Multi-contact statistics distinguish models of chromosome organization

Janni Harju[1], Joris J.B. Messelink[2], and Chase P. Broedersz[1,2,*]

[1]Department of Physics and Astronomy, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands
[2]Arnold Sommerfeld Center for Theoretical Physics and Center for NanoScience, Department of Physics, Ludwig-Maximilian-University Munich, Theresienstr. 37, D-80333 Munich, Germany
[*]Corresponding author c.p.broedersz@vu.nl

May 17, 2022

Abstract

Whereas pairwise Hi-C methods have taught us much about chromosome organization, new multi-contact methods, such as single-cell Hi-C, hold promise for identifying higher-order loop structures. The presence of such high-order structure may be revealed by comparing multi-contact data with a theoretical prediction based on pairwise contact information. Here, we develop and compare three polymer-based prediction schemes for chromosomal three-point contact frequencies, based on a non-interacting polymer, a polymer with independent cross-linking, and a polymer with weak pairwise interactions between monomers. First, we test these predictions for two distinct simulation models of bacterial chromosome organization: a data-driven model inferred from a Hi-C map and bottom-up simulations of loop-extruding proteins. We find that the most predictive approximation is indicative of how contacts are primarily formed in a model. We then apply our prediction schemes to previously published super-resolution chromatin tracing data for human IMR90 cells. Strikingly, we find that the best prediction is given by the independent cross-linking approximation. This result is consistent with chromosomal contacts being dominantly caused by weakly interacting loop-extruders. Our work could have implications for developing models of chromosome organization from multi-contact data, and for better identifying higher-order loop structures.

1 Introduction

Over the last two decades, chromosomal capture experiments, such as Hi-C, have provided insight into how both eukaryotic and prokaryotic chromosomes are organized [1, 2, 3]. However, since traditional Hi-C methods provide only population-averaged pairwise contact frequency data, new methods must be used to assess to what extent chromosomal contacts are correlated. Patterns of contact correlations could be evidence for higher-order loop structures – structures with frequencies not encoded in pairwise contact data – potentially caused by factors such as transcription factories [4], super-enhancers [5], and interacting loop-extruding SMC complexes [6, 7].

To identify higher-order loop structures, several experimental protocols that track multiple contacts within individual cells have been implemented, yielding so-called "multi-contact data". The most general of these methods are single-cell experiments, where methods based on imaging [8, 9], single-cell Hi-C [10, 11], or single-cell SPRITE [12] are used to track contacts within individual cells. Others have adapted chromosomal capture methods to study the frequency of simultaneous contacts between three or more sites [13, 14, 15, 16, 17, 18]. Whereas these new experimental methods allow us to start the search for contact correlations, we still lack statistical tools to use these data to identify and interpret potential higher-order loop structures.

How can it be shown that the observed frequency of a given set of simultaneous contacts – a contact conformation – implies the presence of a higher-order loop structure? To address this question, the contact conformation frequency should be compared to a prediction based on pairwise contact frequencies. If the observed frequency is different from expected, the contacts in the conformation are presumed to be correlated. After such a contact correlation has been identified, it remains to be shown whether it is caused by higher-order collective effects, or by other factors, such as temporal correlations between loops (e.g. chromosome conformation changes during the cell cycle). Recent studies

have proposed a range of statistical approaches to predict contact conformation frequencies based on pairwise contact frequencies [14, 18, 8, 19]. However, these prediction schemes still lack theoretical motivation, and their performance has not been thoroughly compared or tested. Thus, it remains a challenge to develop a reliable method to identify and interpret higher-order structure in multi-contact data.

Here we formulate prediction schemes for contact conformation frequencies based on several simplified physical models for chromosome organization, including a non-interacting polymer and a polymer with independent cross-linking events, and discuss how they are related to previously used heuristic schemes [8, 14]. In addition, we develop a novel prediction scheme for frequencies of contact conformations on a polymer with short-range pairwise interactions between monomers, motivated by the maximum entropy (MaxEnt) model for a bacterial chromosome [20]. We test these predictions against three-point contact data sampled from the MaxEnt model, and from simulations of loop-extruders on a bacterial chromosome [7]. We find that the best performing prediction scheme corresponds to the dominant mechanism of contact formation in a model. Finally, we compare the performance of all three schemes for previously published data from single-cell experiments on human IMR90 cells [8]. Remarkably, we find that three-point contact frequencies are best predicted by a model where contacts between loci form independently of each other, as expected for a weakly interacting loop-extruder model. We hence offer guidelines for how chromosomal multi-contact data can be used to identify higher-order contact structures, and to gain insight into dominant mechanisms of chromosomal contact formation.

2 Results

Currently, three-point contact frequencies are the most prevalent type of multi-contact data [13, 14, 15, 16, 17, 18]. To investigate how to identify higher-order structure from such data, we start by introducing three prediction schemes for three-point contact statistics from pairwise contact probabilities, and briefly discuss how these predictions can be adapted for other contact conformations.

2.1 Independent loop scheme

As a simple starting point, we consider contact conformations on a non-interacting polymer in a cell, equivalent to a confined random walk. The contact probability $P_0(i,j)$ between any two sites i,j on the polymer is determined by the genomic length $|i-j|$ between them. In this case, the probabilities of contact conformations can be calculated by considering the effective genomic length of a given loop after the formation of another [21]. For three sites $i<j<k$, the effective genomic lengths of the two smallest loops (i,j) and (j,k) are independent of each other, and these two loops are equivalent to a three-point contact (i,j,k) (Figure 1A). Hence in this non-interacting limit, the three-point contact probability is given by

$$P_0(i,j,k) = P_0(i,j)P_0(j,k). \tag{1}$$

By replacing the non-interacting pairwise contact probabilities $P_0(i,j)$ with those found experimentally for a chromosome, $P(i,j)$, this formula provides an approximation scheme for three-point contacts that aims to take into account some of the chromosome structure present *in vivo*, such as variations in effective stiffness along the chromosome, and possible affinities between the pairs (i,j) and (j,k). This prediction scheme can also be adapted to describe other contact conformations by considering which loops have the shortest combined genomic length, and by presuming that these loops form independently [21]. We will hence refer to this approximation as the *independent loop formula*.

When a three-point contact is defined as the contacts (i,j) and (j,k) occurring simultaneously, the independent loop formula is equivalent to $P((j,k)|(i,j)) = P(j,k)$. This corresponds to the hypothesis used by Bintu et.al. to test for "cooperative, higher-order chromatin interactions" [8]. However, we note that this approximation does not take into account any possible pairwise interactions between i and k.

2.2 Independent link scheme

Next, we consider the limit where polymer contacts occur independently of each other. This limit could be approached, for example, if chromosomal contacts were dominantly caused by non-interacting passive cross-linking agents or active loop-extruders. Let $\tilde{P}(i,j)$ be the probability that the loci i and j are linked. The probability that there are at least two independent links between the sites i,j,k, resulting in a linked three-point contact, is then given by (Figure 1B)

$$\tilde{P}(i,j,k) = \tilde{P}(i,j)\tilde{P}(j,k) + \tilde{P}(i,k)\tilde{P}(k,j) + \tilde{P}(j,i)\tilde{P}(i,k) - 2\tilde{P}(i,j)\tilde{P}(j,k)\tilde{P}(k,i). \tag{2}$$

The last term ensures that the case when all three sites are linked is not over-counted. When most chromosomal contacts are caused by independent linking events, this formula can be used to approximate three-point contact probabilities

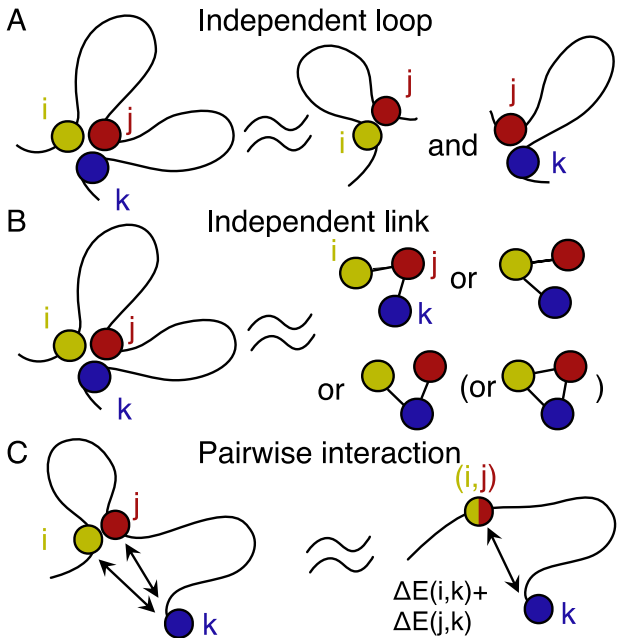

Figure 1: **Visualisation of approximation schemes for three-point contact frequencies**. **A.** When neglecting interactions, the probability of forming the three-point contact (i, j, k) is set by the genomic frequencies of the two smallest loops (i, j) and (j, k). **B.** When the genomic length has a small effect on linking probabilities, the three-point contact probability can be estimated as the probability that at least two links are present. The higher-order term can often be neglected. **C.** In the MaxEnt model or a pairwise interacting polymer, after a contact (i, j) has formed, the three-point contact (i, j, k) is equivalent to a contact $(k, (i, j))$, with an effective length $|j - k|$ and interactions between the neighbourhood of k and the other two points.

$P(i, j, k)$, by substituting all $\tilde{P}$'s by the respective pairwise contact probabilities. However, this approximation assumes that contact probabilities are independent of the effective genomic length between monomers. This assumption leads to an error that is especially relevant for the last higher-order term in Equation 2, since the loop (i, k) has an effective length of zero once the other two contacts have formed. The formula can be extended to other contact conformations by considering all possible combinations of independent links that would explain the contact conformation. We will hence refer to Equation 2, and its extensions for other contact conformations, as the *independent link formula*.

The independent link formula is related to the algorithm used by Olivares-Chauvet et.al. [14] for sampling three-point contacts. In the SI, we show that this algorithm assigns a three-point contact the relative weight

$$P(i, j, k) \propto P(j|i)P(i|k) + P(i|j)P(j|k) + P(i|k)P(j|k). \tag{3}$$

This approximation differs from Equation 2 in two ways: firstly, the higher-order term of Equation 2 is neglected; secondly, Equation 3 uses relative contact probabilities and gives a prediction up to a constant. The second consideration means that the prediction can be calculated using relative contact counts from Hi-C experiments. We note that when relative contact counts are used, the higher-order term from Equation 2 cannot be included, since it scales as P^3 rather than P^2.

We hence expect that when pairwise interactions between monomers are strong, as might be expected on a chromosome, the independent link formula should give better predictions for contact conformation frequencies than the independent loop formula. However, since the independent link formula assumes that linking probabilities are independent of the genomic distance between sites, it is not clear whether this is a good prediction scheme for chromosomal multi-contact data.

2.3 MaxEnt model and the pairwise interaction scheme

Given the simplifications made by the independent loop and independent link approximations, we next consider a data-driven alternative motivated by information theory. The first models that explicitly applied maximum entropy principles to Hi-C data assumed that the chromosome can be represented as a polymer at thermal equilibrium, and

data was used to constrain the number of contacts between predefined loop sites, and between loci of given chromatin types and at different genomic distances [22]. Later, our group developed a fully data-driven maximum entropy model for an entire bacterial chromosome (the MaxEnt model) [20], making no further assumptions than that the model should reproduce the normalized pairwise contact map. Previously, a model with similar effective interactions has been developed for sections of eukaryotic chromosomes [23]. The MaxEnt model can be used to sample full chromosome configurations, and hence also to predict contact conformation frequencies. However, since a MaxEnt model can be computationally demanding to train and sample, we here develop a simple approximation for its contact conformation frequencies. Such an approximation could be used to efficiently predict contact conformation frequencies from pairwise contact frequencies for systems with no MaxEnt model, providing an alternative to the easily computable independent loop and independent link formulae.

The maximum entropy distribution for chromosomes can be shown to assign a polymer configuration $\{\mathbf{x_i}\}_{i\in\{1,2,\dots N\}}$ with N monomers a probability

$$P(\{\mathbf{x_i}\}_i) \propto e^{-E(\{\mathbf{x_i}\}_i)}, \tag{4}$$

where $E(\{\mathbf{x_i}\}_i)$ is the effective energy of the polymer. The form of this energy function is determined by experimental constraints imposed on the model. When we constrain the pairwise contact frequencies of the monomers, and choose a lattice polymer representation, each contact (n, m) has an effective energy $\epsilon_{n,m}$, so that

$$E(\{\mathbf{x_i}\}_i) = \sum_{(n,m)} \epsilon_{n,m}\delta^3(\mathbf{x_n} - \mathbf{x_m}). \tag{5}$$

Here the sum runs over pairs of monomers $n < m$, and $\delta^3(\mathbf{x_n} - \mathbf{x_m})$ is the discrete delta-function, equal to one if the monomers (n, m) occupy the same lattice site, and zero otherwise. To train a model, an inverse learning algorithm iteratively adjusts the effective contact energies $\epsilon_{n,m}$ until the model quantitatively reproduces the experimentally measured Hi-C map.

To circumvent the training of a full MaxEnt model, we wish to predict how often three-point contacts in the model form, given that we have access to the pairwise contact probabilities $P(i, j)$, but not the effective contact energies $\epsilon_{i,j}$ (Figure 1C). Here we provide an intuitive overview of our approximation scheme and refer the reader to the SI for a detailed derivation.

We start by noting that the distribution of chromosome conformations in the MaxEnt model is mathematically equivalent to a pairwise interacting polymer in equilibrium, with Equation 4 assigning each configuration an effective Boltzmann weight. This equivalence allows us to use tools from statistical physics, and furthermore, any results we derive for the MaxEnt model will also hold for the associated equilibrium system. Using the terminology of statistical physics, the probability of any contact conformation is determined by its effective *energetic* and *entropic* cost. A combination with a higher energy will have a lower statistical weight, as seen from Equation 4, and in the limit of low energies, entropy can be seen as a measure of the number of possible polymer configurations with the contact conformation.

We first assume that the energetic costs of contacts in a three-point contact add, or that the energy corresponding to the three-point contact (i, j, k) is approximately the sum of the energies of (i, j), (j, k) and (i, k) [24]. In reality, this may result in over-counting energetic contributions when two points in the three-point contact are near each other. In the limit of a MaxEnt model with low contact energies, we can also find an approximation for the entropic cost of a three-point contact. On the non-interacting polymer, the probability of a three-point contact is given by the independent loop formula, and the entropic cost of the three-point contact is hence the sum of the entropic costs of the two independent loops. When contact energies in the MaxEnt model are sufficiently small, we therefore expect that a three-point contact's entropic cost approaches the entropic cost of its two smallest loops. Similarly, the entropic cost of the largest loop also approaches its entropic cost on the non-interacting polymer. These considerations lead to the following (low energy) pairwise interaction approximation:

$$P(i, j, k) \approx \frac{P(i, k)}{P_0(i, k)} P(i, j)P(j, k). \tag{6}$$

The terms $P(i, j)$ and $P(j, k)$ add the energetic and entropic costs of the two independent loops in the three-point contact, whereas the term $\frac{P(i,k)}{P_0(i,k)}$ adds an approximation for the energetic cost of the contact (i, k). The formula can be extended for other contact conformations by multiplying the independent loop approximation with a factor of P/P_0 for every contact not already included as a factor. However, the error is expected to grow as more energies are estimated via P/P_0. Nevertheless, the pairwise interaction formula takes into account both the energetic cost of the largest loop, unlike the independent loop formula, and changes in effective genomic length, unlike the independent link formula.

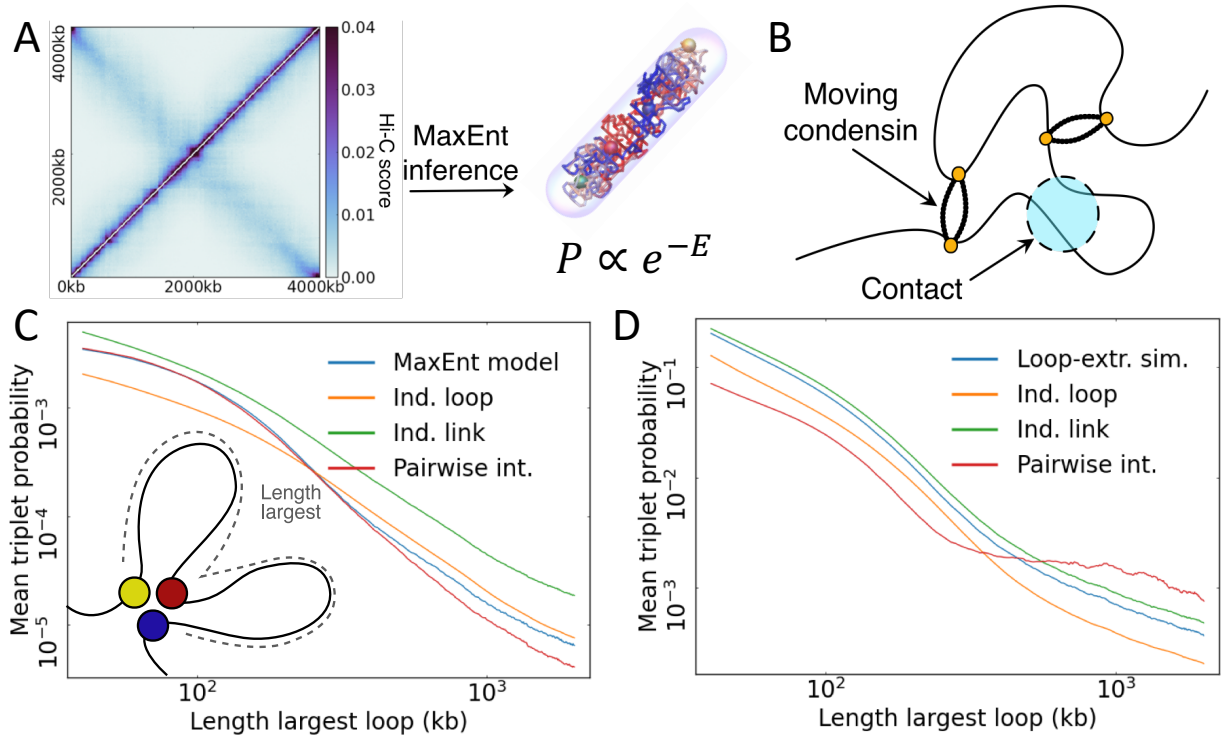

Figure 2: **MaxEnt and loop extruder simulations. A.** A Hi-C map for *C. crescentus* [3] is used to infer the effective contact energies $\epsilon_{i,j}$ for the MaxEnt model. The effective energies define a distribution of chromosome configurations that is sampled via a Monte Carlo algorithm. **B.** Loop-extruder simulations [27, 7] are conducted by updating the locations of moving condensins on a polymer, and letting the polymer configuration relax given the constraints. Contacts are defined as loci within a predefined range. **C, D.** Plots show $P_3(s)$ curves for data sampled from the MaxEnt model [20] for *C. crescentus* or from loop-extruder simulations, and predictions calculated using pairwise contact frequencies. The parameters used for the loop-extruder simulations are described in the Methods section.

2.4 Best prediction scheme reflects dominant mechanism for contact formation

We have introduced three formulae as prediction schemes for multi-contact statistics. To benchmark these methods, we next compare our predictions to simulated three-point contact data. We expect that the best matching prediction will reflect the dominant contact formation mechanism in a model. If this is the case, multi-contact data could be used to discriminate between models of chromosome organization. To test this idea, we examine simulated three-point contact data from the full MaxEnt model constrained by Hi-C experiments [20] (Figure 2A); and a model of loop-extruding condensins on a bacterial chromosome [7] (Figure 2B). We expect alternative simulation schemes for loop-extrusion [25, 26] to yield similar results, as long as loop-extruders interact weakly and/or rarely. We base both simulations on the chromosome of *Caulobacter crescentus*, a well-studied model organism with one circular chromosome over 4 Mb in length. This bacterium's Hi-C map is marked by an off-diagonal line of contacts emanating from the origin of replication (0 kb), as well as by rectangular regions of increased contacts around the primary diagonal, called chromosomal interaction domains (CIDs) [3]. The fully data-driven MaxEnt model reproduces all structure in the measured Hi-C map within experimental error, whereas the bottom-up loop-extruder model captures basic features of the Hi-C map such as the off-diagonal. These two models differ in their philosophies and the way in which contact formation is modelled, allowing us to test whether three-point contact statistics can be used to discriminate between them.

To easily visualise how well our three-point contact frequency prediction schemes perform, we adapt $P(s)$ curves for pair-wise contacts to describe three-point contacts. We define $P_3(s)$ as the average probability of three-point contacts where the largest loop is of genomic length s (Materials and Methods). We find that s is more indicative of prediction scheme performance than the size of either smaller loop in a three-point contact (Supplementary Information, Supplementary Figure 1). In addition, to avoid averaging over genomic location, we follow [13, 14, 17] and visualize

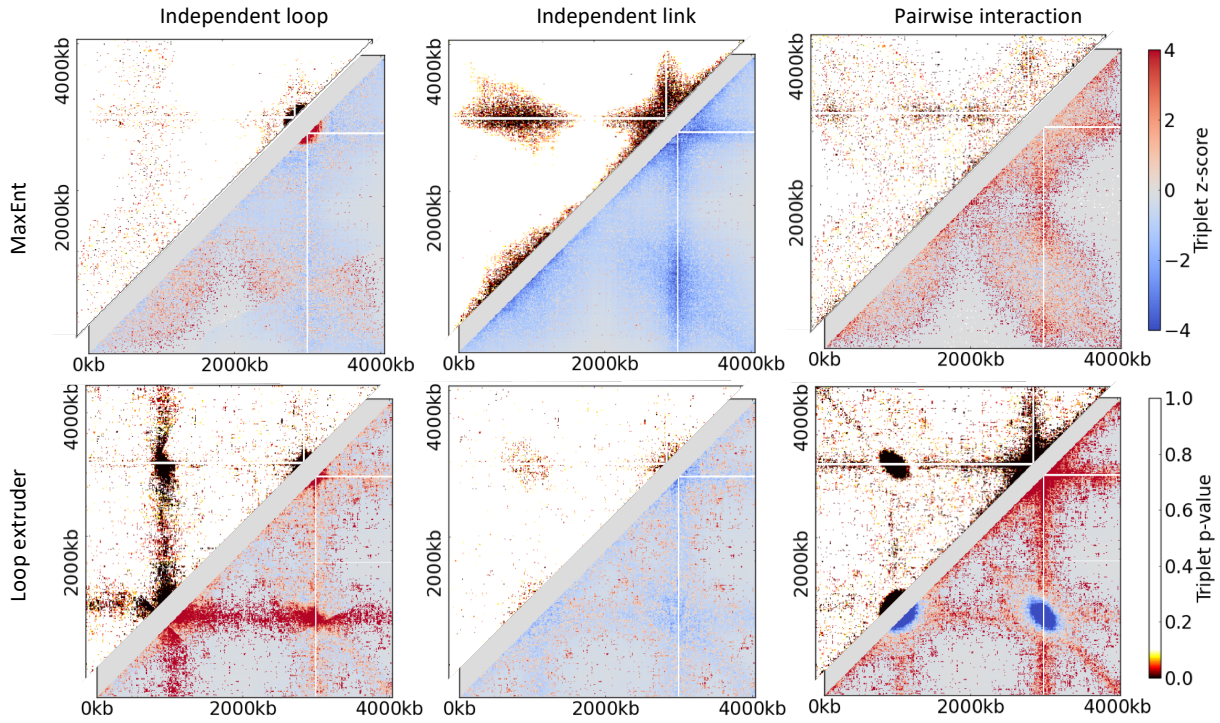

Figure 3: **Three-point contact predictions applied to simulated data for bacterial chromosomes.** Comparison of the MaxEnt/loop-extrusion simulation data used in Figure 2 C,D to the independent link, independent loop, and pairwise interaction formula. Top-left: p-values calculated by presuming the three-point contact data is binomially distributed with a probability given by the independent link, independent loop, or pairwise interaction prediction. Bottom-right: Z-values calculated for the three-point contact map compared to predictions. All plots are constructed for the bait point corresponding to 3000 kb.

three-point contact probabilities for a given bait point k as heatmaps where the intensity at point (i, j) reflects the frequency of three-point contacts (i, j, k).

As expected, the $P_3(s)$ curve (Figure 2C) of the MaxEnt model is best predicted by the pairwise interaction formula. The three-point contact map (Figure 3, first row) shows that the data mostly differs from the pairwise interaction prediction where three-point contact counts are very low, and the energy estimator has a larger error (Supplementary Figure 2). The results are similar for different bait points (Supplementary Figure 3), and when multiple hypothesis testing is accounted for, most deviations from the pairwise interaction formula do not appear significant (Supplementary Figure 4). By contrast, the independent loop formula gives significant under- and overestimates. The sign of the error is determined by whether the largest loop – neglected by the formula – has an attractive or repulsive effective energy. Finally, the independent link formula gives a persistent overestimate, mostly because it presumes a three-point contact can occur in four, rather than one, different ways. We hence conclude that the pairwise interaction scheme gives the best prediction of our MaxEnt model data, consistent with the model mapping chromosomal interactions into effective pairwise interactions

The three-point contact frequencies of the loop-extrusion simulations, on the other hand, are best predicted by the independent link scheme (Figure 3, second row). This scheme predicts only a slight over-estimate for loops of all sizes (Figure 2D), and after correcting for multiple hypothesis testing, few p-values are significant (Supplementary Figure 4). The independent loop formula, by contrast, fails to predict some of the lines on the three-point contact map (Figure 3). These missing lines correspond to contact triplets where the largest loop – again ignored by the formula – lies on the condensin trajectory. The pairwise interaction formula also gives an inaccurate prediction, partially because it presumes simultaneous interactions between all three points. We conclude that the simulated loop-extruder data is most accurately described by the independent link scheme, reflecting that the model's contact formation is driven by the localisation of weakly interacting loop-extruders.

So far, we have used absolute contact frequencies to predict three-point contact statistics. However, Hi-C experiments only provide relative contact counts, and hence all above formulae give predictions up to a constant prefactor. Previously,

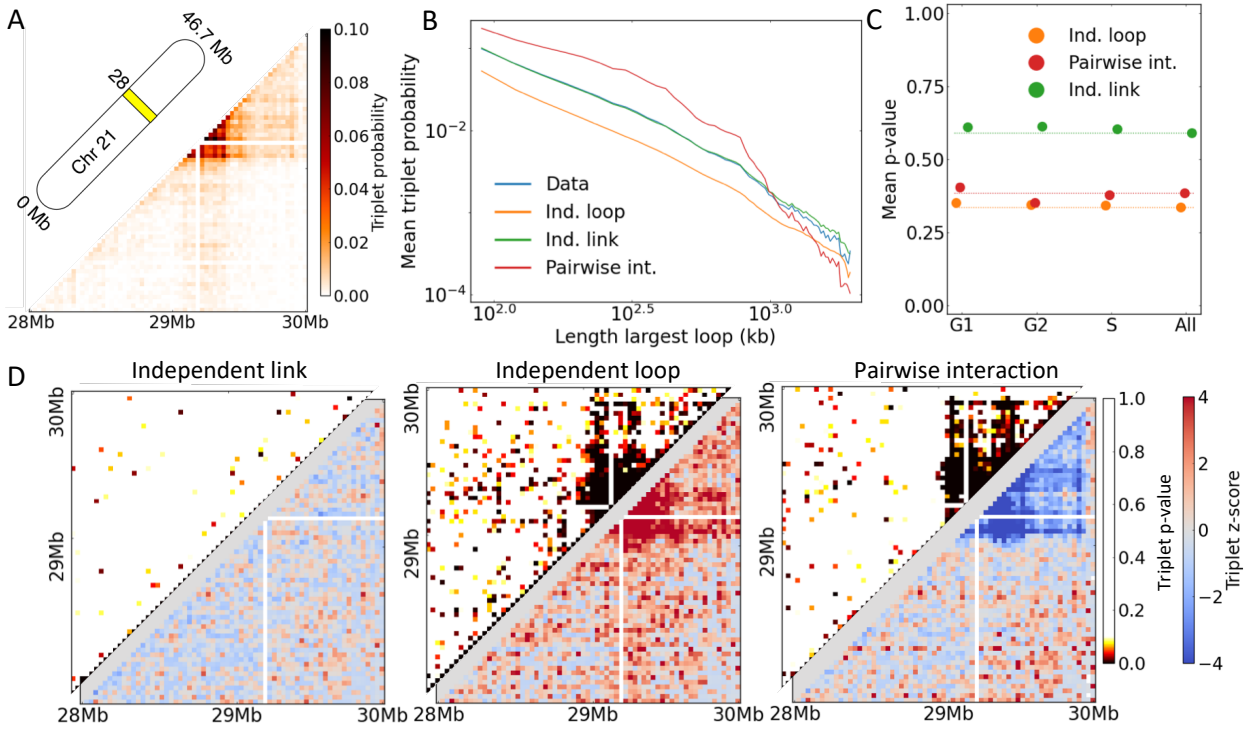

Figure 4: **Three-point contact predictions applied to experimental data on human chromosomes from Bintu et.al.. A.** Three-point contact probabilities. The arbitrary bait point is located at 29.1 Mb on chromosome 21. **B.** Similar to Figure 2C,D. Blue line describes the experimentally found contact triplet probability for chromosome 21 of 1574 IMR90 cells in the G1, G2 or S phase. The predicted curves were calculated using the experimentally found pairwise contact probabilities. Contacts were defined as spatial separations of < 150 nm between the centers of labelled 30 kb regions. **C.** Comparison of the p-value averaged over three-point contacts for data divided by cell cycle phase or combined together. Dashed lines indicate the p-values for the combined data. 288 samples were used to calculate all p-values, to ensure that results are comparable. **D.** P-value and z-score plots for the independent link, independent loop, and pairwise interaction formula, constructed as in Figure 3.

this prefactor has been set so that the expected total number of three-point contacts matches observations [14, 18]. We find that such a scaling can significantly improve three-point contact frequency predictions (Supplementary Figure 5), even if the formula does not reflect the underlying contact formation mechanisms of the system. Furthermore, scaling affects where three-point contacts appear to deviate from prediction, and hence potentially leads to misinterpretation of where contact correlations occur. We hence find that when absolute contact frequencies are used, the pairwise interaction and the independent link formula are able to predict background levels of three-point contacts of the MaxEnt and loop-extruder models, respectively. This is expected based on the underlying contact formation mechanisms of the two models. This supports the idea that the prediction scheme that best describes multi-contact data can indicate the most appropriate physical model for the system.

2.5 Experimental multi-contact data on human chromosomes are well described by independent link prediction

Having shown that simulated multi-contact data are best described by predictions that reflect the underlying contact formation mechanisms, we next explore which prediction scheme is most accurate for experimental multi-contact data. As three-point contact frequencies for *C. crescentus* or other bacterial systems are not yet available, we chose to analyse the super-resolution chromatin tracing data published by Bintu et.al. [8] (Materials and methods). These single-cell imaging data can be used to find the absolute frequencies of pairwise contacts and three-point contacts, avoiding the need for scaling predictions, which we found to distort predictions for simulated data. For simplicity, we first focus on cell-cycle averaged data. Although the data represent a small part of a human chromosome (Figure 4B), their analysis in terms of multi-contact statistics serves as an illustrative example. We hence extract pairwise contact frequencies from

the data, apply our formulae for the three physical scenarios to make predictions for the three-point contact frequencies, and then compare these predictions to actual three-point contact frequencies also extracted from the data.

Strikingly, the independent link formula describes the experimental data the best, and at an accuracy comparable to our results for loop-extruder simulations. Both the $P_3(s)$ curve and the averaged z-score plots show that the independent link formula tends to give slight overestimates for three-point contact frequencies, but the difference is smaller than for loop-extrusion simulation data (Figure 4A,D, Supplementary Figure 6). Since the independent loop formula always yields a lower prediction than the independent link formula, it here gives a persistent underestimate of the three-point contact frequencies. The pairwise interaction formula gives persistent overestimates for three-point contacts within TADs (Supplementary Figure 7). This could be a similar error as seen for the simulated loop-extruder data (Figure 3); if a three-point contact within a TAD occurs due to two cohesins that collide, there is no "attractive force" associated with the third contact, as presumed by the pairwise interaction formula. However, for three-point contacts across TADs, the pairwise interaction formula performs significantly better than the independent loop formula. When controlling for multiple hypothesis testing, hardly any deviations from the independent link formula can be considered significant (Supplementary Figure 6). We thus conclude that the three-point contact data would be consistent with a model where contacts are dominantly caused by weakly interacting cross-linkers, such as loop-extruders capable of bypassing one another.

To test whether the change of the pairwise contact probabilities throughout the cell cycle has an impact on the quality of our prediction schemes, we repeated the analysis for contact data separated by cell cycle stage. This separation of data does not noticeably improve the mean p-value for our predictions (Figure 4C), and we hence conclude that at this sample size contact correlations due to contacts being enhanced during the same cell cycle stage are not prominent.

3 Discussion

Experimental multi-contact data hold great potential for identifying higher-order chromosomal structure. However, to test the null hypothesis that observed multi-contact statistics are merely a result of pairwise chromosomal contact frequencies, prediction schemes for higher-order contact structures are needed. We discussed how three such schemes – the independent link, the independent loop, and the low-energy pairwise interaction formula – can be physically motivated. Since each formula corresponds to a different simplified picture of chromosomal organization, we hypothesized that the best prediction scheme for given multi-contact data could be reflective of dominant mechanisms of contact formation in the studied system.

We tested the three prediction schemes against data simulated using a MaxEnt model with effective pairwise interactions and a model of weakly interacting loop-extruders. As expected, the pairwise interaction formula described the MaxEnt data most accurately, whereas the loop-extrusion data were best described by the independent link formula, despite that the model features some condensin interactions upon collision. Our results hence illustrate that the three approximation schemes can give predictions that differ significantly, and that an approximation can be accurate when applied to data from an appropriate physical model.

By applying our prediction schemes to previously published experimental data for IMR90 chromosomes [8], we showed that the independent link formula best described the data at the megabase scale. Our results are consistent with the hypothesis that chromosomal contacts are dominantly caused by weakly interacting loop-extruders. We note that if loop-extruders stalled upon collision for significant periods of time, three-point contacts should be stabilised, and deviations from the independent link formula would be expected. This illustrates that testing multi-contact data against different predictions can lead to insight on the nature of the mechanisms driving chromosomal contact formation.

We repeated our analysis for these experimental data using the three-point contact frequency prediction recently proposed by Liu et al. [19]. Their method is based on inferring a set of spring constants from Hi-C data, and is hence computationally more complex than the independent link or independent loop approximations, which calculate a set of three-point contact frequencies with optimal N^3 scaling. Nevertheless, for the Bintu et.al. data analysed, we found the independent link formula performed slightly better than the Liu et.al. prediction (mean p-values 0.59 vs. 0.52; Supplementary Figure 8). Furthermore, unlike the independent link prediction, the Liu et.al. prediction had to be scaled to match the observed number of three-point contacts. For simulated MaxEnt data, we found that such scaling significantly improved predictions. The comparatively good performance of the independent link approximation suggests that the simple prediction schemes we have described provide a useful benchmark when constructing more complicated models for predicting multi-contact frequencies.

Our findings suggest that simple approximations based on pairwise contact data can be used to predict background levels of three-point contact frequencies in different scenarios. However, to claim that deviations from such predictions result from inherently higher-order loop structures, different predictions based on pairwise contact frequencies should

be tested, appropriate methods for multiple hypothesis testing should be used, and ambiguities arising from the scaling of predictions or the use of unsynchronized data should be addressed. Such practices could both offer insight into how a system's chromosomal contacts are best described, and improve the accuracy at which higher-order loop structures are identified, potentially leading to the discovery of new mechanisms of chromosome organization.

4 Methods and Materials

4.1 The MaxEnt model

The MaxEnt model for bacterial chromosomes represents the least assuming model for chromosome organization given a Hi-C map [20]. A converged model was used to sample full chromosome configurations and the scripts are accessible at [28]. Contacts were defined as two monomers occupying the same lattice site, and three-point contacts as three monomers occupying the same lattice site. The model was sampled using a Monte Carlo algorithm (76800 samples), and contacts in each sample were saved. Raw data and a Julia script for analysis is available in github.com/PLSysGitHub/chromosomal_multi_contact_data.

4.2 Loop-extruder simulations

The script used for [7], available at [29], was adapted to resemble the Hi-C map of *C. crescentus*. A single loop-extruder loading site at 1 kb was used, and loop-extruders were assumed to move at equal speeds in both directions on the chromosome (parameter "wind" set to zero in simulations). Raw configuration data and Julia scripts for analysis are available in github.com/PLSysGitHub/chromosomal_multi_contact_data. Contacts were defined as a distance of less than 5 simulation units between monomers, and three-point contacts as events where at least two contacts were present between three monomers. We note that using this definition, the pairwise interaction scheme makes a further approximation; it neglects the fact that two loci in a three-point contact can be more than a cross-linking radius apart, which can change both the entropic and energetic cost of this secondary contact. Contact and three-point contact frequencies were calculated by sampling 3000 polymer configurations.

4.3 Data from Bintu et.al.

We analysed super-resolution chromatin tracing data published by Bintu et.al. [8], available at github.com/BogdanBintu/ChromatinImaging. The authors imaged 65 neighbouring 30 kb intervals of human chromosome 21, and thus gathered data on their relative positions. The data files containing relative positions of loci were used to calculate the frequencies of contacts (two loci separated by a distance less than 150 nm) and three-point contacts (three loci with at least two contacts between them), using a Julia script available in github.com/PLSysGitHub/chromosomal_multi_contact_data.

4.4 Non-interacting simulations

The pairwise interaction formula (Equation 6) requires the probabilities of contacts on a non-interacting, ideal polymer, with the same length and confinement volume as the chromosome.

For the MaxEnt model, the simulations were run with the effective energies between all monomers set to zero. For the loop-extrusion model, excluded volume interactions were set to zero, and no loop-extruders or plectonemes were included in the simulations. The models were then sampled for contact probabilities as before.

For the experimental data from Bintu et.al., we required estimates for the shape and size of the confinement volume, the monomer length, b, and the number of monomers each bin is mapped to, n. We chose to consider a spherical volume of confinement, with a radius R given by one half of the mean maximal cross-section of the chromatin region. For each data set, we calculated the distribution for the separation d between neighboring 30 kb regions, and used $\langle d^2 \rangle = nb^2$ to set b. We found that the choice of n did not significantly affect our results (Supplementary Figure 9). Unless otherwise stated, results are shown for $n = 10$. We simulated confined random walks, and tracked how frequently every nth monomer was within a distance < 150 nm of each other. Code is available in github.com/PLSysGitHub/chromosomal_multi_contact_data.

4.5 2D averaged three-point contact plots

Given a three-dimensional array M corresponding to three-point contact data, we defined 2D averaged plots A (Supplementary Figure 1) as follows. For a three-point contact $i < j < k$, let $x = \max(j - i, k - j)$, $y =$

$\min(j-i, k-j)$. Increment $A[x, y]$ with $M[i, j, k]$. Divide each $A[x, y]$ by the number of three-point contacts with $x = \max(j-i, k-j)$ and $y = \min(j-i, k-j)$.

For a circular chromosome of length N, the algorithm must be adjusted. We calculated the minimum genomic distance between each pair of points in the three-point contact, $\min(j-i, N-j+i)$. x and y were chosen as the two lowest minimum genomic distances.

4.6 $P_3(s)$ curves

Given a three-point contact frequency array M, we calculated $P_3(s)$ curves as follows. For each three-point contact $i < j < k$, add $M[i, j, k]$ to $P_3[k-i]$. Divide each $P_3[s]$ by the number of three-point contacts $i < j < k$ with $k - i = s$.

We note that for a circular chromosome of length N, the genomic length of the largest loop is given by either $s = k - i$ or $s = N - k + i$. For consistency with 2D averaged plots, we chose s equal to the sum of the two smaller loops in the three-point contacts, so that lines of $x + y = s$ on the 2D averaged plots still correspond to single points on the $P_3(s)$ curves. Using the minimum genomic length would map points with $N/2 < s < 2N/3$ to $N - s$. We hence display the plots only for $s < N/2$. As long as the same definition is used for comparing predicted $P_3(s)$ curves, the comparison is informative.

4.7 Statistical analysis

Z-scores and two-tailed p-values for three-point contact frequencies were calculated by presuming that the number k of three-point contacts observed in n samples follows a binomial distribution $k \sim B(n, p)$, where p is the predicted frequency of the three-point contact. Values were calculated using the HypothesisTests package for Julia [30].

The Benjamini-Hochberg procedure [31, 32] was used to analyse whether deviations were significant when the large number of possible three-point contacts was taken into account. The Benjamini-Hochberg procedure controls the false discovery rate (FDR), or the probability that we falsely reject the hypothesis for an individual three-point contact. If an FDR of α is required, all hypotheses with an adjusted p-value smaller than α should be neglected. Adjusted p-values were calculated using the MultipleTesting package for Julia [33].

4.8 Algorithm by Liu et.al. for Bintu et.al. data

The code available at github.com/leiliu2015/HLM-Nbody was used to analyse the Bintu et.al. data. The script for Tri-C data was adapted to take in the Bintu et.al. data as input, and to calculate a prediction for three-point contacts around every experimentally tracked locus. The results were combined into a 3D three-point contact probability array. The 3D array was analysed in the same way as for the independent link, independent loop and pairwise interaction schemes, after scaling the results for each viewpoint to match the experimentally observed three-point contact counts.

5 Acknowledgments

We thank Hugo Brandão for discussions and help with loop-extruder simulations, and Pedro Olivares-Chauvet for help accessing experimental data.

6 Data Availability

Contact and three-point contact frequency files extracted from simulations and data from Bintu et.al., as well as Julia scripts for analysis, are available at github.com/PLSysGitHub/chromosomal_multi_contact_data.

References

[1] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, Feb 2002.

[2] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009.

[3] Tung B. K. Le, Maxim V. Imakaev, Leonid A. Mirny, and Michael T. Laub. High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science*, 342(6159):731–734, Nov 2013.

[4] Lucas Brandon Edelman and Peter Fraser. Transcription factories: genetic programming in three dimensions. *Curr. Opin. Genet. Dev.*, 22(2):110–114, Apr 2012.

[5] Frank Grosveld, Jente van Staalduinen, and Ralph Stadhouders. Transcriptional Regulation by (Super)Enhancers: From Discovery to Mechanisms. *Annu. Rev. Genomics Hum. Genet.*, 22(1), May 2021.

[6] Eugene Kim, Jacob Kerssemakers, Indra Shaltiel, Christian Haering, and Cees Dekker. DNA-loop extruding condensin complexes can traverse one another. *Biophysical Journal*, 118(3):438–442, 2020.

[7] Hugo B. Brandão, Zhongqing Ren, Xheni Karaboja, Leonid A. Mirny, and Xindan Wang. DNA-loop-extruding SMC complexes can traverse one another in vivo - Nature Structural & Molecular Biology. *Nat. Struct. Mol. Biol.*, 28(8):642–651, Aug 2021.

[8] Bogdan Bintu, Leslie J. Mateo, Jun-Han Su, Nicholas A. Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N. Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413):eaau1783, Oct 2018.

[9] Andrés M. Cardozo Gizzi, Sergio M. Espinola, Julian Gurgo, Christophe Houbron, Jean-Bernard Fiche, Diego I. Cattoni, and Marcelo Nollmann. Direct and simultaneous observation of transcription and chromosome architecture in single cells with Hi-M. *Nat. Protoc.*, 15(3):840–876, March 2020.

[10] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L. Gunderson, Frank J. Steemers, Christine M. Disteche, William S. Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell Hi-C. *Nat. Methods*, 14(3):263–266, Mar 2017.

[11] Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, Oct 2013.

[12] Mary V. Arrastia, Joanna W. Jachowicz, Noah Ollikainen, Matthew S. Curtis, Charlotte Lai, Sofia A. Quinodoz, David A. Selck, Rustem F. Ismagilov, and Mitchell Guttman. Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nat. Biotechnol.*, 40(1):64–73, January 2022.

[13] Tingting Jiang, Ramya Raviram, Valentina Snetkova, Pedro P. Rocha, Charlotte Proudhon, Sana Badri, Richard Bonneau, Jane A. Skok, and Yuval Kluger. Identification of multi-loci hubs from 4c-seq demonstrates the functional importance of simultaneous interactions. *Nucleic Acids Research*, 44(18):8714–8725, 2016.

[14] Pedro Olivares-Chauvet, Zohar Mukamel, Aviezer Lifshitz, Omer Schwartzman, Noa Oded Elkayam, Yaniv Lubling, Gintaras Deikus, Robert P. Sebra, and Amos Tanay. Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. *Nature*, 540(7632):296–300, 2016.

[15] Ferhat Ay, Thanh H Vu, Michael J Zeitz, Nelle Varoquaux, Jan E Carette, Jean-Philippe Vert, Andrew R Hoffman, and William S Noble. Identifying multi-locus chromatin contacts in human cells using tethered multiple 3c. *BMC Genomics*, 16(1):121, 2015.

[16] Amin Allahyar, Carlo Vermeulen, Britta A. M. Bouwman, Peter H. L. Krijger, Marjon J. A. M. Verstegen, Geert Geeven, Melissa van Kranenburg, Mark Pieterse, Roy Straver, and Judith H. I. Haarhuis. Enhancer hubs and loop collisions identified from single-allele topologies. *Nature Genetics*, 50(8):1151–1160, 2018.

[17] Emily M. Darrow, Miriam H. Huntley, Olga Dudchenko, Elena K. Stamenova, Neva C. Durand, Zhuo Sun, Su-Chen Huang, Adrian L. Sanborn, Ido Machol, Muhammad Shamim, Andrew P. Seberg, Eric S. Lander, Brian P. Chadwick, and Erez Lieberman Aiden. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U.S.A.*, 113(31):E4504–E4512, Aug 2016.

[18] Filipe Tavares-Cadete, Davood Norouzi, Bastiaan Dekker, Yu Liu, and Job Dekker. Multi-contact 3C reveals that the human genome during interphase is largely not entangled. *Nature Structural & Molecular Biology*, 27(12):1105–1114, 2020.

[19] Lei Liu, Bokai Zhang, and Changbong Hyeon. Extracting multi-way chromatin contacts from Hi-C data. *PLoS Comput. Biol.*, 17(12):e1009669, December 2021.

[20] Joris J. B. Messelink, Muriel C. F. van Teeseling, Jacqueline Janssen, Martin Thanbichler, and Chase P. Broedersz. Learning the distribution of single-cell chromosome conformations in bacteria reveals emergent order across genomic scales. *Nat. Commun.*, 12(1963):1–9, Mar 2021.

[21] Edward J Banigan, Aafke A van den Berg, Hugo B Brandão, John F Marko, and Leonid A Mirny. Chromosome organization by one-sided and two-sided loop extrusion. *eLife*, 9, 2020.

[22] Michele Di Pierro, Bin Zhang, Erez Lieberman Aiden, Peter G. Wolynes, and José N. Onuchic. Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. U.S.A.*, 113(43):12168–12173, Oct 2016.

[23] Luca Giorgetti, Rafael Galupa, Elphège P. Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, and Edith Heard. Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. *Cell*, 157(4):950–963, May 2014.

[24] A. Ben-Naim. Statistical potentials extracted from protein structures: Are these meaningful potentials? *J. Chem. Phys.*, 107(9):3698–3706, Sep 1997.

[25] Christiaan A. Miermans and Chase P. Broedersz. A lattice kinetic Monte-Carlo method for simulating chromosomal dynamics and other (non-)equilibrium bio-assemblies. *Soft Matter*, 16(2):544–556, January 2020.

[26] Andrea Bonato and Davide Michieletto. Three-dimensional loop extrusion. *Biophys. J.*, 120(24):5544–5552, December 2021.

[27] Hugo B. Brandão, Payel Paul, Aafke A. van den Berg, David Z. Rudner, Xindan Wang, and Leonid A. Mirny. RNA polymerases as moving barriers to condensin loop extrusion. *Proc. Natl. Acad. Sci. U.S.A.*, 116(41):20489–20499, Oct 2019.

[28] Joris J B Messelink. Jorisjjb/maxent-chromosome-caulobacter v0.1, January 2021.

[29] Hugo B. Brandão. hbbrandao/bacterialSMCtrajectories: Bacterial_SMC_complex_Simulations, June 2021.

[30] Andreas Noack. HypothesisTests.jl. https://github.com/JuliaStats/HypothesisTests.jl, 2021.

[31] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29(4):1165–1188, Aug 2001.

[32] Jelle J. Goeman and Aldo Solari. Multiple hypothesis testing in genomics. *Stat. Med.*, 33(11):1946–1978, May 2014.

[33] Julian Gehring and Nikos Ignatiadis. MultipleTesting.jl. https://github.com/juliangehring/MultipleTesting.jl, 2017 – 2021.