a guide to the eye. In both cases, the exponents are close to $-1/4$. But the interesting element here is the variability (scatter) in the data which shows that for a given contact probability value, there can be multiple values of 3D distances, with deviation of many units. From the colors that represent the bead-pair distance along the contour ($|\mu - \nu|$), it is clear that the mean 3D distance is a function of both the contact probability and contour distance $|\mu - \nu|$. It can also be seen that for beads far-apart along the contour, the 3D distances are nearly independent of the contact probabilities while for nearby beads, the average distance is strongly dependent on the contact probability.

To understand this variability better, we bin the same data and plot it as violin plots that display the mean 3D distance for a given small range of contact probabilities, as shown in Figs. 8 (a) and (b). It is clear that the distribution of points around the mean is very diverse – bimodal in a few cases and with an extended tail in many cases – suggesting that a simple functional form between the mean 3D distance and the contact probability may not be feasible. It must be reiterated here that many previous studies have assumed power law relations such as $d_{\mu\nu} \propto p_{\mu\nu}^{\tau}$ can be used, with exponents $\tau = -1$ (25, 26) and $\tau = -1/2$ (28), independent of $|\mu - \nu|$. Some groups have also assumed exponential (27) and logarithmic decay of distance with probability (29). As shown above, the results reported here do not support the usage of such simple functional forms.

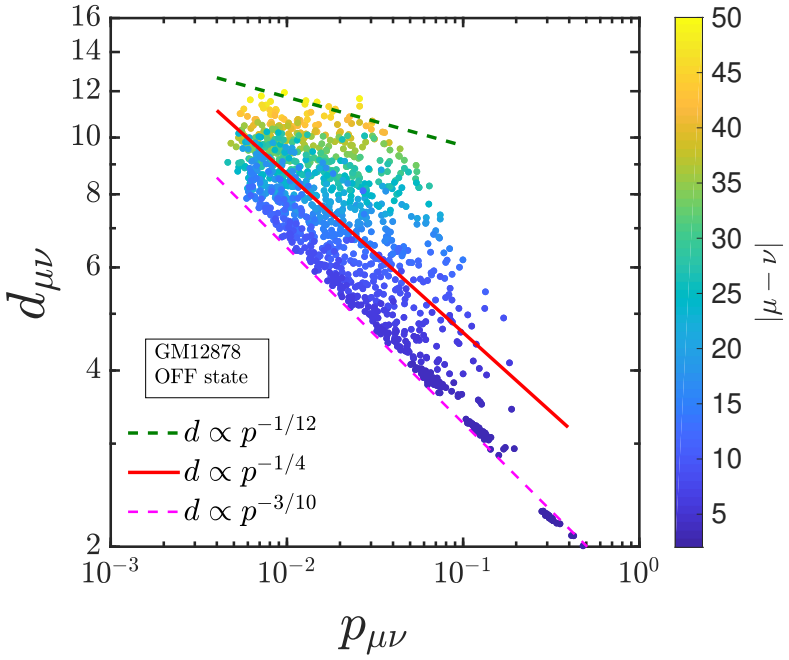
4 CONCLUSION

The 3-dimensional organisation of chromatin based on publicly available chromatin conformation capture experimental data has been investigated. Unlike many existing models, the current work treats this as an inverse problem where interactions between different chromatin segments are computed such that the experimentally known contact probabilities are reproduced. A polymer model and an Inverse Brownian dynamics (IBD) alogirthm has been developed for this problem which has the following advantages: (i) it does not assume any *a priori* relation between spatial distance and contact probability, (ii) it optimizes the interaction strength between the monomers of the polymer chain in order to reproduce the target contact probability, and (iii) since hydrodynamics interactions are included, it is capable of investigating the dynamics of the chromatin polymer.

The main results of this work are as follows: (i) The IBD method was validated for a bead-spring chain comprising of 45 beads. It was observed that IBD reproduced the contact probability and the interaction strength (within 5% of error), reflecting its reliability. (ii) Three different coarse-graining procedures – *independent*, *dependent* and *average* were used to map between the experimental and coarse-grained contact matrices. For the gene locus studied in this work (α-globin gene), no significant differences between the three cases was observed, both for the gene extension and the density profile. (iii) A procedure for normalizing the contact count matrix was introduced with a parameter N_f varying from 0 to 1, that reflected the two different extreme scenarios for estimating the sample size. For the GM12878 (OFF state), the gene extension increases rapidly initially with increasing N_f, while for the K562 (ON state) on the other hand (which is already in an extended state), there is a very little scope for further extension with increasing N_f. (iv) Since there is a relationship between the normalization factor N_f and physically measured properties such as the radius of gyration, it is conceivable that the value of N_f can be estimated from experiments such as FISH, Chip-seq, etc. (v) The structural properties of the α-globin gene locus were investigated in terms of shape functions, bead number density distributions, and 3D snapshots. In the ON state (K562), α-globin appears to lack any



(a)



(b)

Figure 7: Dependence of mean 3D distances $d_{\mu\nu}$ on contact probabilities $p_{\mu\nu}$ for (a) K562 (ON state) and (b) GM12878 (OFF state) cell lines, respectively. For the K562 (ON state) cell line, the contact probabilities are bounded by power laws, $d_{\mu\nu} \propto p_{\mu\nu}^{\tau}$, where τ varies from $-1/20$ (upper bound) to $-1/4$ (lower bound). Similarly, in the GM12878 (OFF state), τ varies from $-1/12$ to $-3/10$.

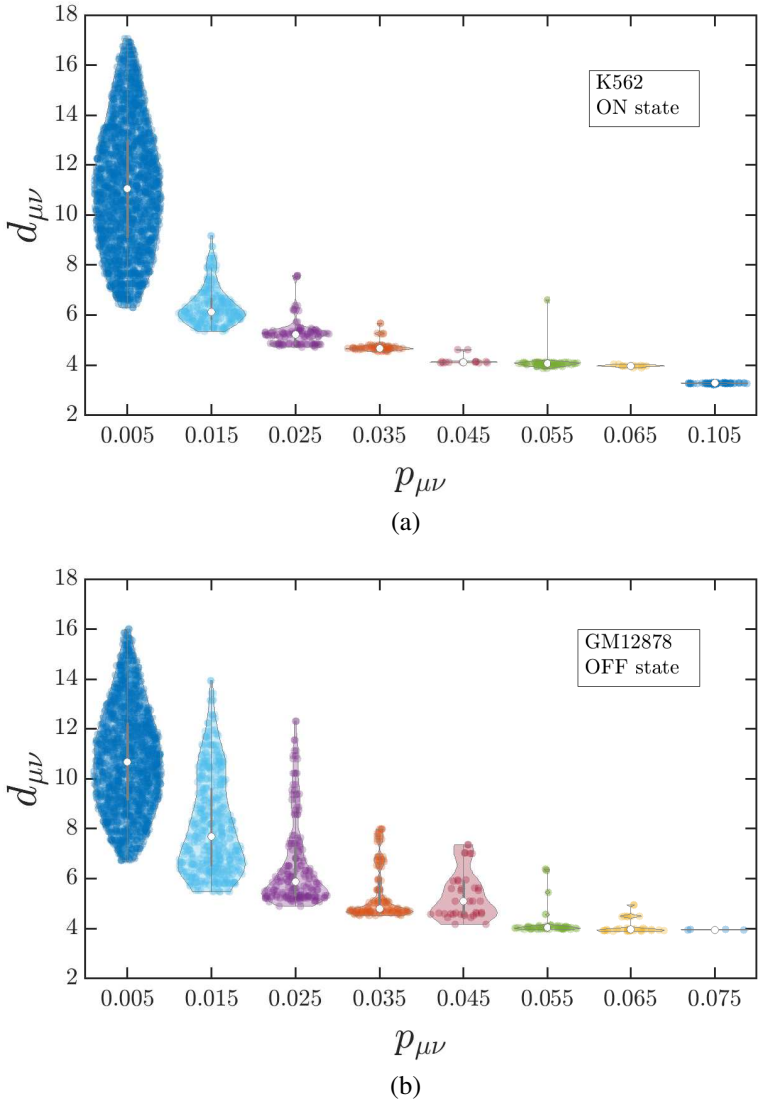


Figure 8: Violin plots which display the probability distribution of mean 3D distances for selected ranges of contact probabilities in (a) the K562 (ON state) cell lines, and (b) the GM12878 (OFF state) cell lines.

prominent interactions, and exists in an extended structure. Whereas in the case of GM12878 (OFF state), the gene appears to be in a folded state. This is also consistent with theory, as in the ON state (K562) the transcription factors need to access the gene, while the structural status of the OFF state (GM12878) should be to avert the transcription factor, resulting in gene silencing. (vi) The density profile along the major axis of the radius of gyration tensor also supports the extended structure in cell line K562 (ON state) and a sharp cluster of monomers at the core of GM12878 (OFF state). (vii) The dependency of spatial distance on contact probability has been investigated, and it is shown that the usage of simple functional relationships may not be realistic.

Most of the results in this work are predictions that may be tested in suitably designed experiments. We predict that the spatial segmental distance is not only dependent on the contact probability but also on the segment length along the contour. One of the ways to test our prediction is to perform 3D FISH on segment-pairs having the same contact probability but different segment length. A difference in distance obtained from the FISH experiment will validate the predictions made in the current work. Shape properties of the α-globin are also predicted and can be tested using techniques like super-resolution microscopy.

Since the model has dynamics, with hydrodynamics interactions built in, it has the potential to be used to address problems in the future involving dynamics of the 3D chromatin polymer between different chromatin states. Currently in this model only chromatin configuration capture data has been considered. However, the model may be extended to incorporate more data (histone modification data, CHIP-Seq data of certain proteins) and address chromatin organization on the length scale of genes in more detail. With the capability of analysing the 3D configuration along with chromatin dynamics, IBD can complement experimental research and also provide deeper and useful insights based on the same.

AUTHOR CONTRIBUTIONS

K.K., B.D., R.P., and J.R.P. designed the research. K.K. wrote code, carried out simulations and analysed the data. K.K., B.D., R.P., and J.R.P. wrote the manuscript.

ACKNOWLEDGEMENTS

We appreciate the funding and support from the IITB-Monash Research Academy, and from SERB, DST India via grant number EMR/2016/005965. We gratefully acknowledge the computational resources resources provided at the NCI National Facility systems at the Australian National University through the National Computational Merit Allocation Scheme supported by the Australian Government, the MonARCH facility maintained by Monash University, and by IIT Bombay.

REFERENCES

1. Alberts, B., 2014. Molecular Biology of The Cell. Garland Science, Taylor and Francis Group, New York, 6 edition.
2. Ecker, J. R., W. A. Bickmore, I. Barroso, J. K. Pritchard, Y. Gilad, and E. Segal, 2012. Genomics: ENCODE explained. *Nature* 489:52.
3. Larson, A. G., and G. J. Narlikar, 2018. The role of phase separation in heterochromatin formation, function, and regulation. *Biochemistry* 57:2540–2548.
4. Gilbert, N., S. Gilchrist, and W. A. Bickmore, 2005. Chromatin organization in the mammalian nucleus. *International review of cytology* 242:283–336.
5. Fraser, P., and W. Bickmore, 2007. Nuclear organization of the genome and the potential for gene regulation. *Nature* 447:413.
6. Bickmore, W. A., 2013. The spatial organization of the human genome. *Annual review of genomics and human genetics* 14:67–84.
7. Dekker, J., K. Rippe, M. Dekker, and N. Kleckner, 2002. Capturing Chromosome Conformation. *Science* 295:1306–1311.
8. Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat, 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* 38:1348–1354.
9. Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker, 2006. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 1299–1309.

10. Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326:289–293.

11. Nora, E. P., B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen, J. Dekker, and E. Heard, 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485:381–385.

12. Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485:376–380.

13. Nora, E. P., J. Dekker, and E. Heard, 2013. Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? *BioEssays* 35:818–828.

14. Rao, S. S., 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, and E. L. Aiden, 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping . *Cell* 159:1665 – 1680.

15. Schiessel, H., 2003. The physics of chromatin. *Journal of Physics: Condensed Matter* 15:R699–R774.

16. Teif, V. B., and K. Bohinc, 2011. Condensed DNA: condensing the concepts. *Progress in biophysics and molecular biology* 105:208–222.

17. Ganai, N., S. Sengupta, and G. I. Menon, 2014. Chromosome positioning from activity-based segregation. *Nucleic acids research* 42:4145–4159.

18. Bascom, G. D., C. G. Myers, and T. Schlick, 2019. Mesoscale modeling reveals formation of an epigenetically driven HOXC gene hub. *Proc. Natl. Acad. Sci. U. S. A.* 116:4955–4962.

19. Dans, P. D., J. Walther, H. Gómez, and M. Orozco, 2016. Multiscale simulation of DNA. *Current opinion in structural biology* 37:29–45.

20. Gehlen, L. R., G. Gruenert, M. B. Jones, C. D. Rodley, J. Langowski, and J. M. O'Sullivan, 2012. Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions. *Nucleus (Austin, Tex.)* 3:370–383.

21. Yan, J., R. Kawamura, and J. F. Marko, 2005. Statistics of loop formation along double helix DNAs. *Physical Review E* 71:061905.

22. Di Pierro, M., B. Zhang, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, 2016. Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences of the United States of America* 113:12168–12173.

23. Mirny, L. A., 2011. The fractal globule as a model of chromatin architecture in the cell. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* 19:37–51.

24. Bancaud, A., C. Lavelle, S. Huet, and J. Ellenberg, 2012. A fractal model for nuclear organization: current evidence and biological implications. *Nucleic acids research* 40:8783–8792.

25. Fraser, J., M. Rousseau, S. Shenker, M. A. Ferraiuolo, Y. Hayashizaki, M. Blanchette, and J. Dostie, 2009. Chromatin conformation signatures of cellular differentiation. *Genome Biol.* 10:R37.

26. Duan, Z., M. Andronescu, K. Schutz, S. McIlwain, S. Mcllwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and S. William, 2010. A Three-Dimensional Model of the Yeast Genome. *Nature* 465:363–367.

27. Tanizawa, H., O. Iwasaki, A. Tanaka, J. R. Capizzi, P. Wickramasinghe, M. Lee, Z. Fu, and K. I. Noma, 2010. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research* 38:8164–8177.

28. Rousseau, M., J. Fraser, M. A. Ferraiuolo, J. Dostie, and M. Blanchette, 2011. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC bioinformatics* 12:414.

29. Bau, D., A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom, 2011. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology* 18:107–114.

30. Paulsen, J., M. Sekelja, A. R. Oldenburg, A. Barateau, N. Briand, E. Delbarre, A. Shah, A. L. Sørensen, C. Vigouroux, B. Buendia, and P. Collas, 2017. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biology* 18:21.

31. Paulsen, J., T. M. Liyakat Ali, and P. Collas, 2018. Computational 3D genome modeling using Chrom3D. *Nature Protocols* 13:1137.

32. Meluzzi, D., and G. Arya, 2013. Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic Acids Res.* 41:63–75.

33. Soddemann, T., B. Dünweg, and K. Kremer, 2001. A generic computer model for amphiphilic systems. *The European Physical Journal E* 6:409–419.

34. Santra, A., K. Kumari, R. Padinhateeri, B. Duenweg, and J. R. Prakash, Under review. Universality of the collapse transition of sticky polymers. *Soft Matter* .

35. Öttinger, H. C., 1996. Stochastic Processes in Polymeric Fluids. Springer.

36. Kuhn, W., 1934. Über die Gestalt fadenförmiger Moleküle in Lösungen. *Kolloid-Zeitschrift* 68:2–15.

37. Šolc, K., 1971. Shape of a Random-Flight Chain. *The Journal of Chemical Physics* 55:335–344.

38. Zifferer, G., 1999. Monte Carlo simulation studies of the size and shape of linear and star-branched polymers embedded in the tetrahedral lattice. *Macromolecular Theory and Simulations* 8:433–462.

39. Haber, C., S. A. Ruiz, and D. Wirtz, 2000. Shape anisotropy of a single random-walk polymer. *Proceedings of the National Academy of Sciences of the United States of America* 97:10792–10795.

40. Steinhauser, M. O., 2005. A molecular dynamics study on universal properties of polymer chains in different solvent qualities. Part I. A review of linear chain properties. *The Journal of chemical physics* 122:94901.

41. Theodorou, D. N., and U. W. Suter, 1985. Shape of unperturbed linear polymers: polypropylene. *Macromolecules* 18:1206–1214.

42. Bishop, M., and J. P. J. Michels, 1986. Polymer shapes in three dimensions. *The Journal of Chemical Physics* 85:5961–5962.

43. Soysa, W. C., B. Dunweg, and J. R. Prakash, 2015. Size, shape, and diffusivity of a single Debye-Huckel polyelectrolyte chain in solution. *The Journal of chemical physics* 143:64906.

44. Prabhakar, R., and J. Prakash, 2004. Multiplicative separation of the influences of excluded volume, hydrodynamic interactions and finite extensibility on the rheological properties of dilute polymer solutions. *Journal of Non-Newtonian Fluid Mechanics* 116:163–182.

45. Lyubartsev, and Laaksonen, 1995. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics* 52:3730–3737.

46. Lyubartsev, A. P., M. Karttunen, I. Vattulainen, and A. Laaksonen, 2002. On Coarse-Graining by the Inverse Monte Carlo Method: Dissipative Particle Dynamics Simulations Made to a Precise Tool in Soft Matter Modeling. *Soft Materials* 1:121–137.

47. Lyubartsev, A. P., A. Naome, D. P. Vercauteren, and A. Laaksonen, 2015. Systematic hierarchical coarse-graining with the inverse Monte Carlo method. *The Journal of chemical physics* 143:243120.

48. Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992. Numerical Recipes in C. Cambridge University Press, Cambridge, USA, second edition.

49. Fill, J. A., and D. E. Fishkind, 1998. The Moore – Penrose Generalized Inverse for Sums of Matrices Background and Main Result. *Education* 1–14.

50. Marrucci, G., R. B. Bird, C. F. Curtiss, R. C. Armstrong, and O. Hassager, 1989. Dynamics of polymeric liquids. Volume 2: Kinetic Theory. John Wiley & Sons, Ltd.