Polymer Chain Dynamics - Summary of Methods and Findings

September 2, 2014

Contents

Chapter 1

Thoretical Materical

1.1 The Topologically associating domains (TADs)

The Topologically Associating Domains (TADs), which belong to the X inactivation center (Xic), are discrete adjacent chromosomal regions prominently interacting among themselves. By interaction we mean close physical proximity of one region to another part of the same chromosome, determined by the resolution of the 5C technique [4][1], for which chromosomal regions are fixed in formaldehyde, and contact are defined as region close enough to be considered as having protein-protein interactions.

The Xic is a 4.5 Mb region located on embryonic stem cells' X chromosome and is believed to be conserved throughout the mammalian family. Nora et. al[13], recognized 9 TADs within the Xic region, and termed them in alphabetical order A to I (see Figure 1.1) These region span 200kb to 1Mb in size.

The Xic orchestrate the inactivation of the X chromosome by controlling the transcription of its coding sequences. The major player in the inactivation process is the Xist sequence.

The Xist sequence is located on the TAD termed E. The repressive anti-sense of Xist, termed Tisx, is located on the adjacent D TAD (Figure 1.2). In this report we focus only on TADs D and E. The TADs D and E are approximately of size 324 kb and 537 kb respectively (see Figure 1.1) and contain non-protein and protein coding sequences(For a graphical view see Figure 1.2), among which, the TAD D harbors the Xite regulator of Xist gene, and Tsix which is the repressive anti-sense transcript of Xist.

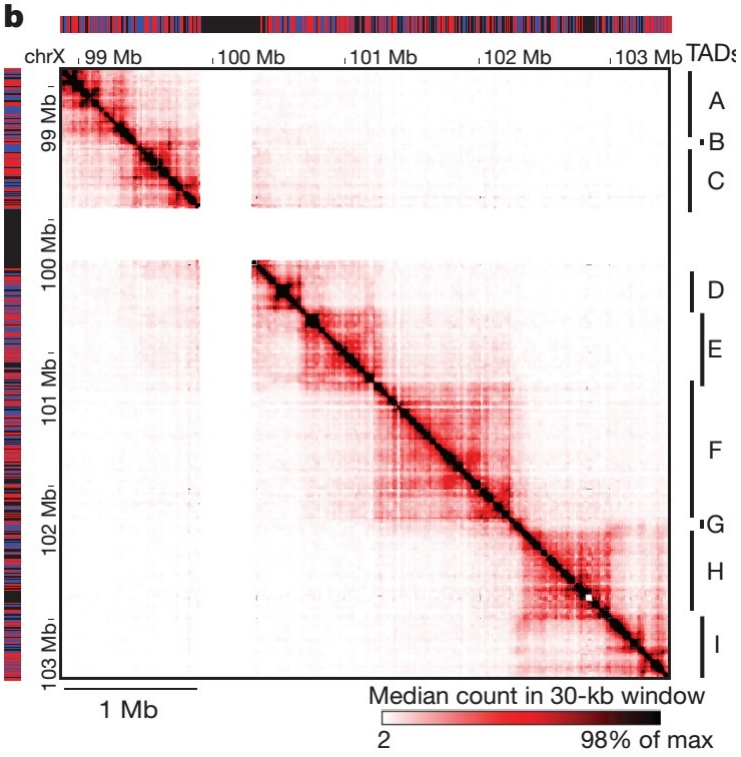


Figure 1.1: A display of the TADs in the region of the X chromosome. Taken from Nora et al.Nature. 2012 Apr 11;485(7398):381-5 [13]. The red-blue bar on top and left show the restriction segments cut by HINDIII (forward and reverse segments)

Figure 1.2: The genes and non protein coding regions of TAD D and E. Image taken from from Nora et al.Nature. 2012 Apr 11;485(7398):381-5 [13]

1.1.1 A short review of the 5C method

The 5C method [4] relies on the first steps of the seminal 3C method proposed by Dekker et al. 2002 [3]. From the supplementary material of the article by Dekker et al. The steps of the 3C can be described as follows (see also Figure 1.3):

1. Roughly 10^8 intact nuclei are isolated

2. Nuclei are cross-linked by using 1% formaldehyde or its relative paraformaldehyde for 10 minutes, which induces protein-protein and DNA-protein cross-linking. The cross-link is made by formaldehyde in the guanine nucleotide, between DNA-bound proteins and the nucleotide.

3. the cross-linking distance is between $10 - 100nm$ [2]

4. All non cross-linked proteins are removed. It is important to note that the DNA is purified only After the cross-links were made.

5. A site-specific restriction enzyme (ECOr or HindIII, 6bp cuter) is used to digest the cross-linked DNA for 1 hour.

6. The restriction enzymes are inactivated using SDS and incubation for 20 minutes.

7. The reaction is 15 times diluted to favor relevant DNA end ligation.

8. The free ends of the cleaved DNA are ligated (DNA is still cross-linked with proteins) for 45 minutes using ligase.

9. Reverse cross-linking by overnight incubation: the cross-links are destroyed, leaving segments of ligated DNA. The reverse cross-linking is done by incubation at 70^0.

10. DNA is purified from non cross-linked proteins.

11. The products of the ligation are detected and quantified using PCR.

12. a control template is created by using the same restriction enzyme on non cross-linked DNA and ligation of DNA fragments without dilution

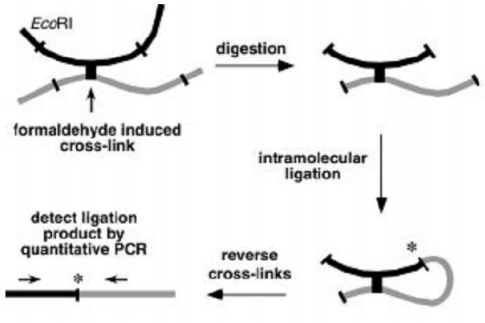


Figure 1.3: A schematic representation of the 3C method. Image taken from Dekker et al. 2002[3]

1.2 The Experimental Data

Two replicates of 5C experiments were conducted taking into account the Xic region. The data was coarse-grained as explained in 1.2.1. The encounter data includes the number of encounters between 14,508 genomic segments. Each genomic segments is defined by the start and end bead.

1.2.1 Coarse-graining of the data

The following paragraphs are an adapted version of the text in from Giorgetti et al. [7] supplementary material.

In the genomic region of the TAD D and E, 124 forward (FOR) and 126 reverse (REV) primers anneal to alternate HindIII restriction fragments of variable size. Since the average length of FOR and REV fragments in this region is 3078 bp, a 3000-bp beads was chosen, resulting in polymers of N=108 beads in the case of TAD D alone (chrX:100378306-100699670, total length 321364 bp) and $N = 307$ beads in the case of TAD D+E (total length 920432 bp). Thirty five percent of restriction fragments are longer then 3000 bp and will be mapped on multiple beads, while the remaining 65% will contribute together with their nearest neighbor to define the expected contacts of single beads.

To simulate 5C experiments, the uneven sampling of the genomic region provided by restriction fragments of different length must be mapped onto the even sampling provided by equally spaced beads. This is performed as follows:

1. The 5 end of the first bead in the chain is positioned at the 5 end of the first restriction fragment in the region of interest

2. Each restriction fragment in the region is assigned two indices i and j $(i, j = 1N)$ corresponding to the beads to which its 5 and 3 ends overlap.

3. 5C counts corresponding to each pair of restriction fragments are assigned to the corresponding pairs of indices (i,j) and (h,k), for example:

4. If two or more consecutive FOR or REV restriction fragments map to the same bead, their contributions are summed ($< 10\%$ of all interactions).

5. 5C counts of two experimental replicates are averaged for each pair of restriction fragments and their standard deviation is taken as a measure of experimental uncertainty.

1.2.2 The encounter matrix

The data provided by Giorgetti et al.[7] contains the encounter frequency of beads (not segments!) constructed as explained in 1.2.1. Sorting the encounter data into a 2D matrix results in a 307 by 307 array. Reminder: beads are numbered sequentially from the 5-end to the 3-end of the polymer.

1.3 Analysis of the data

1.3.1 From encounter frequency to encounter probability

We start by transforming the encounter data into encounter probability. Each line of the encounter matrix is divided by the sum of all encounters of the line.

1.3.2 Showing the TADs

TADs are seen by displaying the normalized encounter data, after removing closest beads data from the matrix of encounters. The normalization is done

by taking each row k of the encounter matrix, removing the data for the closest neighbors of bead k and dividing by the sum of the row. A median filter of size 30 beads is then applied to the image (see methods in [13]).

1.3.3 Symmetry of encounter frequency data

We next examined whether the data can be regarded as symmetric in the sense of encounter probabilities for each bead. That is, whether bead k has the same probability to meet its closest neighbors from the right and left in the linear chain, or not.

The symmetry of encounter frequency data is shown by measuring the left and right encounter frequencies of each one of the beads comprising the chain. The left and right encounter data are defined in terms of bead index, where for the bead j the left encounter frequencies are defined as a vector of length $j-1$ with the encounter data of bead j with $1, .., j-1$, and $j+1, .., N$ for the right encounter frequencies.

For bead index $j = 1, ..., N$ we have different number of beads on the left and right. For comparison, we need to match the number of beads from the left and right for each bead. The number of beads on the left and right of bead j is then taken to be $n_j = min(j-1, N-j+1)$. Let L_j be the left encounter data, R_j be the right encounter data vectors of lengths n_j

$$L_j = [e(j-1), e(j-2), ...e(j-n_j)]$$

and

$$R_j = [e(j+1), e(j+2), ..., e(j+n_j)]$$

where $e(k)$ is the encounter frequency of the k^{th} closest bead to the bead j. Note the reverse order of L_j.

We normalize each vector by dividing it by the sum of its encounters

$$\hat{L_j} = [e(j-1), e(j-2), ..., e(j-n_j)] \,/ \sum_{k=j-n_j}^{j-1} e(k)$$

and similarly for $\hat{R}_j$.

To test the difference between left and right frequencies, we calculate the mean difference by

$$d_j = \frac{1}{n_j} \sum_{k=1}^{n_j} \left(\hat{L}_j(k) - \hat{R}_j(k) \right)$$

The Figure 1.4 below shows the results of plotting the mean difference for the two replicates of the 5C data

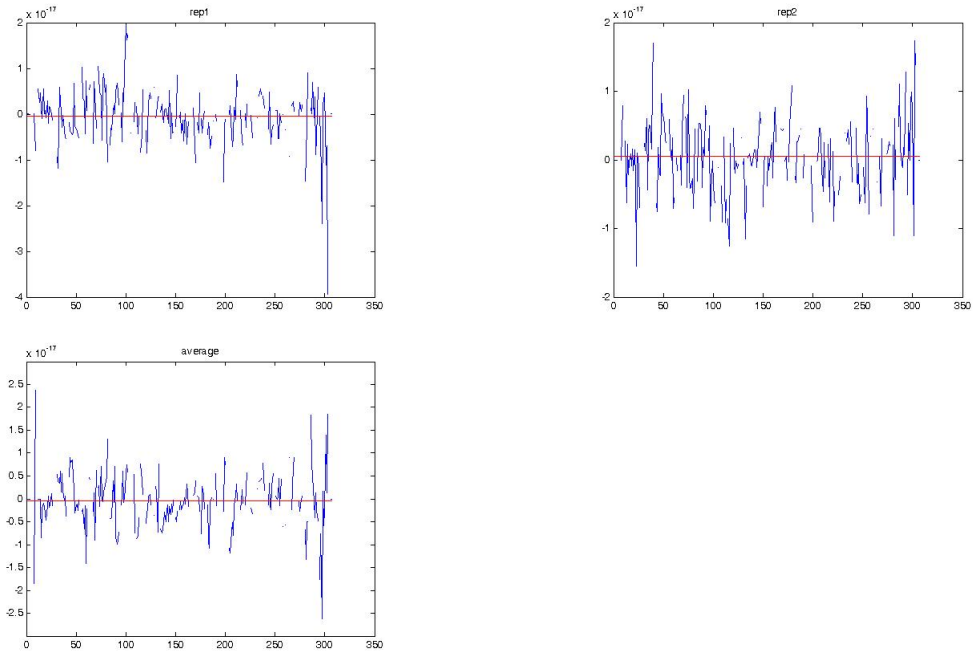


Figure 1.4: The mean difference (y axes) between left and right encounter frequencies for each bead number (x axes) is shown for Rep1 (top left) Rep2 (top right) and average (bottom)

The red lines in the Figure 1.4 signify the mean of the difference data. For the three figures it is of the order 10^{-18}. When points are missing, it means that the segment (restriction segment) occupies more than one bead, therefore the encounter data is zero there.

We can therefore treat the right and left data in the same manner and 'fold' the encounter data such that from now on we test the encounter data by bead distance. the 'folded' encounter data will be called just the encounter data from now on.

1.3.4 Peaks of the encounter data

For some beads the encounter data shows peaks at some distance from the origin. These peaks represent a frequent encounter with a distal part of the chain. The list of bead numbers and the frequent encounter they have with

the distal chain parts (represented by bead numbers) is summarized in Table 1.3.4

bead numbers	encountered beads	TAD
23-26	280-290	$D \leftrightarrow E$
49-53	148-155	$D \leftrightarrow E$
56-59	80-90	$D \leftrightarrow D$
115-117	165-170	$E \leftrightarrow E$
161-162	187 190	$E \leftrightarrow E$
182-184	260-264	$E \leftrightarrow E$
185-186	253-255	$E \leftrightarrow E$
234-236	184-189	$E \leftrightarrow E$
234-236	4-11	$E \leftrightarrow D$
243	88	$E \leftrightarrow D$
264	89-90	$E \leftrightarrow D$
274-277	113-120	$E \leftrightarrow D(?)$

bead numbers	encountered beads	TAD	rep
1	86	$D \leftrightarrow D$	1
6	81	$D \leftrightarrow D$	1
6	227-229	$D \leftrightarrow E$	1
6	243	$D \leftrightarrow E$	1
7	81	$D \leftrightarrow D$	1
7	226-229	$D \leftrightarrow E$	1
7	247	$D \leftrightarrow E$	1
10	77	$D \leftrightarrow D$	1
10	77	$D \leftrightarrow D$	1
23	293	$D \leftrightarrow E$	1
24	264	$D \leftrightarrow E$	1
25	285	$D \leftrightarrow E$	1
26	285	$D \leftrightarrow E$	1
26	86	$D \leftrightarrow D$	1
26	59-74	$D \leftrightarrow D$	1
49-53	150	$D \leftrightarrow E$	1
64-66	287	$D \leftrightarrow E$	1
65-66	34	$D \leftrightarrow D$	1
64,67-69	26	$D \leftrightarrow D$	1
71	24	$D \leftrightarrow D$	1
72-75	26	$D \leftrightarrow D$	1

The question arises as to whether these frequent encounter represent a loop in the polymer chain or is it the results of the 3D structure of the folded polymer? Since the 5C data is essentially taken over millions of cells it remains to show that these encounters (or loops) are indeed conserved elements of the Xic region.

1.3.5 Fitting the encounter data

In this and subsequent sections we work with the first replicate of the data.

To examine the behavior of the polymer as a whole, we fit for each bead's encounter data (row or column) a function of the form

$$\log(e_j(t)) = \log(At^{-B}) = \log(A) - B\log(t) \tag{1.1}$$

where t represents the distance in bead units to bead j, and $e_j(t)$ is the

normalized encounter data by distance for bead j, A and B are parameters to be determined by the fitting process.

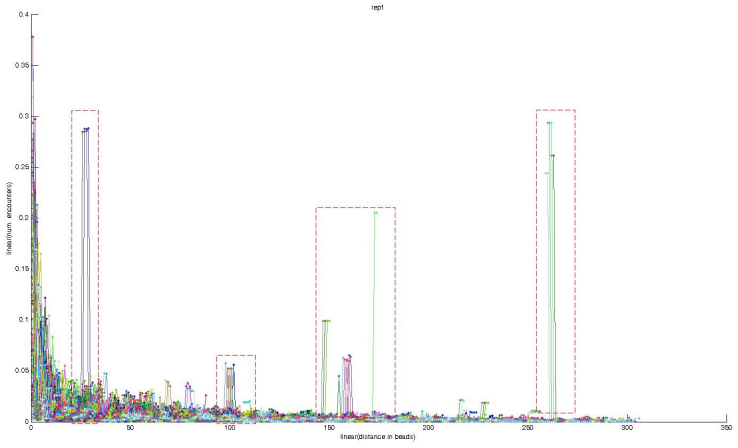


Figure 1.5: The encounter probability of each bead (Y axis) as a function of bead distance (X axis). The dashed red rectangles shows places of prominent non-nearest neighbors encounters

The fitted exponent (B) and bias (A) values obtained by minimizing the error in the model 1.1 are shown in Figure 1.6

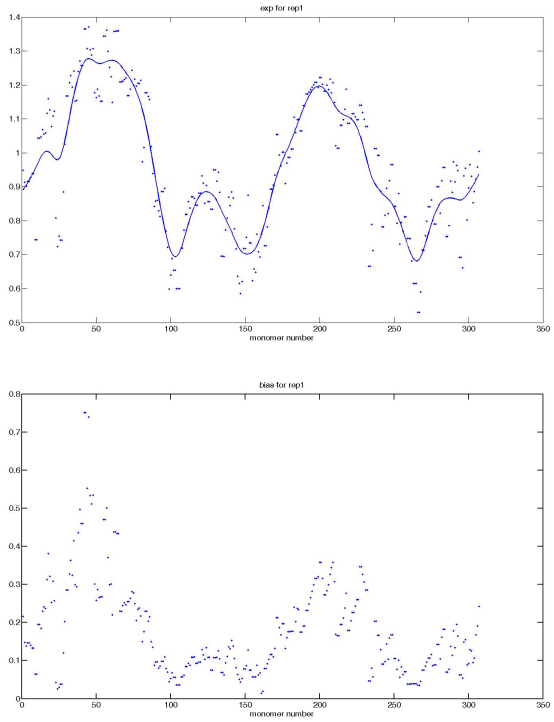


Figure 1.6: The fitted exponent (top) and bias (bottom) for each bead index. In the upper plot, the continuous curve represents a smoothing of the fitted exponent values using a smoothing spline with tolerance of 2 (see Matlab documentation on spaps function)

It is apparent from Figures 1.6 that there is some correlation between the fitted exponent values and the bias (indeed, since the integral of each function should be 1). To see it more clearly, we plot the fitted exponent and bias values on the same axes (Figure 1.7).

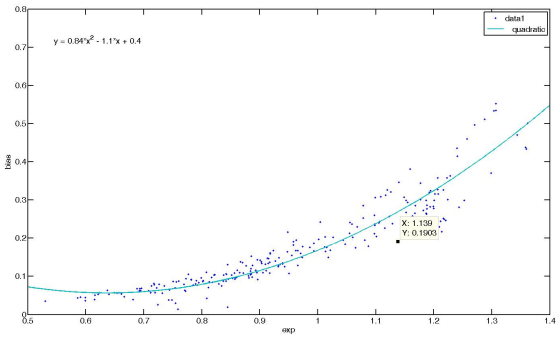


Figure 1.7: The fitted bias values are plotted as a function of fitted exponent values. The continuous cyan curve is a quadratic fit to the data. The equation that represents the fit appears in the top left corner

In Figure 1.6 top box, we see that the fitted exponent curve is somewhat tri-modal. With one peak located in roughly at bead 54, one at 125, and the left most at 200. The first and the last places correspond the mid points of the two TADs of the encounter.

From Figure 1.7 we see that the bias values tend to increase positively with the exponent. A quadratic curve was fitted to the fitted exponent vs. fitted bias data and resulted in values

$$B^* = 0.84A^2 - 1.1A + 0.4 \tag{1.2}$$

An important issue to address no is what is the validity of the fitted values? Comparing the actual bias values to the prediction of the fitted quadratic curve 1.2 we get the difference curve as appears in Figure 1.8. The mean difference between model predictions and model fit is 0.0274. We see that for the beads 35-50 there is a disagreement between prediction of model 1.2 and model 1.1 (green dashed box in Figure 1.8).

We therefore conclude that further in-depth evaluation of parameters is needed in specific regions of the polymer ot gain insight into the nature of the departure from the 'average' model fitted.

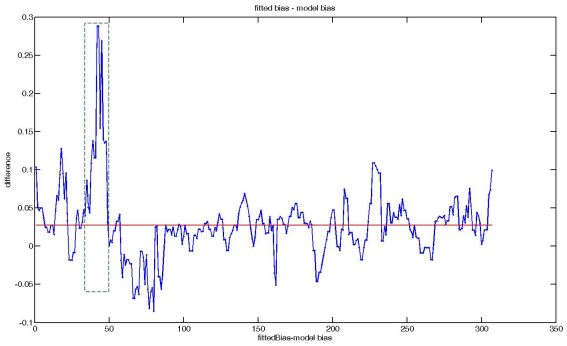


Figure 1.8: The difference between the fitted bias values B and the values predicted by model 1.2. The red line shows the mean of the difference and equals 0.0274. The green dashed rectangle shows the beads for which the model prediction and the data fitting do not agree. It corresponds to roughly the beads 35-50

1.3.6 Heterogeneity of the parameters over the beads

The question arises as to whether the beads behave in a similar manner in terms of parameter values (i.e obey the same model rules), and if not, would it be possible, just by assuming a model, to identify beads belonging to the same group, and elucidate the similarity or difference between groups of beads. We start by looking at adjacent beads properties according to the resulting parameter fitted. To simultaneously see the heterogeneity of the fitted parameters over all beads, we calculate the pairwise difference between parameters of the beads. the resulting matrices take the shape

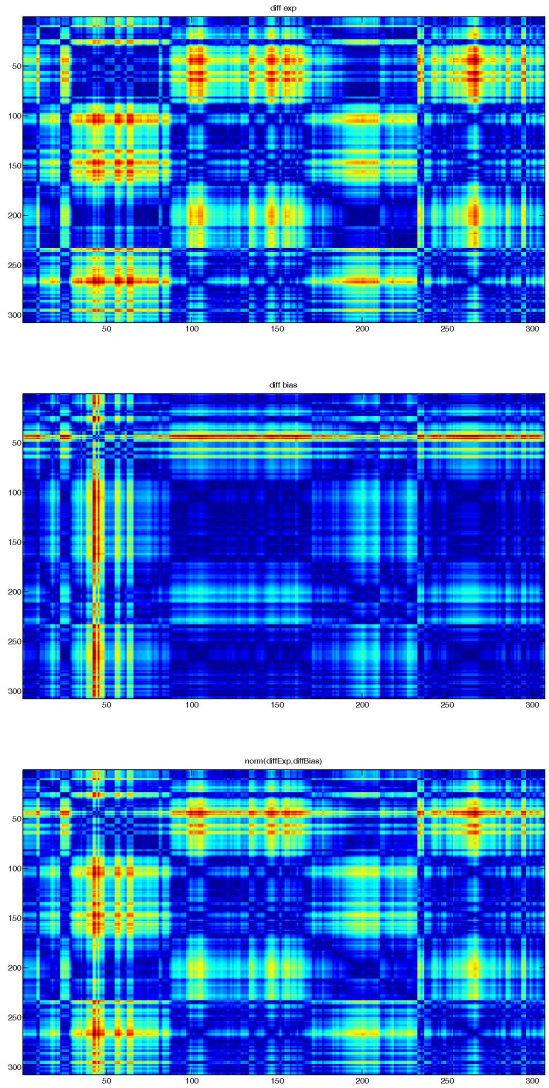


Figure 1.9: The norm of the difference between the fitted parameters from model 1.1 in a pairwise manner. the columns and rows represent bead numbers. The pairwise exponent difference norm (top), the pairwise difference norm for bias (middle); the norm of the two difference matrices (bottom)

It is now also apparent from Figure 1.9 that beads 35-50 have different bias values then expected. To find beads for which the exponent values agree but the bias values do not, we search for a threshold for the difference of exponents and bias below which it could be said that values are similar. For this end, we can consider the data as having 'low' and 'high' difference values.

We search for a values v_{exp} and v_{bias} for which the 'within' group variance is low and the 'between' group variance is the highest. Such approach leads naturally to employ the Otsu method. Using this method on the matrix, the value of v_{exp}, v_{bias} are found to be

$$v_{exp} = 0.262 \tag{1.3}$$

$$v_{bias} = 0.18 \tag{1.4}$$

1.4 Simulations

1.4.1 Initial settings

For the subsequent simulations, unless stated differently, we determine the simulation time step,Δt and the simulation time as follows. the simulation step Δt was determined such that it will prevent simulation 'blow-ups'

$$\Delta t = \frac{b^2}{12D}$$

Which is the values that stabilizes the numerical solution to the ode with no noise. when noise is present, this values can be increased.

The chain relaxation time is given by

$$\tau_p = \frac{b^2 \Delta t}{12 d^2 \sin^2(\frac{p\pi}{2N})}$$

where p is the mode number. For simulations we use the relaxation time of the slowest mode $p = 1$.

1.4.2 Verification of the validity of the output of the simulation framework

Here we make sure that the output simulation framework obeys the expected behavior of the Rouse polymer. A polymer of 64 beads, with no cross-linking is simulated for $40,000$ steps, the STD of link length was set to $b = 0.1$. The result is presented in Figure 1.10

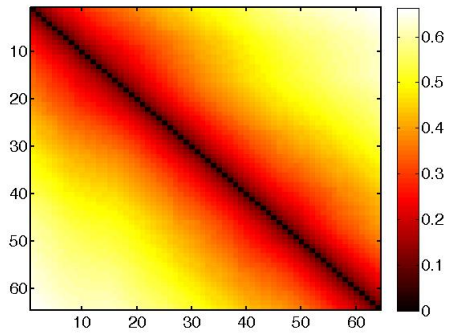


Figure 1.10: The mean bead distance matrix after 50,000 steps of the Rouse model

For an easier comparison of the simulation output with theoretical results, we work from now on with the Root-Mean-Square (RMS) distance between beads.

To see the expected increase in distance, we plot the cross-section of the RMS distance of the first bead to beads $2, 3, .., 64$. The distribution of distances between any two beads n and m of the Rouse polymer, have the following distribution

$$\Phi(R_n - R_m, n - m) = \left[\frac{3}{2\pi b^2 |n-m|}\right]^{3/2} \exp\left[-\frac{3(R_n - R_m)^2}{2|n-m|b^2}\right] \tag{1.5}$$

Where, R_n is the vector of coordinates of bead n. The equation above is the consequence of the fact that the sum of normally distributed random variables is again a normally distributed random variable. In addition, it can be shown that

$$< (R_n - R_m)^2 >= |n-m|b^2 \tag{1.6}$$

Using the result above we compare the RMS distance between bead 1 and beads $2, 3, .., 64$ with the function $|n-m|b^2$, appearing in equation 1.6, with $b = 0.1$. As can be seen from Figure 1.11, there is an excellent agreement between the output of the simulations and the theoretical results.

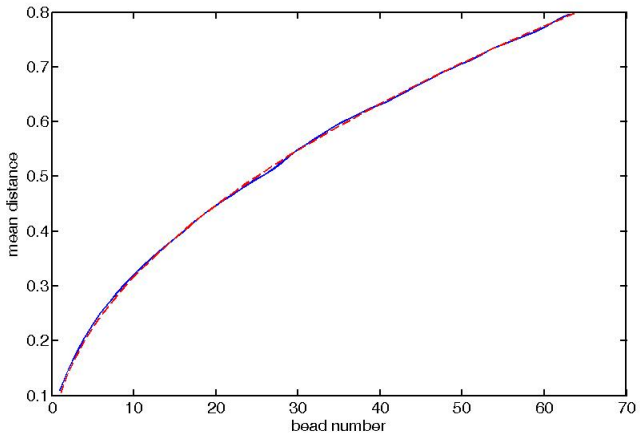


Figure 1.11: The RMS distance of bead 1 to all other 63 beads is plotted (blue curve) vs. the theoretical result (dashed red curve) $dist(1,n) = nb^2$. As can be seen, there is an excellent agreement between the curves, indicating that the simulation framework operates as expected

The RMS distance between adjacent bead was found to be 0.105, in agreement with the bond length $b = 0.1$. The mean square end-to-end distance was found to be 0.6319, in agreement with the expected 0.64. Overall we can conclude that the system operates as expected.

The encounter frequency

According to the theory, the encounter frequency of a bead m in a Rouse chain as a function of the distance of beads from it on the linear chain, should be proportional to $1/|m-n|$, where $n \neq m$ is the distance in bead units. To see that the simulation framework obeys this rule, the encounter frequency matrix was calculated. The encounter distance was set to 0.1, the matrix is presented in Figure 1.12. To see the decay of the encounter matrix, we plot the normalized encounter events of bead 1, and present the fit of the function

$$y = ax^{-b} \tag{1.7}$$

The encounter data was normalized prior to fitting by dividing it by the sum of encounter of bead 1. The parameters fitted were found to be $a = 0.2436$ and $b = 1.036$, in agreement with the expected $e(1) \propto 1/n$

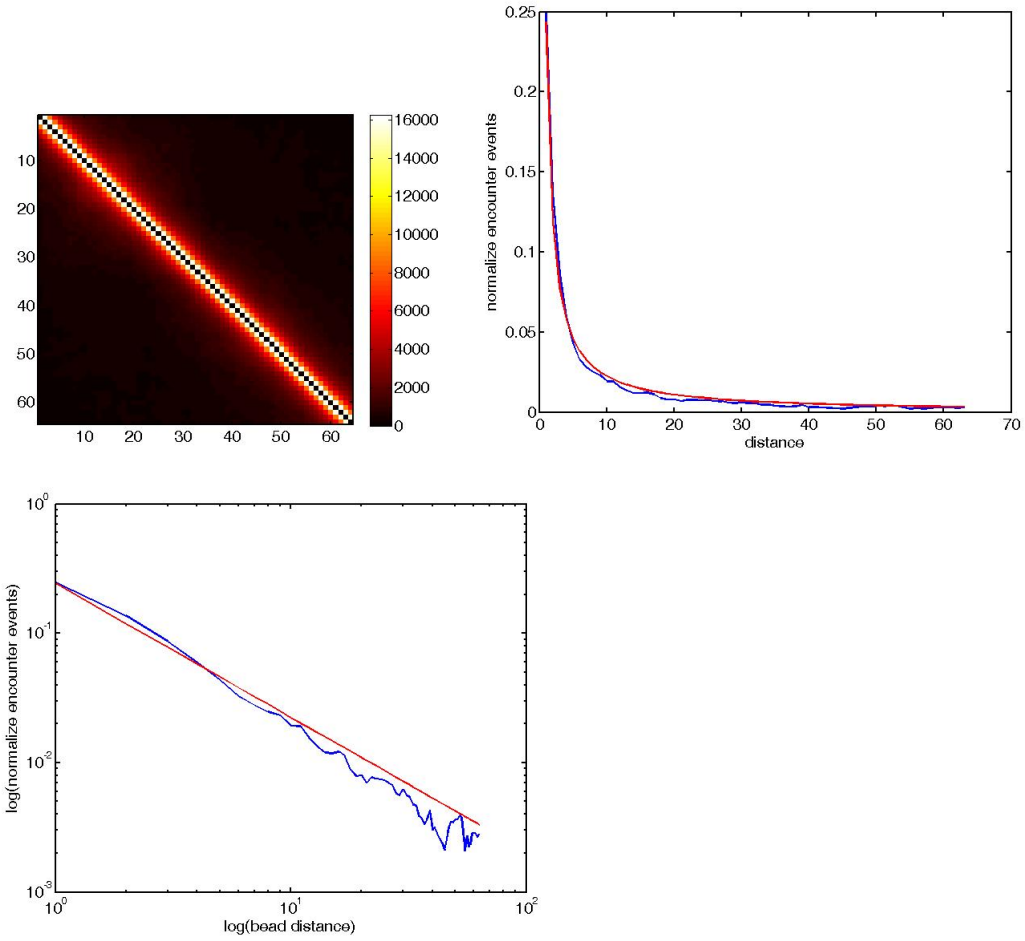


Figure 1.12: The encounter frequency and verification of theoretical results. The encounter matrix (top left) for a 64 beads Rouse chain. The decay of the rows of the encounter matrix is presented for the first bead (top right) for which we have the most number of beads (blue line) and a fit to the encounter curve (red line). the encounter frequency of bead 1 in a log-log scale (bottom left, blue curve) and a fit according to the theory (red line).

1.4.3 Simulating cross-linked polymer with a single link

Next, we test the behavior of the distance matrix for a 64 bead polymer for which bead 24 and 44 are cross-linked, creating a loop of size 20 beads. The simulation is run from relaxation time, a total of $50,000$ steps. The

relaxation time is determined according to the formula

$$\tau_p = \frac{b^2 \Delta t}{12 d^2 \sin^2(\frac{p\pi}{2N})} \tag{1.8}$$

For $\Delta t = 10^{-4}$, $p = 1$, $d = \sqrt{2D\Delta t}$, $N = 64$, we get

$$\tau_1 = 57.68 \tag{1.9}$$

minutes, which is the relaxation time of the first Rouse mode derived from the cross-correlation function.

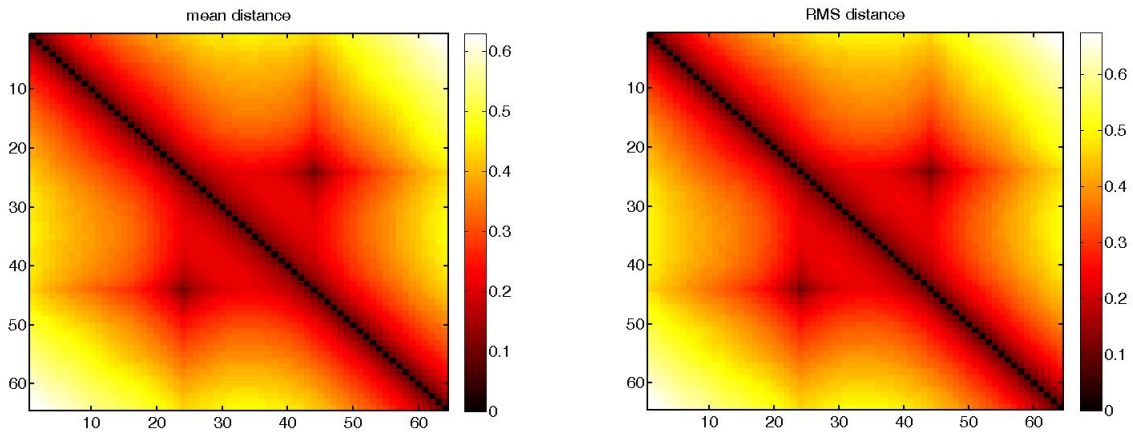


Figure 1.13: mean Distance matrix (left) and RMS distance matrix (right) over 50,000 steps of the simulation. Beads 24 and 44 are connected in a polymer of length 64 beads. As expected,the two matrices display similar qualitative behavior but with slightly different values

To see the decay of bead distance, we plot the cross-sections of the bead-distance matrix presented in Figure 1.13 at the rows 24 and 44. It

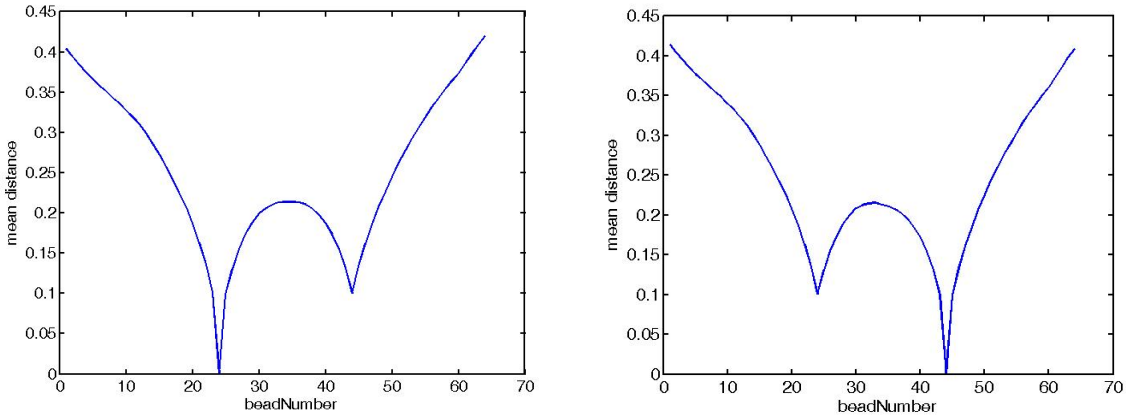


Figure 1.14: The cross section of the mean bead distance matrix presented in Figure 1.13 for bead 24 (left) and bead 44 (right). The height at point 44 (left panel) and point 24 (right panel) is the bond length $b = 0.1$ as expected.

Increasing the encounter distance

Next we examined whether changing the encounter length in the cross-linked chain can explain the appearance of TADs. The encounter frequency between any two beads of the chain is shown in Figure 1.15. The encounter length varied between half and twice of the connector length (0.1).

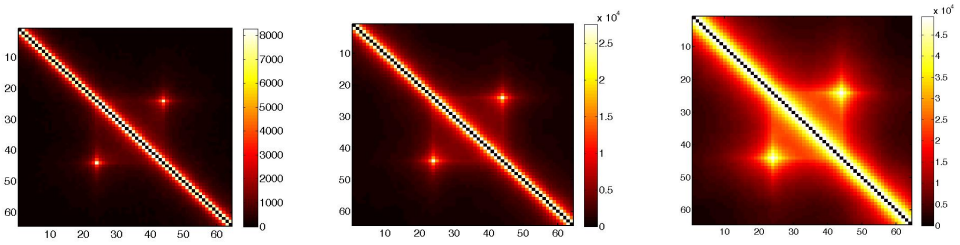


Figure 1.15: Encounter frequency matrices of a Rouse polymer with 64 beads with bead 24 and 44 connected by harmonic potential when varying the encounter distance from 0.05 (left), 0.1 (middle), and 0.2 (right). The bond length was set to $b = 0.1$, the encounter distance was set to 0.1. Simulation ran over 50,000 steps of 10^{-4} seconds. Each pixel represent bead pair (i, j) of the polymer.

The encounter histogram of bead 24 is displayed in Figure 1.16.

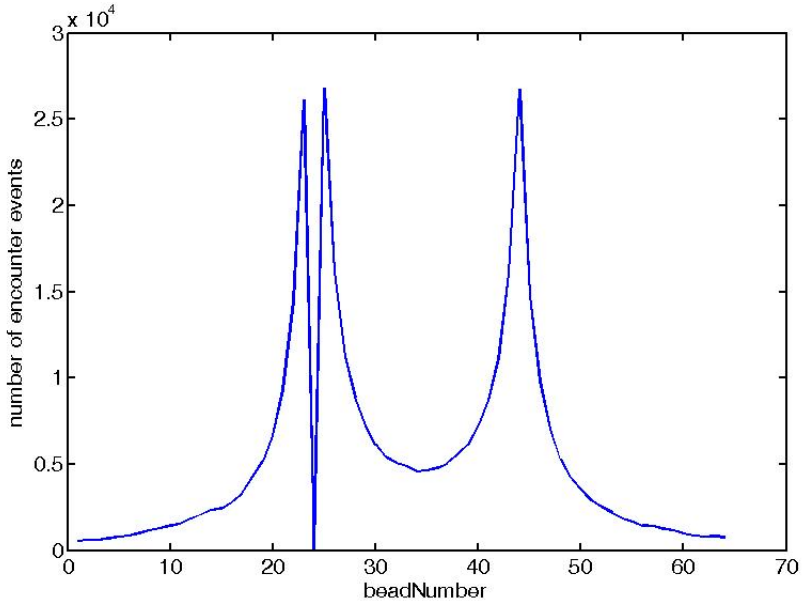


Figure 1.16: the encounter histogram of bead 24 as a function of bead number. The polymer included 64 beads with Kuhn length of 0.1, the encounter distance was set to be 0.1. the self encounters were trivially zeroed out

1.4.4 Simulating cross-linked polymer with subsequent cross-link

To explore the characteristics of the polymer dynamic system with one link, we construct a Rouse polymer with 32 beads, and subsequently link bead i to bead j, where $i = 1, 2, ...14$, $j = i + 2, i + 3, .., 30$. The time step was determined to be $\Delta t = 5 \times 10^{-4}$ sec, over a series of 3500 steps.

The eigenvalues and eigenvectors of the *mean* bead distance matrix were calculated. Eigenvalues were sorted in ascending order. In all simulation eigenvalues were negative except one, which took on a high positive value in relation to the other eigenvalues.

A characteristic pattern can be seen in Figure 1.17

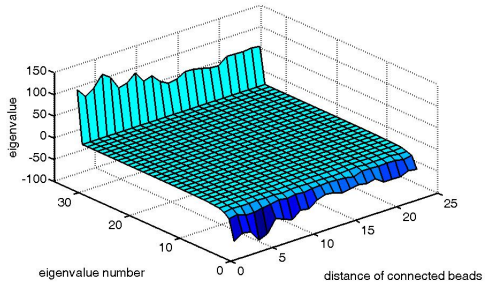


Figure 1.17: a surface representing the sorted eigenvalues

1.4.5 Simulation of variable loop number

1.4.6 One TAD

To examine the effect of increasing the number of stable loops on the encounter probability, a chain of 32 beads was constructed. The number of stable loops varied from 0 to 16. the bond length $b = 0.1$, the encounter distance was set to $b/2 = 0.05$, the time step was set to prevent simulation blow ups by the formula above, $\Delta t = 8.3 \times 10^{-4}$. For each number of loops, random pairs of beads were chosen to be connected.

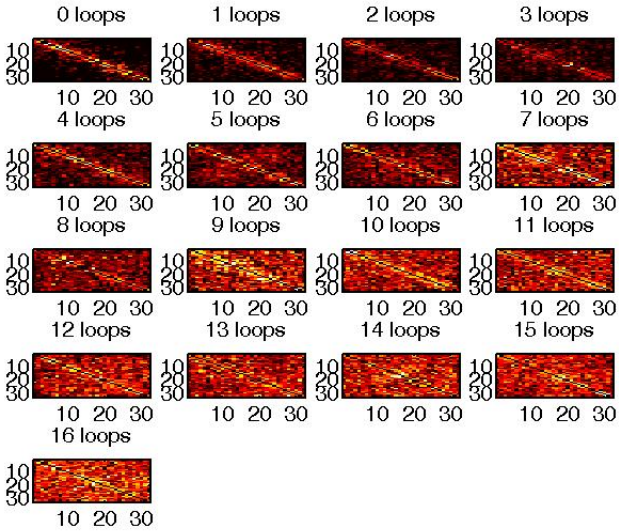
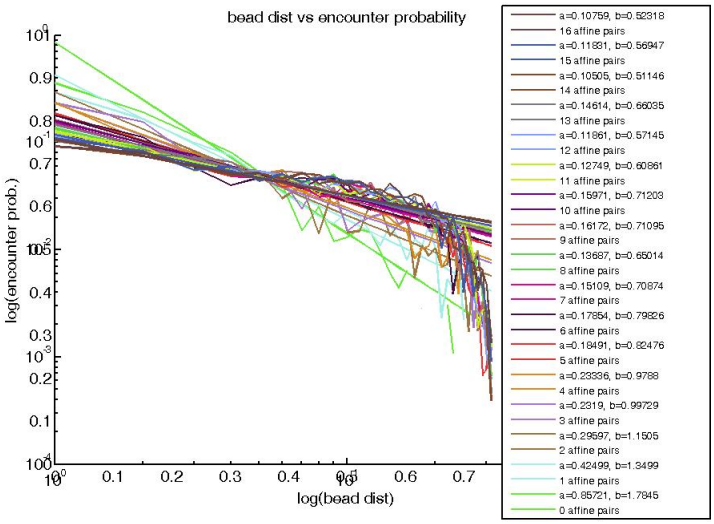
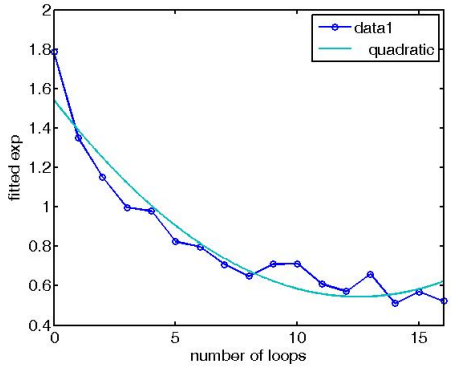


Figure 1.18: The encounter histogram of a chain of 32 beads. When the number of loops varies from 0 to 16 an area resembling a TAD emerges

for each choice of loop number, we fit a curve to the encounter probability as a function of distance.



The fitted exponent values are decreasing roughly quadratically.



One TAD with 'tail'

In this simulation we test the affect of the increase in the number of stable loops on the bead encounter probability. An ensemble of chains comprised of 64 beads were simulated to 1.2 times the relaxation time of the first Rouse mode, resulting in about $16,000$ steps for each chain. For each loop number 300 simulation were carried out. The bond length was set to be $b = 0.1$, the encounter distance $\epsilon = b/2$, time step was set to be $\delta t = 10^{-4}$. Stable loops were defined between bead 1 and 32 which we expect to resemble a TAD. At each simulation round we increase the number of loops L by choosing at random $2L$ bead indexes of bead to be connected, L is varied from 0 to 15. the encounter histogram shows an increasing encounter pattern in the region the loops are defined in (see Figure 1.19)

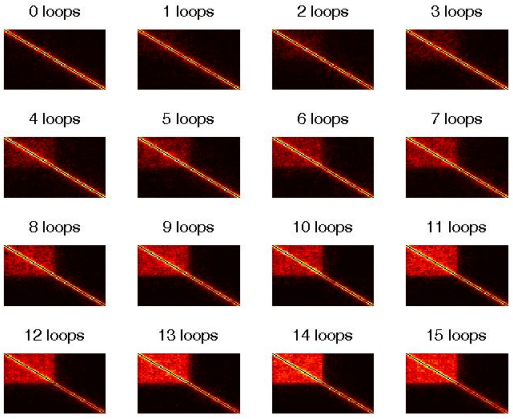


Figure 1.19: The affect of increasing the stable loops in a polymer of 64 beads. Loops were created at random between beads located in the first half of the polymer. Increasing the number of loops from 0 to 15 reveals a region resembling a TAD as seen in the data

The encounter probabilities for the chains when increasing the loops from 0 to 15, can be seen in Figure 1.20 For each encounter probability curve we fit a function of the form ad^{-b}. The change in the exponent b as a function of the number of loops is seen in Figure 1.21

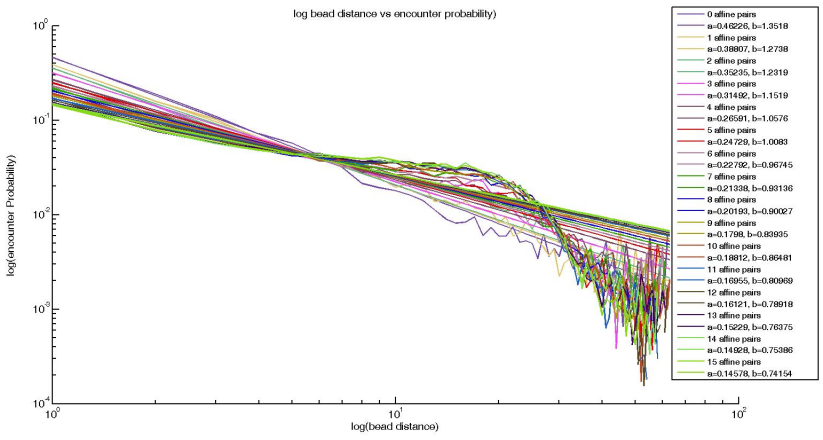


Figure 1.20: Encounter probability vs bead distance for the 16 models representing variable stable loops number if the first half of the chain of 64 beads (top)

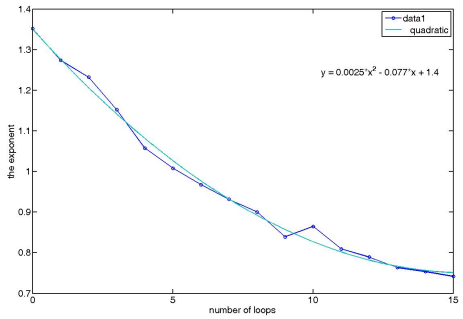


Figure 1.21: The change in the value of the fitted exponent to the encounter probability curves of models with increasing number of stable loops. The decrease in the values of b quadratic as can be seen by the cyan line fitted to the data.

Two TADs

1.4.7 Simulating dynamic loops model

It can be assumed that stiff fixed loops in the chromatin are not always stable, but can open and close transiently. We turn to examine whether a model with loops that can be formed, held fixed, for some period, and be released, can explain the appearance of the TADS. The rational behind is that the variability of cells used in the experiments some in different states, i,e. are found with different initial chromatin loops. When these cells are fixed and the chromatin is cross-linked, the cross-links that will form will depend on the proximity of the beads conditional on the loops present at the moment the experiment began.

A polymer with 128 beads was simulated. We define two region with high affinity of their member beads to a specific point in them. These regions can be considered as the TADs. Each region is of size 25 beads. The first region is defined to be from bead 1 to 25, the second from bead 77 to 102.

To simulate the state of experiment initiation, we define two beads in each TAD, which have high probability to be found together, and run simulations until steady state. Only one pair from each TAD was examined in each simulation, which we vary from simulation to simulation. At the end of each simulation we retrieve the encounters frequency for all the beads.

Beads that come in close proximity to one another, are considered attached, and can be released and some rate k_{off}. The attachment distance was defined as half of the connector length between adjacent beads.

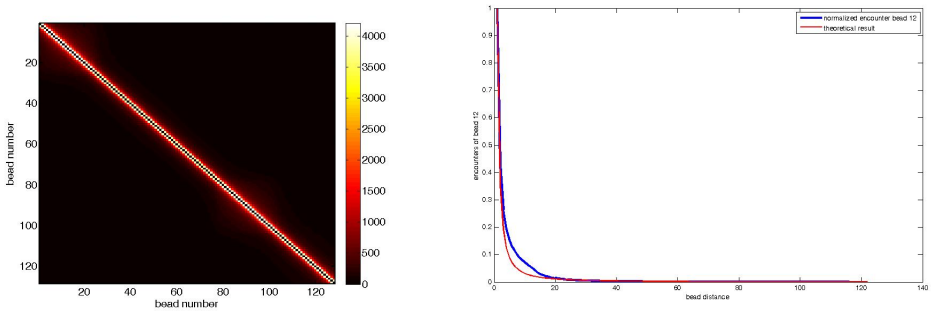


Figure 1.22: The mean encounter matrix for the transient loop model. the chain included 128 beads, with connector length of 0.1. Two regions were defined as the TADs: Bead 1 to bead 25, and bead 77 to 102. at each simulation, a pair of beads from each region were considered as having some affinity to one another. The model was simulated for each choice of affine pair until steady-state, 30000 steps. The mean encounter of all simulation in presented.

For beads that are not affine to any other, like the beads 26 to 77, the encounter frequency retrieved matches the theoretical result.

1.4.8 Simulating the 3C experiment

Next simulated the conditions of the 3C experiment to examine whether we can retrieve the TADs. The simplified version of the experiment is simulated, i.e. we take a Rouse chain comprised of 128 beads, we simulate the chain until steady-state, and we 'freeze' the picture. We then find all beads located within a distance ϵ from one another. At the end of the simulation we sum over all realizations and call it the encounter matrix.

The main difference between simulating a Rouse chain many times and performing this experiment is that in the the former, we simulate until steady state and from steady state onward we recored the encounter frequency at each step. The result is the encounter matrix of a *single* chain evolving in time. In this experiment, we simulate different chains, with different initial condition, and the end result is the sum of encounters of *different* chains.

The results show a strong deviation from the theoretical result. The encounter frequency of each bead decayed much slower than $N^{-1.5}$. Although this result can not explain the appearance of two different regions which have lower interactions among them, it can be form the basis for the understanding of the behavior of encounters in each one of the TADs. In later experiments we will try to change the architecture of the chain, making it heterogenic, by placing stiff boundaries between TADs.

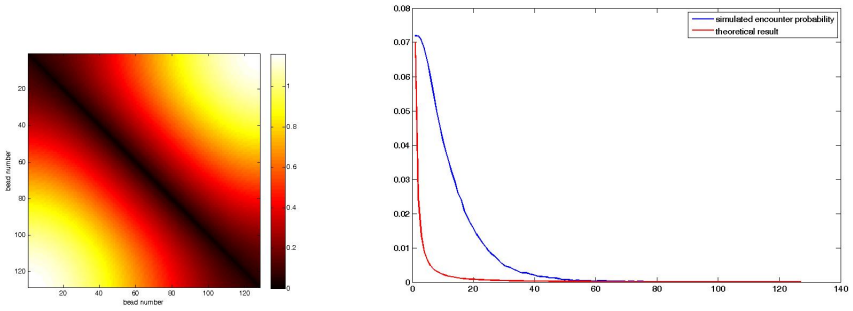


Figure 1.23: The encounter matrix and encounter probability profile for the simulated 3C experiment. The encounter matrix (left) still shows a decaying pattern of encounter.

1.4.9 Simulating dense regions

To verify that regions in the boundaries of the TADs are responsible for the interaction maps viewed in the article by Nora et al. We simulate a chain comprised of 120 beads. We cross-link a group of adjacent beads at two locations along the chain, the group 30-40 and the group 80-90. 3500 simulation steps are performed with a step size of $\Delta t = 5\dot{1}0^{-4}$ sec.

1.4.10 Simulating TAD D and E with cross-links corresponding to the peaks of 5C data

We next wanted to verify whether using a Rouse polymer and connecting beads for which the peaks of the experimental data, would allow us to reproduce the TADs.

A polymer with 307 beads was created, corresponding to the TAD D (107 beads) and E (200 beads). The cross-links were made at the mid point of bead numbers displayed in Table 1.3.4. Thirty simulations were performed and their average distance matrix after relaxation time is presented in Figure 1.24. As can be seen the location of the cross-links create 'hot spots' in the mean distance matrix. In comparison to Figure 1.1, we see that there is little resemblance. Thus concluding that creating cross-links according to the encounter matrix, is not sufficient to reproduce the TADs.

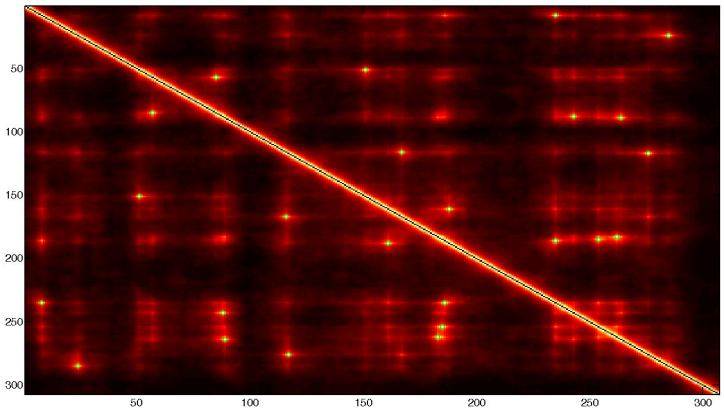


Figure 1.24: The mean of 30 simulation of the polymer with 307 beads, connected corresponding to the peaks of the encounter data (Luca et al). The small green circles represent locations of connected bead pairs. These circles overlap with peaks of the simulated encounter matrix. It is evident that the TAD structure was not reproduced with this approach. The x and y axes represent bead numbers. Each pixel represents the encounter number of bead i and j

1.4.11 Simulating the conditional encounter probability

In these simulation we empirically find what is the probability that a bead A meets a bead B before it meets bead C. For this end, a chain of 32 beads was constructed. Each simulation contained 5 times the relaxation steps, in a total of 150 simulations for each choice of bead n. We vary bead B, from 2 to 31.A was set to 1, C was set to 32.

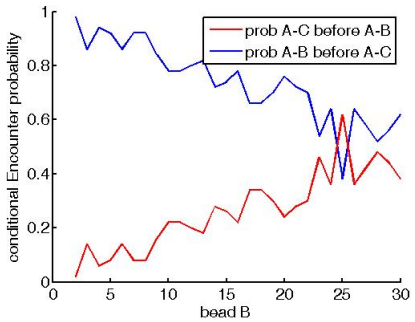


Figure 1.25: The conditional probability that bead 1 will meet bead 2 to 32

1.4.12 simulating the effect of increasing the encounter distance

In a chain of 32 beads, we increase the encounter distance and examine the difference in the encounter probability. The bond length was set to $b = 0.1$, the encounter distance was set to increase from $b/4$ to $2b$ with an increment of 0.01, a total of 18 simulation rounds. For each choice of encounter distance 50 simulation were preformed. simulations were ran to 5 times the relaxation time of the first Rouse mode.

Chapter 2

Review of Literature

Here we give several key points from articles bearing important principals from the understanding of chromosome internal interaction and organization

2.1 From Nora et.al 2012 [13]

- The X inactivation center (Xic) orchestrate the initiation of the X chromosome inactivation by controlling the expression of non-protein-coding Xist transcript.
- 5C techniques were used to analyze the *spatial* organization of the region including the Xist. The region analyzed is 4.5 Mb in size.
- 200 kb to 1 Mb regions were discovered and named TAD.
- TADs remain consistent between male and female Embryonic stem cells (ESC) and betwee different cell lines.
- TADs remain persistent before and after cell differentiation and before and after X inactivation. Internal TAD connections remain high even after X inactivation. However, the intra-TAD connection are lower than those of the same TAD on the active X.
- on the inactive X, the long range interaction($> 50kb$) inside the TAD are lost.
- disrupting the regions between TADs caused an ectopic chromosomal contacts

- 5C-Forward and 5C-Reverse oligonucleotides were designed to conduct test on undifferentiated mouse embryonic stem cells (ESCs)
- it was found that long range contacts (¿50kb) happen within a series of discrete genomic blocks , roughly 0.2-1 Mb in size
- The chromosomal domains are self-associating.
- FISH experiment on 3D chromosomal conformation were conducted to verify the results.
- in the FISH, the distance between probes lying in the same 5C domain were significantly shorter than of probes lying in different domains.
- a strong correlation was found between 5C count and 3D distance.
- bacterial artificial chromosome probes showed in experiments that large DNA segments belonging to the same 5C domain co-localize much more than DNA segment on adjacent domains *throughout cell cycle*.
- a correlation was shown between the existence of H3K27me and the TADs. However, for a mouse with a knockout of these proteins, a significant change was *not* observed in the composition and size of the TADs.
- it was shown that H3K27me - an antibody signifying the presence of Histone 3- is not related to the folding of the genomic regions into TADs.
- when the boundary between TAD E and D (which include the Xist and Tsix), ectopic folding of TAD E was observed. It was assumed that boundary elements have an influence on the conformation of the TADs.
- after removing boundary elements between TADs, the TADs did not merge completely, hence concluded that there are other, intra-TADs element that control conformation.
- TAD did not change during cell differentiation (from embryonic stem cell to neuronal progenitor and embryonic fibroblast), but the *internal* connections inside each TAD did change.

- for the TADs that did change, some have become Lamina Associated Domains (LADs)
- the transcription dynamics showed positive correlation for TADs for which the promoter of the genes was located inside the TAD.
- the correlation between gene expression of genes in the same TAD did not depend on the distance between genes in the TAD.

2.2 From Dekker et al 2013 [2]

1. This is a review of methods to analyze Chromosome Capture data and to infer mechanical properties of the DNA.
2. The 3C methods determines the encounter frequency of genomic segments, probably in the range of $10 - 100nm$
3. Chromosome are highly variable among cells [12]
4. There are some organizational principals at the scale of the whole nucleus
5. chromosomes occupy separate territories that do not usually mix[11]. When they do mix, they can potentially create functional interactions between loci on different chromosomes.
6. transcription does not occur diffusively throughout the nucleus, but happen at sub-nuclear sites enriched in components of the transcription and RNA processing machinery.
7. Therefore, actively transcribed genes tend to co-localize
8. transcriptionally inactive segments tend to associate with each other and can often be found localized at the nuclear periphery
9. sub-nuclear positioning of the loci is correlated with gene expression
10. imaging approaches do not allow analysis of the 3D folding of a complete genome at the kb resolution.

11. the 3C based methods do not distinguish functional from non-functional association of loci

12. the loci encounter frequency is affected by

    (a) direct loci encounter

    (b) encounters mediated by proteins

    (c) indirect co-localization to the same sub-nuclear structure (e.g lamina)

    (d) result of chromosome packing and folding

    (e) random encounters

    (f) interactions defined by the polymer physical characteristics

13. the 3D chromatin is highly variable even among identical cells[11].

14. analytical tools to interpret the 3c data focus first on specific point by-point looping interactions, e.g. between promoters and gene regulatory elements

15. in previous article, thousands of long range ($> 4mb$) interactions were identified between promoters and enhancer-like regions [14]

16. one abundant class of long range interaction involves promoters looping to sites bound by the insulator protein CTCF. The hypothesis is that these looping events have architectural role.

17. genes are regulated by multiple distal elements [5]

18. average pattern of looping interactions around promoters is asymmetric. Looping interaction are most frequently observed $\sim 120kn$ upstream

19. human and mouse genome are each composed of over 2000 TADs covering 90% of the genome.

20. TADs are defined by genetically encoded boundary elements [13].

21. boundary elements are enriched with CTCF-bound loci

22. the mechanisms that establish TAD boundaries are still unclear.

23. analysis of mouse genome suggests tat enhancer-promoter interactions are particularly frequent within TADs [15]

24. genes within the same TAD tend to be coordinately expressed during cell differentiation [13]. Possibly because those genes share the same set of regulatory elements.

25. the 3D structure of the Igh locus was inferred by polymer modeling and genome tagging to give discrete areas of looping [9].

26. the 3D structure of bacterial genome was determined by a combination of 5C data, imaging and modeling [17]

27. in yeast it was demonstrated that a small set of spatial constraints is sufficient to yield a highly organized genome structure, i.e volume exclusion [16]

28. HiC data shows lack of specific interactions for loci $> 1Mb$ apart

29. HiC data for non-synchronized human cells show 3 regimes, each exhibiting a power law decline in the contact probability $P(s) \sim s^{-\alpha}$. for $s < 0.7Mb$ $\alpha \approx 0.7$, corresponding to the TADs. For $0.7Mb < s < 10Mb$, $\alpha \approx 1$. For $s > 10Mb$, α is not well measures due to poor statistics [10].

30. a polymer model with excluded volume can give rise to an encounter probability with $\alpha = 1$

31.

2.3 The Hierarchy of the 3D genome. Gibcus 2013 [6]

1. neighboring chromosomes can overlap considerably

2. the nuclear envelope (double lipid bilayer) restricts genomic DNA to confined 3D space and provides a solid anchor point that allows for specific chromatin interactions.

3. The nuclear lamina is the inner nucleus membrane is associated with inactive heterochromatic chromatin directly or indirectly via lamina associated proteins.

4. almost half of the genome in a given cell population is composed of lamina associated domains [8].

5. activated genes move to the nuclear interior and inactive genes are found in the lamina associated domains

6. loci gather near sub-nuclear structures such as the nucleoli

Bibliography

[1] Elzo de Wit and Wouter de Laat. A decade of 3c technologies: insights into nuclear organization. *Genes & development*, 26(1):11–24, 2012.

[2] Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403, 2013.

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

[4] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309, 2006.

[5] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, et al. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, 2012.

[6] Johan H Gibcus and Job Dekker. The hierarchy of the 3d genome. *Molecular cell*, 49(5):773–782, 2013.

[7] 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, 2014.

[8] Lars Guelen, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B Faza, Wendy Talhout, Bert H Eussen, Annelies de Klein, Lodewyk

Wessels, Wouter de Laat, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, 2008.

[9] Suchit Jhunjhunwala, Menno C van Zelm, Mandy M Peak, Steve Cutchin, Roy Riblet, Jacques JM van Dongen, Frank G Grosveld, Tobias A Knoch, and Cornelis Murre. The 3d structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell*, 133(2):265–279, 2008.

[10] 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, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[11] WF Marshall, A Straight, JF Marko, J Swedlow, A Dernburg, A Belmont, AW Murray, DA Agard, and JW Sedat. Interphase chromosomes undergo constrained diffusional motion in living cells. *Current Biology*, 7(12):930–939, 1997.

[12] Iris Müller, Shelagh Boyle, Robert H Singer, Wendy A Bickmore, and Jonathan R Chubb. Stable morphology, but dynamic internal reorganisation, of interphase human chromosomes in living cells. *PloS one*, 5(7):e11560, 2010.

[13] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, 2012.

[14] Amartya Sanyal, Bryan R Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, 2012.

[15] Yin Shen, Feng Yue, David F McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V Lobanenkov, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2012.

[16] Harianto Tjong, Ke Gong, Lin Chen, and Frank Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome research*, 22(7):1295–1305, 2012.

[17] Mark A Umbarger, Esteban Toro, Matthew A Wright, Gregory J Porreca, Davide Bau, Sun-Hae Hong, Michael J Fero, Lihua J Zhu, Marc A Marti-Renom, Harley H McAdams, et al. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Molecular cell*, 44(2):252–264, 2011.