Comparison of 3-dimensional genome structures based on single cell Hi-C data

by

Moritz Gmeiner

Supervisor: PD Dr. Peter Virnau

Bachelor Thesis in Physics
presented to the faculty physics, mathematics, and computer science (FB 08)
of the Johannes Gutenberg University Mainz
June 30, 2022

1. Reviewer: PD Dr. Peter Virnau
2. Reviewer: Prof. Dr. Friederike Schmid

Ich versichere, dass ich die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Moritz Gmeiner Mainz, June 30, 2022

Moritz Gmeiner
KOMET
Institut für Physik
Staudingerweg 7
Johannes Gutenberg-Universität D-55099 Mainz
mgmeiner@students.uni-mainz.de

Contents

1. Introduction

1.1. DNA and the genome

DNA is the most important long-term information storage in biological systems. It is so fundamental that it is a shared property of all known life and even some non-living things like viruses[1]. In eukaryotes, most of the genome[2] is stored away in the cell's nucleus, a specialized compartment that is separated from the rest of the cell by the nuclear envelope. It is split up into multiple chromosomes, linear, i.e. non-circular, chains of DNA of varying length. The number of different chromosomes in a cell is dependent on the particular species, for example humans have 23 different chromosomes while mice, the primary subject of this work, have 20. Different species and cell types also vary in the number of copies of each chromosome they carry: haploid cells for example carry only a single copy of each chromosome, while diploid cells have two copies of each chromosome. To fit into the nucleus, the genome needs to be highly packaged: the length of the human genome for example is about $2\,\mathrm{m}$, while the cell nucleus is merely $5\text{-}8\ \mu\mathrm{m}$ in diameter. This packaging occurs at different levels, like for example nucleosomes and the 30-nm fibre. The question of how exactly DNA is being packaged is a very important research topic, as the genome structure is a factor in transcription initiation, the first step of making proteins from DNA information: By making certain regions of DNA more or less accessible, specific parts of the genome can either be activated or silenced, which is an important way of regulating gene expression. [1]

1.2. Hi-C

Hi-C is a method of determining the three-dimensional structure of an entire genome utilising a combination of proximity-based ligation and

[1] The question whether or not viruses are alive or not is somewhat dependent on the definition of life and has not been finally decided (and probably never will), but here they shall be considered non-living.

[2] The term "genome" typically refers to the entirety all of the DNA in the cell, including e.g. mitochondrial and plastid DNA. For simplicity, in this work the term genome will include only the DNA in the nucleus, as it is the main and for this work only relevant part.

DNA deep sequencing. It is a continuation of a number of previous chromosome conformation capture methods (3C, 4C, 5C) with the explicit goal of moving from targeting only specific genetic loci to obtaining a genome-wide map of interactions.

First, the DNA is cross-linked using formaldehyde, making connections between DNA strands that are in close proximity. The DNA is then digested by a restriction endonuclease that leaves a 5' staggered end, which is filled with biotin-marked residues and then ligated under conditions that favour ligation of those cross-linked DNA fragments. The DNA is sheared and selected for the biotin-marked cross-linked DNA fragments. These fragments are then deep sequenced and mapped back onto the genome, giving a list of contacts between positions on the genome.[2].

1.3. Simulation of genome structure

After obtaining a contact map of a particular genome using Hi-C, it can then be used to try to reconstruct the three-dimensional structure of the genome in the original physical cell that was mapped. Various approaches have been used for this reconstruction, including simulated annealing[3] and polymeric Gō-models[4]. One specific area of interest is the comparability of these reconstructed genomes structures across different cells of the same cell line, both as a measure of how well the model can reconstruct the original structure as well as to determine similarities and differences in the genome structure between the actual physical cells. In particular finding certain features that are shared across multiple cells might provide further insight in how exactly the folded structure of the genome influences transcription and, in reverse, how nature might utilise the structure of the genome as a mechanism to influence transcription. To this end, Hi-C data sets extracted from eight haploid mouse embryonic stem cells, obtained from (Stevens et al.[3]), were used to reconstruct the genome structure using a generic polymer model and then compared on both the genome level as well as the chromosomal level. For this first an analysis of the simulation results of each cell for itself has be made and compared to the results of (Wettermann et al.[4]). Then various statistical tools were used to compare the obtained structures of the different cells, first on an entire genome level and then on an individual chromosome level.

2. Simulation

2.1. Model and Simulation Protocol

The simulation protocol for the entire genome simulations was mostly carried over from (Wettermann et al.[4]). Additionally simulations of single chromosomes were carried out; the protocol for these simulations is identical to that of the entire genome, except only the beads of the chromosome in question were modelled. The simulation is a molecular dynamics simulation using the HOOMD-blue[5] toolkit. It utilises a Langevin integrator (`hoomd.md.integrate.langevin`) with a timestep of $\mathrm{dt} = 0.001$, a temperature of $\mathrm{kT} = 1.0$, and a drag coefficient of $\gamma = 1.0$. The neighbour list is a BVH tree neighbour list [6] [7] that was originally chosen as it scales with particle number as opposed to the system volume[4].

Each chromosome is modelled as beads on a string, where each bead represents a bin of $100{,}000\,\mathrm{bp}$. This is the same resolution as was chosen in (Wettermann et al.[4]) and represents a compromise between the resolution of the simulation result and the quality of the contact data that is available: a higher resolution, i.e. a smaller bin size of for example $40{,}000\,\mathrm{bp}$, would increase the resolution of the simulated structure and enable us to see smaller structures, but at the same time would spread the fixed number of contacts across a higher number of beads. This would make in particular the effect of having captured only a fraction of all possible contacts in the cell using Hi-C more prominent, which is estimated to be around 5% for each cell as seen in Table A.3. Vice versa, decreasing the bin size would help mitigate the partial capture of contacts, but limit the spatial resolution of the simulated structure. With this chosen resolution of $100{,}000\,\mathrm{bp}$ per bin the genome is represented by 20 chains varying in length between 500 and 2,000 beads each (the exact lengths for each chromosome can be found in Table A.1), or 25,714 beads in total. This does not represent the entirety of the mouse genome, whose length is approximately $2632\,\mathrm{Mbp}$[1], or 26,321 beads at a resolution of $100{,}000\,\mathrm{bp}$ per bead. The reason for this difference is that beads at the boundary of a chromosome that had no contact in any of the eight cells were dropped from the simulation, since their impact was assumed to be only negligible. Boundary beads that had contacts in some cells but not others were kept in the simulation of all cells in

[1]Mouse Genome Assembly GRCm39 from https://www.ncbi.nlm.nih.gov/grc/mouse/data, visited on 26.02.2022

order to keep the simulation data consistent across cells.

The model is based on a generic bead-spring polymer model in which three kinds of bonds are defined. The first two kinds of bonds are harmonic bonds between two beads of the general form

$$V(r) = \frac{1}{2}\kappa\,(r - r_0)^2$$

where κ is the force constant determining the stiffness of the bond, which is fixed at $\kappa = 2000$ for all harmonic bonds, and r_0 is the preferred bond distance. The first kind of harmonic bonds are the backbone bonds connecting adjacent beads in each chromosome; for these bonds the preferred distance is set to $r_0 = 1.0$. The other kind of harmonic bonds in the simulation are the predefined contacts derived from the Hi-C data set from (Stevens et al.[3]). Here the preferred bond distance is set a little larger compared to the backbone at $r_0 = 1.5$ in accordance with (Wettermann et al.[4]).

The third kind of bond is a Gaussian pair potential of the form

$$V(r) = \begin{cases} \varepsilon \exp\left[-\frac{1}{2}\left(\frac{r}{\sigma}\right)^2\right] & r < r_{cut} \\ 0 & r \geqslant r_{cut} \end{cases}$$

between all beads in the simulation designed to push all non-bonded beads away from each other. This potential is used in 2 forms in different parts of the simulation: a full form with $\sigma = 1.0$ and $r_{cut} = 3.5$ and a reduced form with $\sigma = 0.1$ and $r_{cut} = 0.4$; in both cases $\varepsilon = 100$. On one hand, this mimics the fact that at physiological conditions DNA molecules are negatively charged and thus repels each other. On the other hand it represents an excluded volume potential that pushes all beads away from each other, which is quite significant for the emergence of chromosomal territories[4]. An overview of all the potentials in the simulation can be seen in Figure 2.1.

The system is initialised by distributing all the beads randomly throughout the simulation box (uniform distribution, using `numpy.uniform.random`[8]). The bonds are set and the simulation is repeatedly cycled through the following steps:

- 80,000 time steps with no excluded volume potential
- 50,000 time steps with reduced volume potential
- 50,000 time steps with full excluded volume potential

Bonds and contacts are active at all of those steps. After each cycle the current state is saved to a gsd trajectory file. These saved states will be referred to in the following as **frames**. Frames whose trajectories are

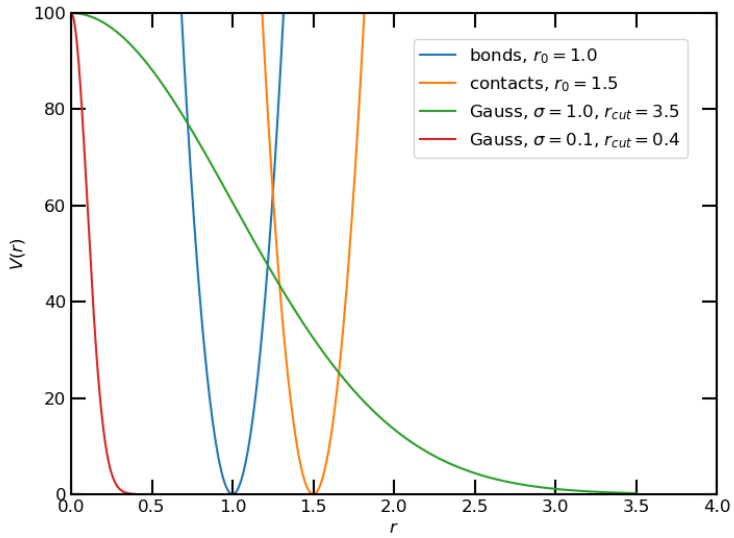

Figure 2.1.
Potentials used in the simulation. Bonds and contacts are harmonic potentials of the form $V(r) = 1000(r - r_0)^2$, with bonds having an r_0 of 1.0 and contacts having an r_0 of 1.5. Gauss potentials are of the form $V(r) = 100 \exp\left[-\frac{1}{2}\left(\frac{r}{\sigma}\right)^2\right]$ for r smaller than r_{cut} and 0 for r greater than the cutoff value r_{cut}.

similar are said to have the same **configuration**. These steps were repeated in each simulation for a total of 105 cycles. The first few cycles have to be discarded as the system takes some time to find its ground state, although certain problems can arise here that will be discussed later in 2.3.

2.2. Simulation results

Each simulation yields 105 sequential frames, i.e. the simulation state is not reset after each simulation cycle, but instead the final state of the last cycle is the initial state of the next cycle. This has the advantage of giving the system time to tune in, but also the disadvantage of the possibility that certain end configurations will never be reached in a particular simulation run after it has tuned in to a different locally minimal configuration. The potential energies of each frame for the simulation run of cell 2 can be seen in Figure 2.2. The first two frames show a potential energy significantly larger than the later ones, then the system quickly converges to a potential energy of about $7{,}950{,}000$ and shows only small deviations of less

than 1% of the mean. Thus both the length of the settling period and the potential energy of the ground configuration match (Wettermann et al.[4]) extremely well. To minimise the effect of the settling period the first 5 frames of each simulation run will generally be excluded in all subsequent analyses where the data is combined over all frames such as averages or standard deviations.

Furthermore, Figure 2.3 shows the distance distribution of the bonds and predetermined contacts combined across all frames of the simulation of cell 2. The results are again very similar to (Wettermann et al.[4]), with both peaks and means shifted slightly to the right of the respective preferred bond length of 1.0 and 1.5. The 99.73th quartile is at 1.71 for bonds and 2.42 for contacts, showing that a substantial portion of the bonds and predetermined contacts are enforced reasonably well.

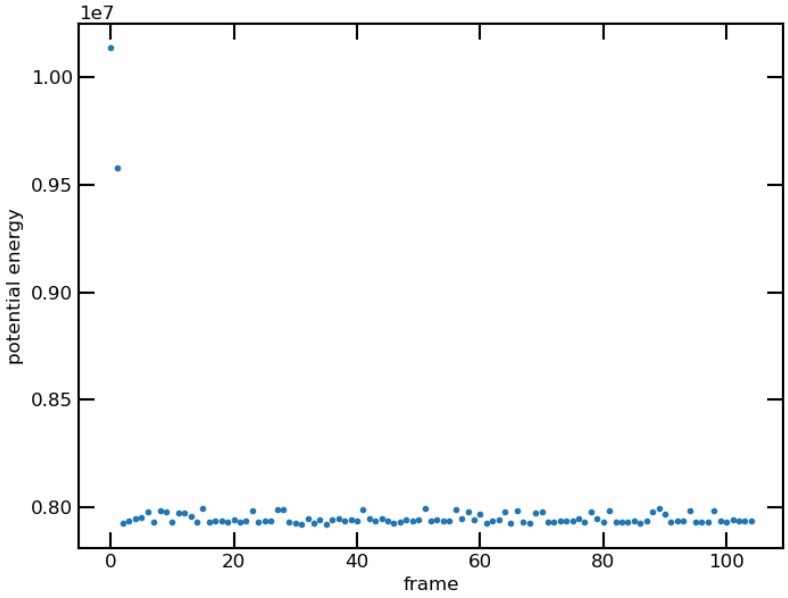

Figure 2.2.
Potential energy of each frame in the simulation of cell 2.

For the other cells the situation is generally similar, except for cell 1 and cell 5, which will be discussed more in-depth in 2.3. The coefficients of variation, i.e. the standard deviations divided by the means, of potential energies are between 0.3% and 1.8%, showing the minimum energy configuration to be quite stable in these cells. Also neither bonds nor contacts are overly overstretched in any cells, including cell 1 and cell 5. A complete overview of potential energy coefficients of variation and means

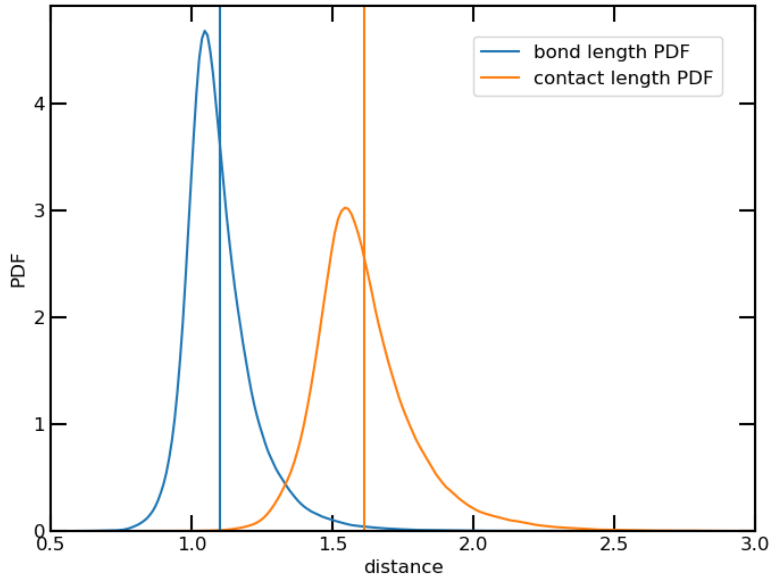

Figure 2.3.
Distance distributions for bonds and predetermined contacts over all frames in the simulation of cell 2 with the means marked by a vertical line. The mean of the bond distances is at 1.10 and the mean of the contact distances is at 1.61.

and 99.73th percentiles for bonds and predefined contacts can be found in Table A.2.

Renderings of all simulated cells can be seen in Appendix C. A few things can be notes from these images visually. First, most simulated genomes have a quite regular, spherical shape, but in particular cell 3, cell 8, and the higher energy configuration of cell 5 (which will be discussed in more detail in 2.3) show clear differences from this. Cell 3 and cell 8 have a more elongated, bean-like shape, as can be seen in Figure C.4 for cell 3 and in Figure C.12 for cell 8. The higher energy configuration of cell 5 on the other hand has a more obloid, donut-like shape. Also notable is the fact that cell 6 and cell 7 are hollow as can be seen for example in Figure C.10 for cell 6 and Figure C.11 for cell 7. More renderings of all simulated cells can be seen in Appendix C.

2.3. Problems with the simulation

While the simulations of most cells quickly reached a stable ground state configuration, the simulations of cell 1 and of cell 5 showed significant deviation from the expected results, warranting further investigation.

2.3.1. Cell 1

In cell 1, a ground state is reached very quickly, but this ground state is very unstable, as indicated by the comparably high coefficient of variation of the potential energy of 8.51% in Table A.2 and seen in Figure 2.4. The RMSDs of all frames with respect to the last frame as seen in Figure 2.5 shows that some of these frames have rather similar configurations while others are very different. One hypothesis would be that some frames represent a ground state configuration while others are higher energy states. This can be tested by selecting for all frames with low energy, defined by being below some cutoff energy, and checking how the RMSDs for those low-energy frames behaves. A cutoff energy of 1.495×10^7, visually displayed in Figure 2.4 as the orange line, was chosen as it captures most of the frames that can be identified visually as being low energy, while excluding all that deviate strongly from this energy baseline. This selects a total of 52 frames, which have been marked in Figure 2.5 by red dot. As can be seen very clearly, these low energy frames do in fact have a low RMSD of 1.00 ± 0.01 with respect to the last frame. This confirms that while it is very unstable, this simulation of cell 1 does in fact have a ground state configuration, and it can be filtered for by selecting for those frames with a low energy. To check if this ground state instability is an intrinsic property of the Hi-C contact data for cell 1 or merely a random artefact of this particular simulation run, cell 1 was simulated two more times. The potential energies for those repeated simulations can be seen in Figure A.1 and Figure A.2 in the appendix and clearly show the same pattern of instability, leading to the conclusion that this instability is in fact a consequence of the predetermined contacts for cell 1.

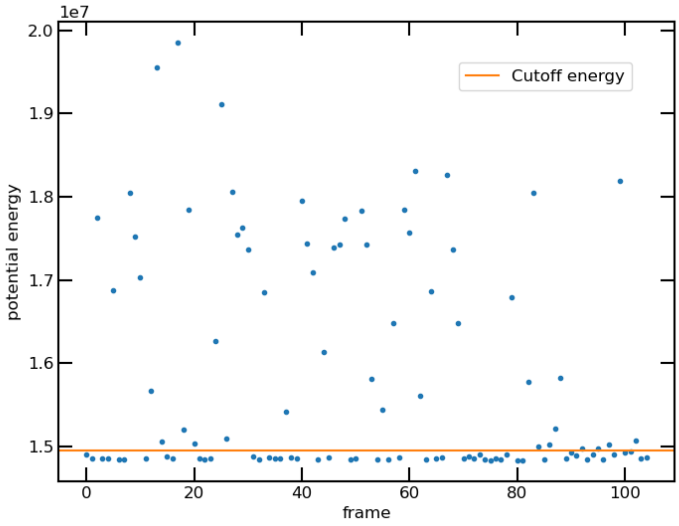

Figure 2.4.
Potential energies of all frames in the simulation of cell 1. Cutoff energy is set at 1.495×10^7, with all frames lower than this threshold being defined as "low-energy frames".

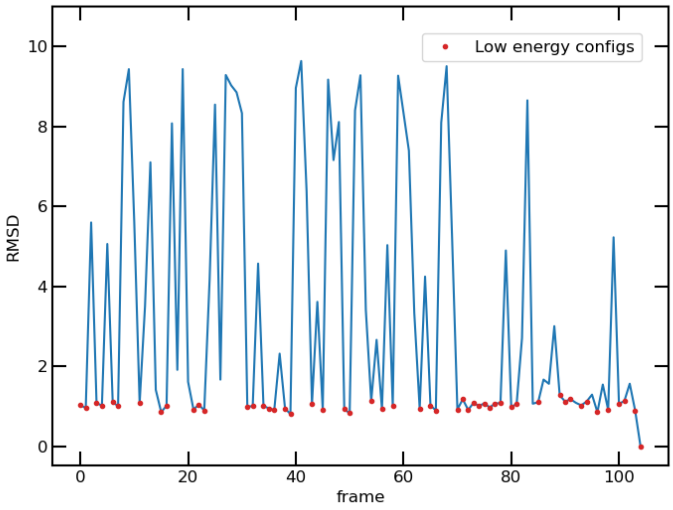

Figure 2.5.
RMSDs of cell 1 with respect to last frame. Low-energy frames, as defined in Figure 2.4 to be frames with a potential energy lower than 1.495×10^7, have their RMSD marked by a red dot.

2.3.2. Cell 5

Figure 2.6 shows the potential energies of all frames in the simulation of cell 5 in blue. As can be seen, after the zeroth frame, which is of higher energy as is typical for the transient phase, the potential energy drops to and stabilises around a value of $(9.805 \pm 0.043) \times 10^6$ for frames 1 through 38. But then for frame 39 and the following ones until the end of the simulation, the energy drops down again to a significantly lower value of $(8.220 \pm 0.008) \times 10^6$. As seen in Figure 2.6, this is a reduction in both the potential energy itself as well as the deviation in potential energy, from a coefficient of variation of 0.5% down to 0.1%, signifying that the second configuration is both energetically more favourable as well as very stable. The fact that the first configuration was held for 38 frames though means that the existance of semi-stable configurations apart from the ground state configuration is possible. To examine this effect further, cell 5 was simulated 2 more times. The resulting potential energies, together with the potential energies of the first simulation, can be seen in Figure 2.6. Clearly the second and third simulation run show the same pattern that most other cells exhibit: the first one or two frames have a higher potential energy, and then it drops into a stable ground state configuration where it remains for the rest of the simulation. The very similar potential energies of the ground state configurations in all three simulations suggests that these ground state configurations are in fact the same across all simulations. An analysis of the RMSDs of each frame in all three simulations from the last frame of the first simulation can be seen in Figure 2.7 and clearly confirms that the ground states of all three simulations are in fact the same one. This means that while the simulation of cell 5 does have a single stable ground state configuration, it also has a semi-stable configuration that arises in some, but not all simulation runs.

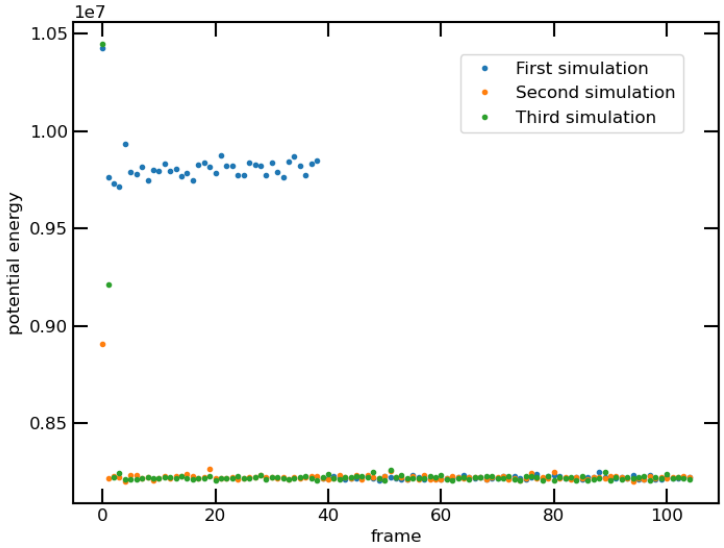

Figure 2.6.
Potential energies of all frames in all three simulations of cell *5*.

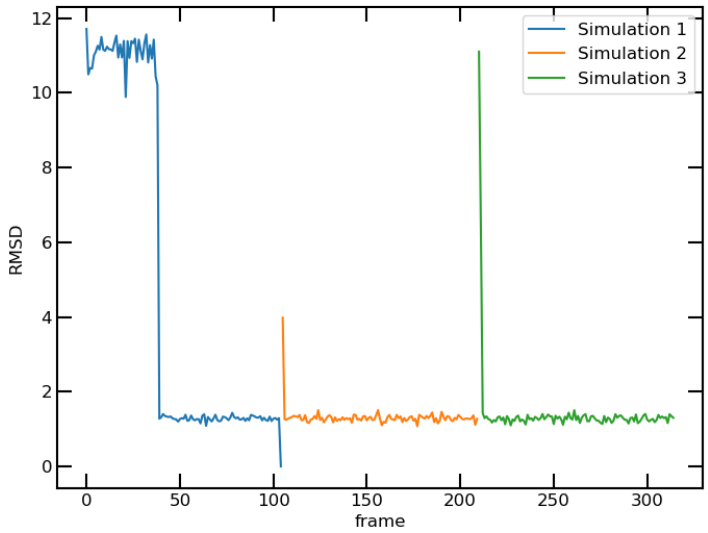

Figure 2.7.
RMSD of all three simulations of cell *5* with respect to the last frame of the first simulation.

3. Comparison of Cells

3.1. Definition of SCC

HiCRep is a mathematical tool introduced in (Yang et al.[9]) for the explicit purpose of comparing Hi-C data sets. It takes as input the contact matrices of two Hi-C data sets (or, more generally, contact matrices of any kind) and outputs a correlation coefficient, that is a number between -1 and 1, with larger numbers signifying a stronger similarity between the data sets and vice versa. This is done in a two-step process: first both contact matrices are smoothed to counteract binning-associated problems and then a stratum-adjusted correlation coefficient (SCC) is calculated between these smoothed contact matrices. The exact procedure will be explained in the following section.

3.1.1. Smoothing

Both contact maps are first smoothed using a square uniform filter of width $2h + 1$ for a chosen smoothing parameter $h \geqslant 0$. This helps compensating a lack of coverage, that is the fact that not all physical contacts are contained in the contact matrices, a common problem for Hi-C data. Mathematically this filter can be defined as

$$X_{ij} = \frac{\sum_{k=i-h}^{i+h} \sum_{l=j-h}^{j+h} C_{kl}}{2h+1}$$

where C is the $n \times n$ raw contact matrix and X is the $n \times n$ smoothed contact matrix. C_{ij} is defined to be 0 for either i or j not in $\{1, \ldots, n\}$.

The uniform filter might seem like an unusual choice compared to other more sophisticated filters, but it has the great advantage of having the representation $X = L \cdot C \cdot R$, where X is the smoothed matrix, C is the raw contact matrix, and L and R are upper and lower triangular matrices respectively. This is useful for computational purposes, especially when using sparse representations of C and X, which is recommended since contact matrices generally can be quite large, $25{,}714 \times 25{,}714$ in this case, but are mostly empty.

h is a parameter for the SCC algorithm and thus has to be chosen appropriately. The HiCRep package includes a function called `htrain`

(`h_train` in the HiCRep.py Python package) that tries to estimate an appropriate h-value heuristically. For a resolution of $100\,\mathrm{kbp}$ an example value of $h = 3$ is given in the original HiCRep paper, which should be kept in mind as a refence when trying to choose an h-value later.

3.1.2. SCC

The stratum-adjusted correlation coefficient aims to be a measure of correlation between two random variables X and Y that are stratified by a third variable into K strata $X_1, \ldots, X_K$ and $Y_1, \ldots, Y_K$ respectively. In each stratum there are the stratified random variables (X_k, Y_k) with N_k observations $(x_{k,1}, y_{k,1}), \ldots, (x_{k,N_k}, y_{k,N_k})$ each. The pearson correlation coefficient between X and Y for the k-th stratum is given by

$$\rho_k = \frac{\mathrm{Cov}(X_k, Y_k)}{\sqrt{\mathrm{Var}(X_k)\mathrm{Var}(Y_k)}} = \frac{\sum_{i=1}^{N_k}(x_{k,i} - \bar{x}_k)(y_{k,i} - \bar{y}_k)}{\sqrt{\sum_{i=1}^{N_k} x_{k,i} - \bar{x}_k}\sqrt{\sum_{i=1}^{N_k} y_{k,i} - \bar{y}_k}}$$

The SCC is the weighted average of the pearson correlation coefficients of the different strata:

$$\rho_s = \sum_{k=1}^{K} w_k \rho_k$$

where the weights w_k are given by

$$w_k = \sqrt{\mathrm{Var}\left(\frac{\mathrm{Rank}(X_k)}{N_k}\right)\mathrm{Var}\left(\frac{\mathrm{Rank}(Y_k)}{N_k}\right)}$$

with $\mathrm{Rank}(X_k)$ and $\mathrm{Rank}(Y_k)$ being the ranked variables[1] of X_k and Y_k respectively. For a thorough derivation of the SCC see the original HiCRep paper[9], Section "Derivation of stratum-adjusted correlation coefficient (SCC)".

3.2. SCC between cells

In Figure A.3 the SCC between the original Hi-C contact matrices and the contact matrices of the simulation can be seen for each cell. Two beads in the simulation were defined to be in contact if their distance is less than 3.0 and contacts were counted across all frames. For all cells the SCC is between 0.68 and 0.80. This is sensible since on one hand, the SCC is expected to be high as the simulated data is based on the Hi-C data, on the other hand it is not surprising that the scores are not perfect since the Hi-C

[1] https://en.wikipedia.org/wiki/Ranking#Ranking_in_statistics

data doesn't cover all contacts in the real genome whereas the simulation data includes all contacts in the simulated genome. Particularly interesting is that the SCCs for cell 1 and cell 5 are very much in line with those of the other cells, regardless of the problems that have arisen during the simulation and were discussed in 2.3.1 and 2.3.2 respectively. This could indicate either that regardless of their problems cell 1 and cell 5 replicated the contacts or their Hi-C data as well or that all simulations replicated the Hi-C data badly so that there is no recognisable difference. This question can unfortunately not be answered here.

The pairwise SCC between the original Hi-C data of all cells can be seen in Table 3.1. It is immediately very clear that all SCC values (except for those of a cell with itself) are very low, especially compared to the values of 0.7 to 1.0 obtained in the original HiCRep paper for hESC (human embryonic stem cells) and IMR90 (human lung fibroblast cells) cell lines (Figure 3A in [9]). This might very likely be related to the fact that each of the Hi-C data sets that have been used for this analysis captured only an estimated 5% of all physical contacts in the actual cell, as detailed in Table A.3, but regardless of the reasons, it sets the expectations for similarity between the simulated cells quite low. The SCC for the contact matrices of the simulated genomes can be seen in Table 3.2 and, as was expected, are similarly low to those of the initial Hi-C contact matrices. This strongly suggests that the structure of the simulated genomes have very little relation with each other. This assumption is also backed by the rendered images in C that show the simulated genomes differ quite strongly in shape, e.g. with some being spherical and others being bean-shaped or obloid, or some of them having some rather big holes inside them making them basically hollow. In summary, the low SCCs are a first strong indicator that the global structure of the different simulated genomes are significantly different.

Table 3.1.
SCC calculated pairwise between the contact matrices of the original Hi-C data. Contacts are binned to a size of 100,000 bp per bin. Multiple contacts between the same bins are only counted once. The smoothing parameter h is set to 7. The SCC between a cell and itself is always 1.0 and thus omitted.

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8
Cell 1	-	0.132	0.093	0.116	0.128	0.105	0.136	0.092
Cell 2	0.132	-	0.098	0.104	0.147	0.116	0.150	0.138
Cell 3	0.093	0.098	-	0.062	0.104	0.093	0.101	0.092
Cell 4	0.116	0.104	0.062	-	0.108	0.101	0.123	0.077
Cell 5	0.128	0.147	0.104	0.108	-	0.121	0.148	0.123
Cell 6	0.105	0.116	0.093	0.101	0.121	-	0.145	0.097
Cell 7	0.136	0.150	0.101	0.123	0.148	0.145	-	0.116
Cell 8	0.092	0.138	0.092	0.077	0.123	0.097	0.116	-

Table 3.2.
SCC calculated pairwise between the contact matrices of the simulated structures. Two beads are defined to be in contact if their distance from the centre is less than 3. The smoothing parameter h is set to 7. The SCC between a cell and itself is always 1.0 and thus omitted.

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8
Cell 1	-	0.171	0.149	0.114	0.173	0.141	0.178	0.104
Cell 2	0.171	-	0.164	0.141	0.183	0.184	0.215	0.184
Cell 3	0.149	0.164	-	0.109	0.161	0.124	0.145	0.122
Cell 4	0.114	0.141	0.109	-	0.135	0.105	0.140	0.098
Cell 5	0.173	0.183	0.161	0.135	-	0.148	0.194	0.143
Cell 6	0.141	0.184	0.124	0.105	0.148	-	0.191	0.116
Cell 7	0.178	0.215	0.145	0.140	0.194	0.191	-	0.157
Cell 8	0.104	0.184	0.122	0.098	0.143	0.116	0.157	-

3.3. RMSD between cells

Another way to gauge the similarity of the configurations between the simulated genomes is to calculate RMSDs between the configurations of the cells. There are generally two straightforward approaches that can be taken when calculating RMSDs between the different cells: first, a reference frame can be chosen for each cell and the RMSDs between these reference frames can be calculated. While this approach is very direct, it does strongly depend on the choice of a good reference frame and is generally much more susceptible to statistical variation between the different frames of a simulation run. The other approach would be to calculate an average trajectory from all (or, as will be discussed, only certain selected) frames in a cell and then calculate the RMSDs between those average trajectories. While this counteracts the statistical problems of using a reference frame, it has a challenge of its own: it only makes sense to take the average of frames that represent the same configuration, i.e. that are already similar to each other. In particular, this not the case for example for the simulation of cell 5, which contains, as discussed in 2.3.2, two configurations, one ground state configuration and one of higher energy, or similarly the simulation of cell 1 which contains besides the ground state frames a number of higher energy frames. To calculate average trajectories, it needs to be ensured that the frames that are averaged over are sufficiently similar and all represent the ground configuration while carefully filtering out the rest.

For cell 1 the same energy-filtering approach was chosen as in 2.3.1 to select only those frames with ground state energy. For cell 5 only frames 39 and onward were included as those represent the ground state frames of the simulation. For the other cells simply the first 5 frames were excluded to avoid the tune-in period as discussed in 2.2. The resulting RMSDs of each frame to both a reference frame, here the last frame of each simulation, and the average trajectory is shown in Figure A.4. There are two major things than can be learned from Figure A.4: first, the RMSD from average is quite consistently smaller than the RMSD from last, implying the averaged trajectories fulfil their purpose as representations of the ground state configuration well. Second, for all cells except cell 1 and cell 5 the mean of the RMSDs from the average trajectory seen in Table 3.3 is between 0.9 and 1.9, with standard deviations mostly between 0.08 and 0.23, with the only exception being cell 4 with a slightly higher standard deviation of 0.66. For cell 1 and cell 5 the raw means and standard deviations are significantly larger due to the presence of the non ground configuration frames, but after applying the filtering from above the means drop to 0.68 ± 0.10 and 0.88 ± 0.07 respectively, although this is expected as precisely those frames get filtered out that showed strong deviation from the ground state configuration the averaged trajectory is representing. The

unexpected outlier here is cell 4 with a mean RMSDs to average of 1.55 ± 0.66 and several pronounced spikes seen in Figure A.4. Looking at the potential energies of the frames with spikes shows that those frames have somewhat elevated potential energy compared to the other frames. Filtering out the nine most promimently spiked frames improves the mean RMSD to average of cell 4 to 1.36 ± 0.19, making it fall more in line with the other cells. This improvement is, especially for the standard deviation, significant enough to use this filtering from here on.

Table 3.3.
Mean and standard deviation of the RMSDs of each frame to the average trajectory for each cell simulation.

Cell	1	2	3	4	5	6	7	8
RMSD Mean	3.54	0.90	0.95	1.55	5.07	1.37	1.88	1.78
RMSD Std	2.46	0.16	0.23	0.66	2.58	0.21	0.12	0.08

Using the filterings described above to generate the average trajectory for each cell, the RMSDs between each of them can be seen in Table 3.4. Except for the self-RMSDs they are all between 18.4 and 25.0, implying the average whole genome trajectories for each cell simulation are significantly different, affirming the same conclusion made in 3.2 using the genome level SCC.

Table 3.4.
RMSDs between average trajectories of each cell. Each average trajectory is generated from the ground state frames of the simulations. The RMSD between a cell and itself is always 0 and thus omitted.

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8
Cell 1	-	19.4	19.8	18.4	19.2	19.6	23.7	24.3
Cell 2	19.4	-	18.9	18.4	19.6	20.1	24.5	24.1
Cell 3	19.7	18.9	-	18.4	19.5	19.9	22.1	23.9
Cell 4	18.4	18.4	18.4	-	19.1	19.3	24.5	24.5
Cell 5	19.2	19.6	19.5	19.1	-	19.6	25.0	23.4
Cell 6	19.7	20.1	19.9	19.3	19.6	-	23.9	22.6
Cell 7	23.7	24.5	22.1	24.5	25.0	23.9	-	24.4
Cell 8	24.3	24.1	23.9	24.5	23.4	22.6	24.4	-

4. Individual Chromosomes

4.1. Comparison of chromosomes in the entire genome simulations

The comparison of the simulated structure between cells can be made on the chromosome level instead of the entire genome level. The results can be seen in Table 4.1. While the mean RMSDs are significantly smaller than for the entire-genome case, they are still high enough to suggest no real similarity between the chromosomes across all cells, and the decrease can be explained simply by the smaller number of constraints when aligning the two trajectories. Considering even the minimum RMSD of each chromosome is never smaller than 4.7, it can be concluded that very likely there is no significant similarity between each chromosome across the different cells.

Table 4.1.
Mean and minimum of RMSDs of a particular chromosome pairwise between the averaged trajectories of all cells, excluding self-comparisons.

Chrom	1	2	3	4	5	6	7	8	9	10
Mean RMSD	9.6	8.1	8.4	8.3	8.1	7.5	8.3	8.0	8.1	7.9
Min RMSD	6.8	6.5	5.6	5.3	6.1	5.9	5.9	6.0	6.3	4.9
Chrom	11	12	13	14	15	16	17	18	19	X
Mean RMSD	7.8	8.2	8.6	8.4	7.1	7.4	7.3	6.9	7.0	8.7
Min RMSD	6.6	5.6	5.6	5.6	5.4	5.3	5.5	5.0	4.7	6.7

An analysis that is somewhat complementary to the above is to study the relative chromsome interaction strengths in each cells, that is how strongly each pair of chromosomes interacts with each other. A naive way to calculate this interaction strength is to simply count the number of beads between each pair of chromosomes that are in contact and divide by the length of both chromosomes involved. The result is displayed in Figure A.5. While a deeper comparison of these interaction strengths between cells is beyond the scope of this work, a simple visual analysis yielded no obvious relationships between the interaction strengths of the different cells. On a positive note, a certain resemblence can be seen to

the same kind of image in (Stevens et al.[3]) in Figure 1b, although a more detailed comparison could not be made due to a lack of the data from (Stevens et al.).

4.2. Simulation of individual chromosomes

Instead of simulating the entire genome, only individual chromosomes can be simulated in isolation. This can show how much of the structure of each chromosome is dependent on intrinsic interactions in the chromosome itself opposed to extrensic interactions with other chromosomes. Simulations were carried out for chromosome 1 and chromosome 19, as those are the largest and smallest chromosome respectively.

4.2.1. Chromosome 1

In Figure 4.1 the potential energy for the simulation of chromosome 1 of cell 3 is displayed. Compared to the potential energy of cell 2 in Figure 2.2 the potential energies for the invdividual chromosome look a lot more spread out and unstable, but the coefficient of variation of the potential energy of 1.24% is actually comparably low in reference to the coefficients of variation of the cell simulations as seen in Table A.2.

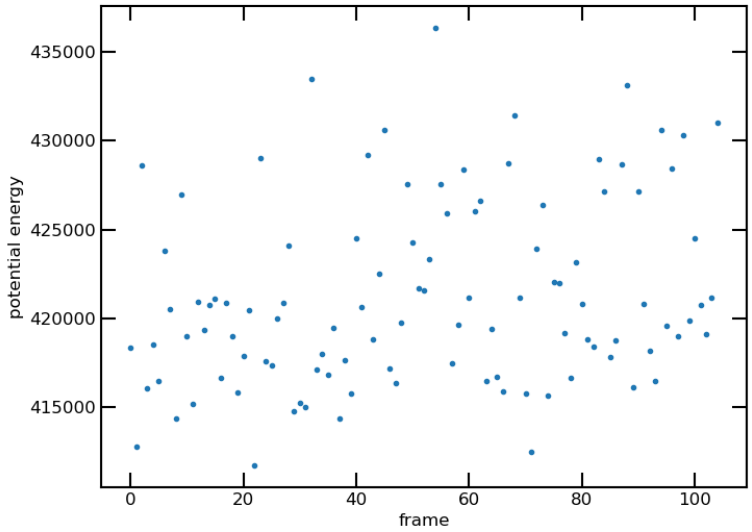

Figure 4.1.
Potential energy of all frames for simulation of chromosome 1 of cell 3 indiviually.

Looking at the RMSDs of the chromsome simulation though, seen in Figure 4.1, the mean RMSD from frame 22, which is the lowest potential energy frame, is quite high at 4.6 ± 1.0. Even worse, even though frame 1 and frame 71 have a similarly low energy compared to frame 22, their respective RMSDs from frame 22 are still larger than 3.0, implying that neither of these frames represents a true ground state. The mean RMSDs from the respective lowest energy frame for the simulations of chromosome 1 for all cells in Table 4.2 show that this is the case for all cells. Unsurprisingly, when comparing these individually simulated chromosomes with their respective counterparts in the simulations of the entire genome, by calculating the RMSD from the chromosome in the average trajectory, the results are similarly bad. This implies that the structure of chromosome 1 when simulated individually is quite dissimilar to the structure when simulated in the context of the entire cell.

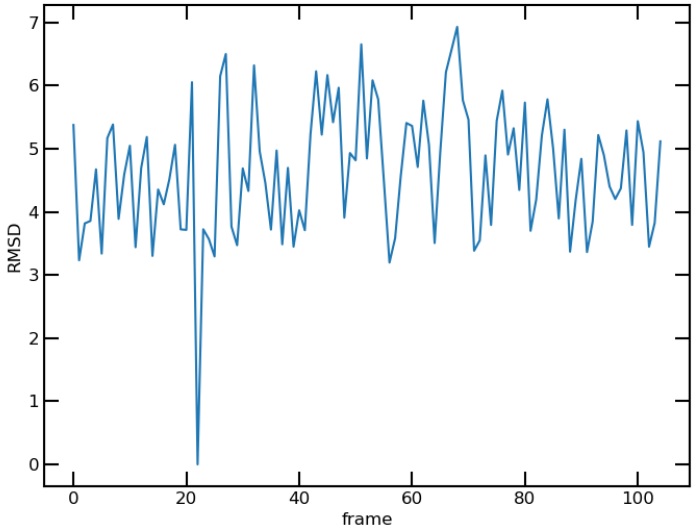

Figure 4.2.
RMSD of each frame of the simulation of chromosome 1 of cell 3 with the lowest-energy frame, frame 22.

Table 4.2.
Mean of RMSDs between each frame of chrosomome 1 simulation to lowest energy frame of this simulation and to chromosome 1 in the average trajectory of the entire cell simulation for each cell.

Cell	1	2	3	4	5	6	7	8
Mean RMSD to lowest energy frame	4.6	3.4	4.6	2.5	4.2	3.2	4.9	8.2
Mean RMSD to avg cell	4.6	3.9	5.6	3.1	4.6	6.8	5.1	7.8

4.2.2. Chromosome 19

The same analysis has been repeated for chromosome 19, which is the shortest chromosome with only 584 beads compared to chromosome 1 with 1924 beads. For the RMSDs from the minimum energy frame, displayed in 4.3, we see the same pattern as already for chromosome 1, with the RMSD being both very high and unstable. In particular no stable ground state configuration is reached again. The RMSDs to the average trajectory from the entire genome simulation can be seen in Table 4.3. While most of the RMSDs are similarly high, for cell 1 and cell 4 the mean RMSD to the average simulated cell drops to 2.3 and 1.5 respectively, which is indicative of decent resemblance. Nevertheless, generally the difference between the individually simulated chromosomes and the chromosomes in the cell is still quite high, suggesting that the results of cell 1 and cell 4 are merely the exception from the rule and that the stabilising effect of other chromosomes is a non-neglegible factor for determining a chromosome's structure.

Table 4.3.
Mean of RMSDs between each frame of chrosomome 19 simulation to lowest energy frame of this simulation and to chromosome 19 in the average trajectory of the entire cell simulation for each cell.

Cell	1	2	3	4	5	6	7	8
Mean RMSD to lowest energy frame	1.8	3.7	4.9	1.2	3.6	2.5	4.6	5.9
Mean RMSD to avg cell	2.3	4.1	6.0	1.5	4.4	3.1	4.8	6.1

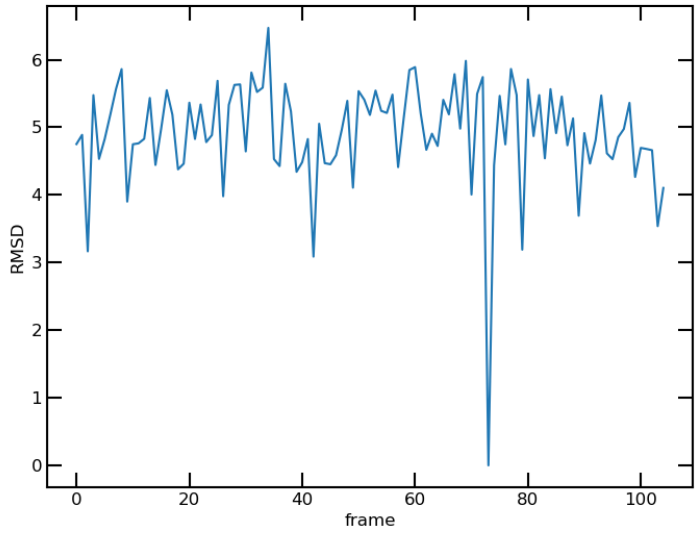

Figure 4.3.
RMSD of each frame of the simulation of chromosome 19 of cell 3 with the lowest-energy frame, frame 73.

5. Conclusion

Using the protocol defined in Wettermann et al.[4], a total of eight cells were simulated using a molecular dynamics simulation with a number of predefined bonds and contacts derived from Hi-C data sets. Both the potential energies and the distance distribution of the predefined bonds and contacts show high agreement with (Wettermann et al.[4]) and suggest that in most cases the simulation reaches a stable ground configuration.

The two major exceptions, cell 1 and cell *5*, were analysed more in-depth and at least in the case of cell *5* the problematic features of the simulation could be resolved to be merely an artefact of that particular simulation run, with reruns of the simulation of that particular cell falling very much in line with the results of the other cells.

As for cell 1, while it was shown that a ground configuration still existed, even though it was not stable, the instability of the configuration was consistently observable across multiple simulation runs. A deeper reason for why this effect arises only in cell 1 and not the other cells could not be given, and could be a question for further research.

The analysis of cell *5* on the other hand showed the possibility of semi-stable configurations to exist with higher energies compared to the ground state energy, that appear in some simulation runs, but not in others. Considering the fact that this semi-stable state disappeared by transitioning into the ground state configuration after more simulation cycles suggests that in this case the effect might not be biologically relevant as it disappeared on its own, but implies that the existence of relevant configurations besides the ground state is certainly possible. This raises the possibility of misfolded genome configurations, like misfolded proteins, that could change the transcription behaviour of a cell.

The simulation results from the different cells were compared pairwise for whole-genome similarity using both SCC and RMSD as similarity measures. The SCC between the initial Hi-C contact matrices and the contact matrices of the simulations showed decent correspondence, but considering the fact the Hi-C data set contains only an estimated 5% of all the contacts in the physical cells, these scores are still reasonable. This implies that the simulation model managed to reproduce the Hi-C data set reasonably well.

The global genome structures of the simulations for each cell were compared using both SCC and RMSD as a measure of comparison. Both measures unquestionably implied that the correspondence between

global structures of each cells were significantly different, with the SCCs being in a range of 0.10 to 0.22 and the RMSDs being in a range of 18.4 and 25.0.

Similarly, for comparisons of individual chromosomes from the entire genome simulations between cells no significant similarity could be found, with the RMSD for no chromosome and no cell combination falling below 4.7. This implies that, just like on the whole genome level, the structure of the individual chromosomes varies significantly between cells.

Furthermore, chromosome 1 and chromosome 19 were also simulated in isolation to determine how much the structure of each chromosome depends on the influence of the other chromosomes opposed to the intrinsic interactions within the chromosome itself. For both chromosomes the results quite clearly indicated that the influence of the other chromosomes is not negligible, with significant differences arising between the chromosomes simulated in isolation and in the context of the entire cell.

In summary, both in the entire genome structure as well as the structure of the individual chromosomes no significant similarity could be found between any of the cells in any way. The straightforward explanation for this would be that exact structural resemblences on high levels simply do not exist between cells, even of in cells of the same cell line, which is in accordance with the findings in (Stevens et al.[3]). In the same way of thinking, some current research seems to be focusing more on other kinds of structures such as A/B compartments, topological-associated domains, and loops[3]. The chance of finding simple global patterns in genome structures seems to be rather slim at this point.

The source code used for the analysis can be found at `https://github.com/Lego1120/chromatin-structure`.

Bibliography

[1] A. Uzman, "Essential cell biology, third edition," *Biochemistry and Molecular Biology Education*, vol. 38, no. 1, pp. 59–59, Jan. 1, 2010, Publisher: John Wiley & Sons, Ltd, ISSN: 1470-8175. DOI: 10.1002/bmb.20371. [Online]. Available: https://doi.org/10.1002/bmb.20371 (visited on 03/19/2022).

[2] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science (New York, N.Y.)*, vol. 326, no. 5950, pp. 289–293, Oct. 9, 2009, ISSN: 1095-9203. DOI: 10.1126/science.1181369. PMID: 19815776. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19815776.

[3] T. J. Stevens, D. Lando, S. Basu, *et al.*, "3d structures of individual mammalian genomes studied by single-cell hi-c," *Nature*, vol. 544, no. 7648, pp. 59–64, Apr. 1, 2017, ISSN: 1476-4687. DOI: 10.1038/nature21429. [Online]. Available: https://doi.org/10.1038/nature21429.

[4] S. Wettermann, M. Brems, J. Siebert, G. Vu, T. Stevens, and P. Virnau, "A minimal gō-model for rebuilding whole genome structures from haploid single-cell hi-c data," *Computational Materials Science*, vol. 173, p. 109178, Feb. 15, 2020, ISSN: 0927-0256. DOI: 10.1016/j.commatsci.2019.109178. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092702561930477X.

[5] J. A. Anderson, J. Glaser, and S. C. Glotzer, "HOOMD-blue: A python package for high-performance molecular dynamics and hard particle monte carlo simulations," *Computational Materials Science*, vol. 173, p. 109363, Feb. 15, 2020, ISSN: 0927-0256. DOI: 10.1016/j.commatsci.2019.109363. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025619306627.

[6] M. P. Howard, J. A. Anderson, A. Nikoubashman, S. C. Glotzer, and A. Z. Panagiotopoulos, "Efficient neighbor list calculation for molecular simulation of colloidal systems using graphics processing units," *Computer Physics Communications*, vol. 203, pp. 45–52, Jun. 1, 2016, ISSN: 0010-4655. DOI: 10.1016/j.cpc.2016.02.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010465516300182.

[7] M. P. Howard, A. Statt, F. Madutsa, T. M. Truskett, and A. Z. Panagiotopoulos, "Quantized bounding volume hierarchies for neighbor search in molecular simulations on graphics processing units," *Computational Materials Science*, vol. 164, pp. 139–146, Jun. 15, 2019, ISSN: 0927-0256. DOI: 10.1016/j.commatsci.2019.04.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092702561930206X.

[8] C. R. Harris, K. J. Millman, S. J. v. d. Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, Publisher: Springer Science and Business Media LLC. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[9] T. Yang, F. Zhang, G. G. Yardımcı, *et al.*, "HiCRep: Assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient," *Genome Research*, vol. 27, no. 11, pp. 1939–1949, Nov. 2017, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.220640.117. [Online]. Available: http://genome.cshlp.org/lookup/doi/10.1101/gr.220640.117 (visited on 02/12/2022).

[10] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, Jun. 2007, ISSN: 1558-366X. DOI: 10.1109/MCSE.2007.55.

[11] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021, Publisher: The Open Journal. DOI: 10.21105/joss.03021. [Online]. Available: https://doi.org/10.21105/joss.03021.

[12] W. Humphrey, A. Dalke, and K. Schulten, "VMD – visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.

[13] J. Stone, "\em an efficient library for parallel ray tracing and animation," Master's Thesis, Computer Science Department, University of Missouri-Rolla, Apr. 1998.

A. Tables and Figures

Table A.1.
Number of beads in each chromosome in the simulation. Each bead represents 100 kbp

Chrom	1	2	3	4	5	6	7	8	9	10
Length	1924	1791	1570	1534	1488	1466	1424	1263	1215	1275
Chrom	11	12	13	14	15	16	17	18	19	X
Length	1189	1171	1174	1218	1010	952	919	876	584	1671

Table A.2.
Statistics about the quality each cell simulation. The pot(ential) energy coefficient of variation (CV) is the standard deviation of the potential energy divided by the mean potential energy. Mean and 99.73th percentile are calculated for both bonds and predefined contacts, i.e. the harmonic potentials that were set as part of the simulatin protocol.

		bond lengths		contact lengths	
Cell	pot energy CV	mean	99.73th percentile	mean	99.73th percentile
1	8.51%	1.13	1.98	1.62	2.72
2	0.27%	1.10	1.71	1.61	2.42
3	0.67%	1.14	1.88	1.68	2.52
4	1.83%	1.14	2.08	1.61	2.46
5	8.63%	1.11	1.81	1.62	2.55
6	0.94%	1.16	2.07	1.61	2.38
7	0.29%	1.10	1.58	1.59	2.09
8	0.28%	1.08	1.56	1.60	2.25

Table A.3.
Number of contacts prespecified before the simulation by the Hi-C data set and seen in the resulting structure after the simulation for each cell, as well an estimation of the number of actual contacts captured by the Hi-C procedure by the ratio of the two.

Cell	contacts prespecified	contacts after simulation	% of contacts specified
1	48,962	932,831	5.2%
2	32,243	558,981	5.8%
3	19,112	602,439	3.2%
4	35,514	685,314	5.2%
5	31,180	592,569	5.3%
6	32,862	358,721	4.8%
7	21,126	358,721	5.9%
8	17,581	320,656	5.5%

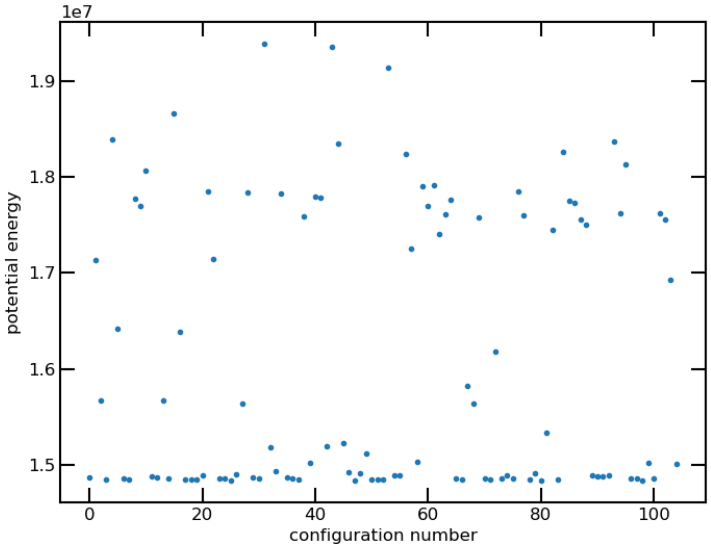

Figure A.1.
Potential energies for the frames of the first repeat simulation of cell 1.

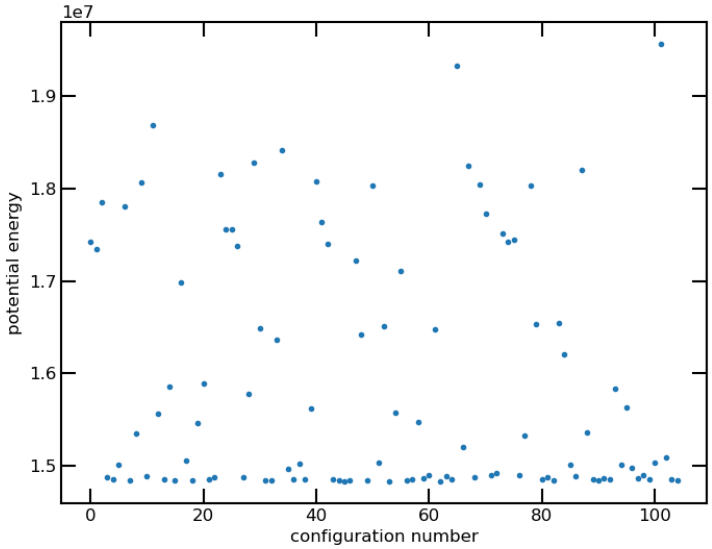

Figure A.2.
Potential energies for the frames of the second repeat simulation of cell 1.

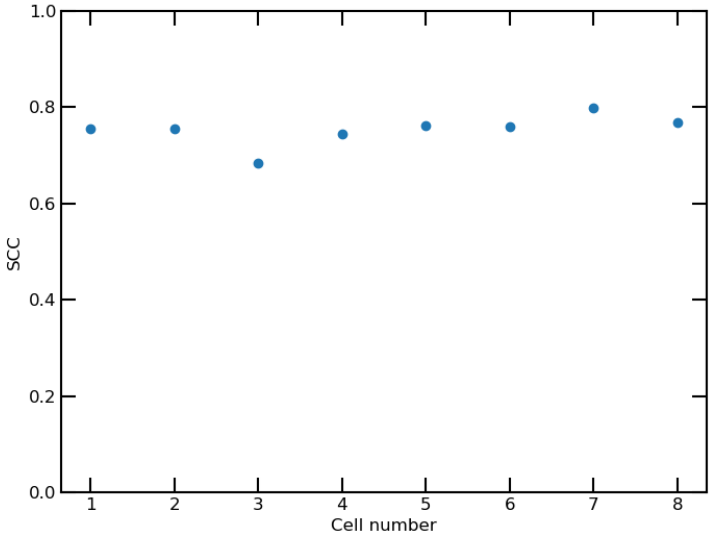

Figure A.3.
SCCs between the original Hi-C contact matrices and the contact matrices of the simulated cell genomes.

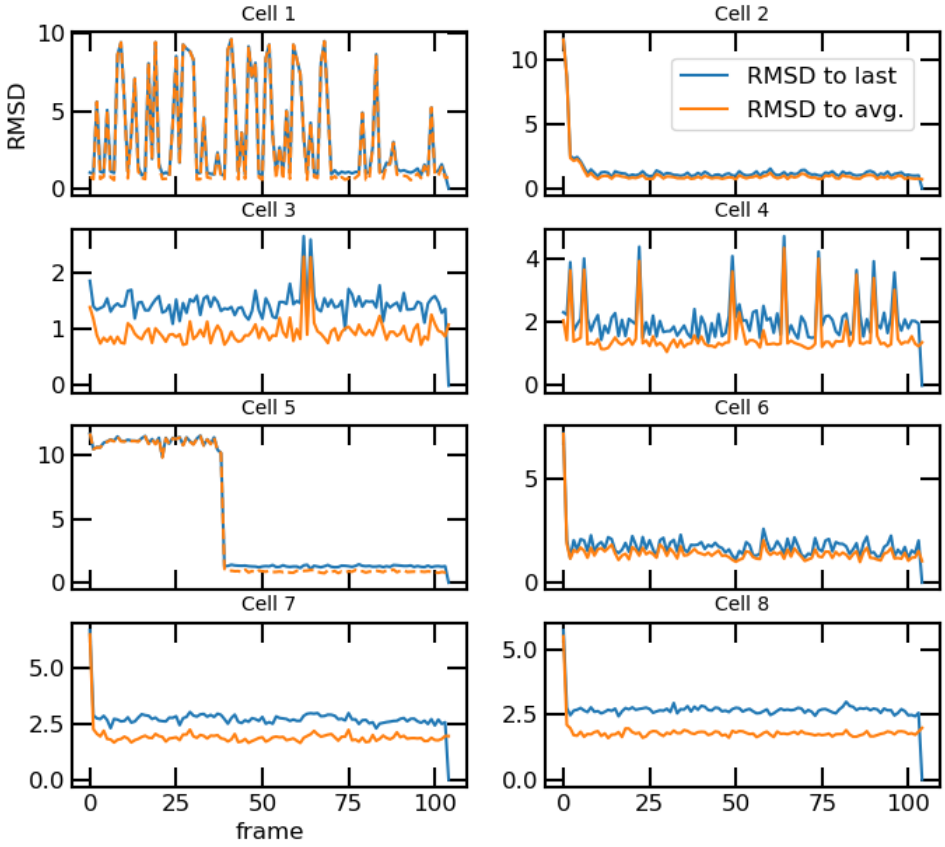

Figure A.4.
RMSD of each frame to both the last frame of the simulation and the average trajectory of the cell for each cell. The average trajectories were generated from a subset of frames from each cell that represents the ground state configuration of that cell in the simulation and is described in more detail in section 3.3.

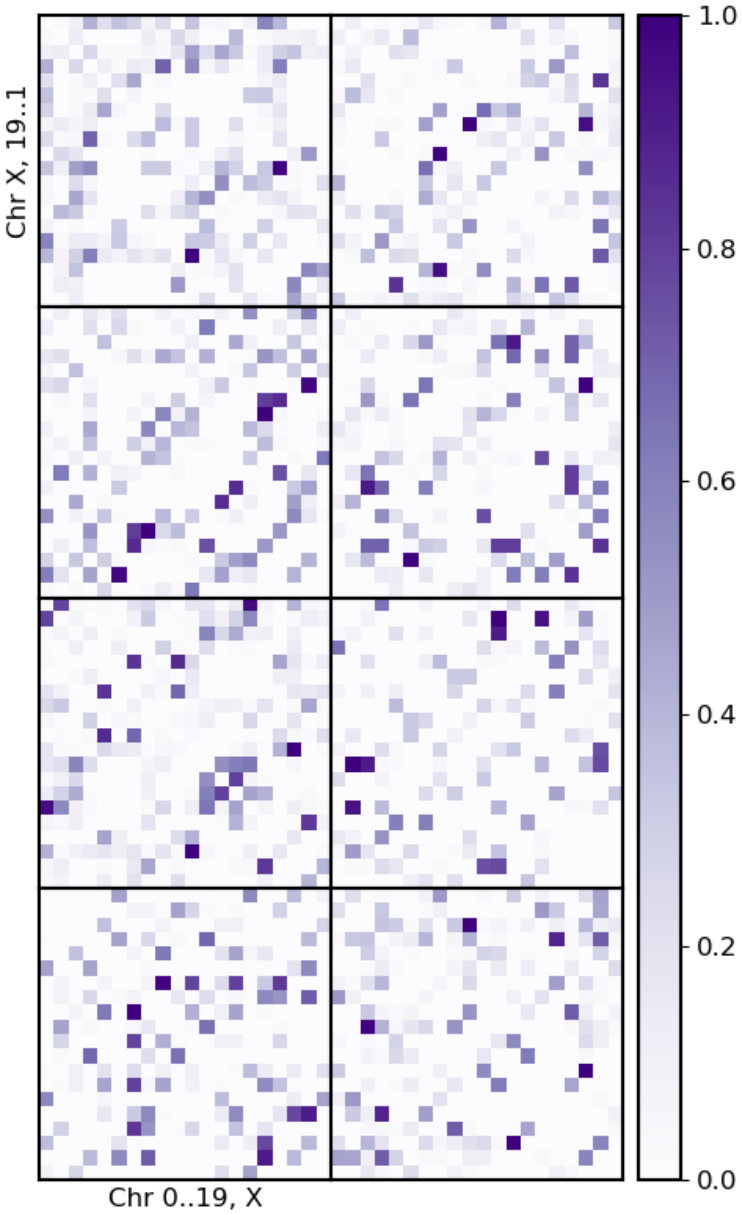

Figure A.5.
Relative interaction strength of chromosomes in each cell. The color scale is relative to the strongest chromosomal interaction and independent for each cell. The ordering of cells is in the direction of reading, with cells 1 and 2 being in the top row.

B. Used Software

Figures were created using matplotlib[10] and seaborn[11].

Table B.1.
Software used for simulation and data analysis. Two systems were used, the top one a Manjaro Linux (based on Arch Linux) system and the bottom one a Microsoft Windows system with Anaconda.

Package	Version	Package Source
Python	3.10.2	arch repo
b HOOMD-blue	1.9.7	built from source
numpy	1.22.2	arch repo
scipy	1.8.0	arch repo
pandas	1.4.0	arch repo
matplotlib	3.5.1	arch repo
gsd	2.5.1	pip
vmd	1.9.3	https://www.ks.uiuc.edu/Research/vmd/
Python	3.9.7	anaconda
numpy	1.20.3	anaconda
scipy	1.7.1	anaconda
pandas	1.3.4	anaconda
matplotlib	3.4.3	anaconda
gsd	2.5.1	pip
hicreppy	771cf72	github

C. Renderings of simuated cells

All renderings rendered using VMD[12] and tachyon[13] (internal). Colorscale is BGR (from blue at the beginning of the trajectory over green in the middle to red at the end of the trajectory).

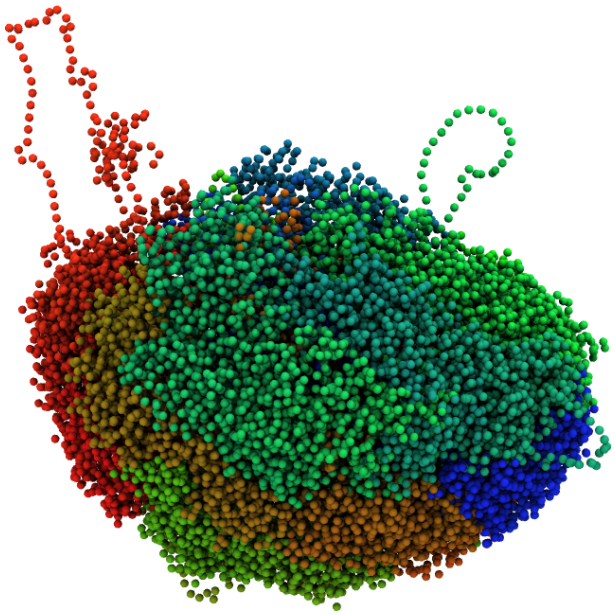

Figure C.1.
Cell 1, frame 104, scene 1

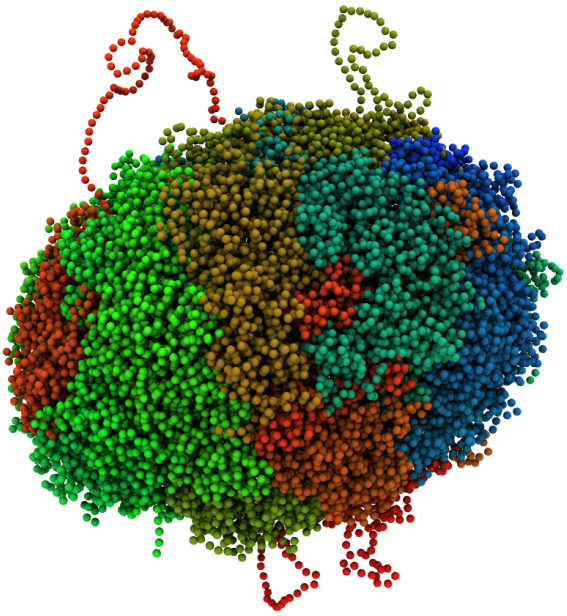

Figure C.2.
Cell 2, frame 104, scene 1

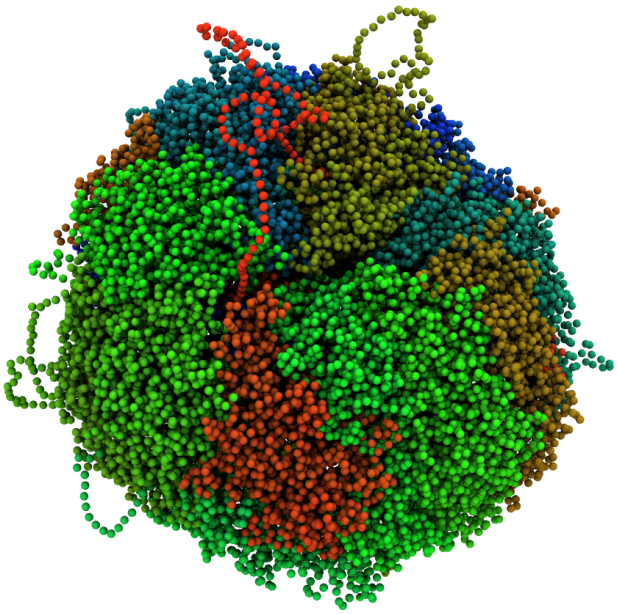

Figure C.3.
Cell 2, frame 104, scene 2

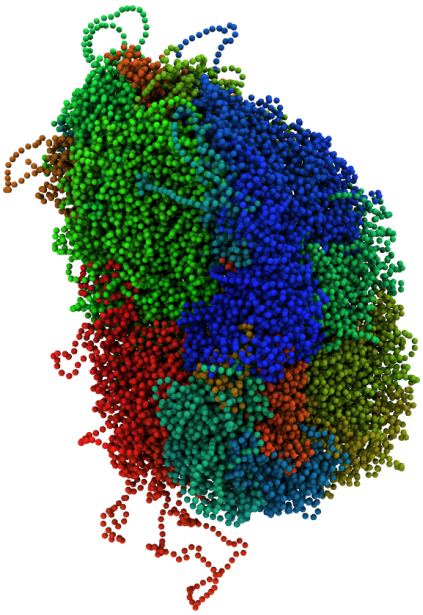

Figure C.4.
Cell 3, frame 104, scene 1

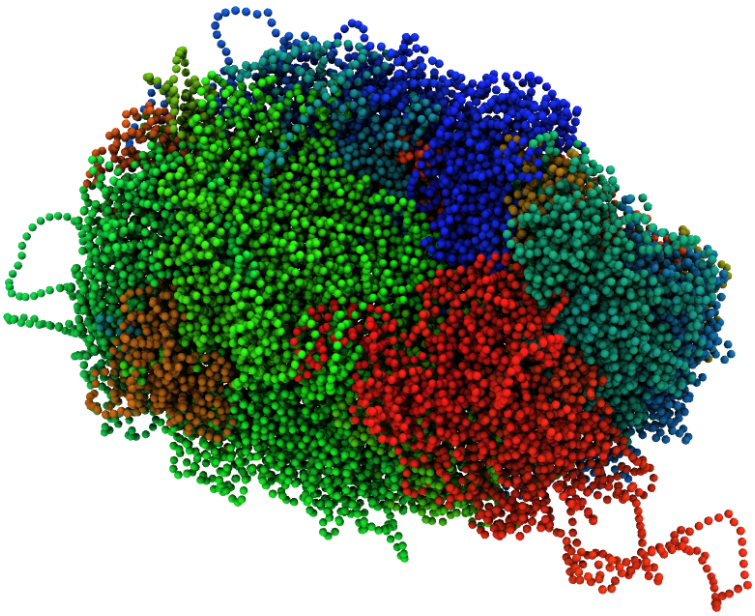

Figure C.5.
Cell 3, frame 104, scene 4

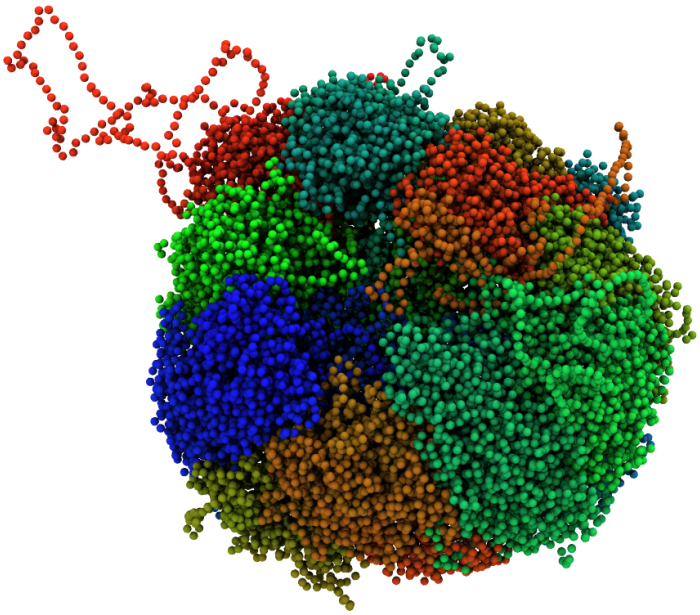

Figure C.6.
Cell 4, frame 104, scene 1

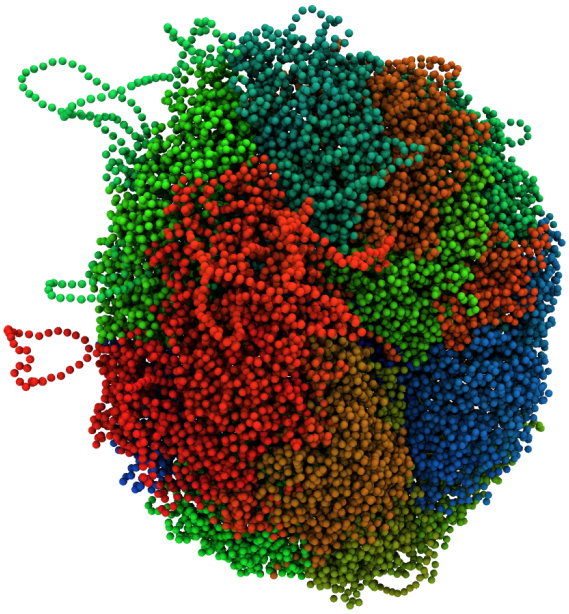

Figure C.7.
Cell 5, frame 36, scene 1

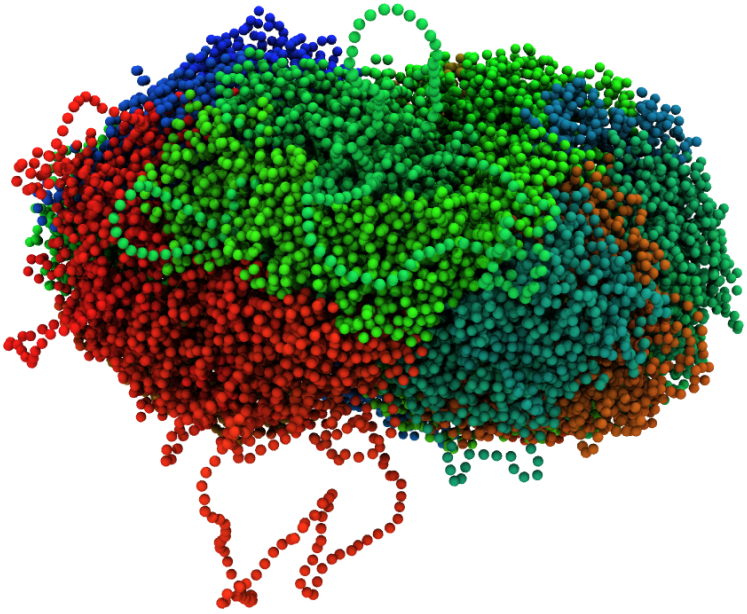

Figure C.8.
Cell 5, frame 36, scene 2

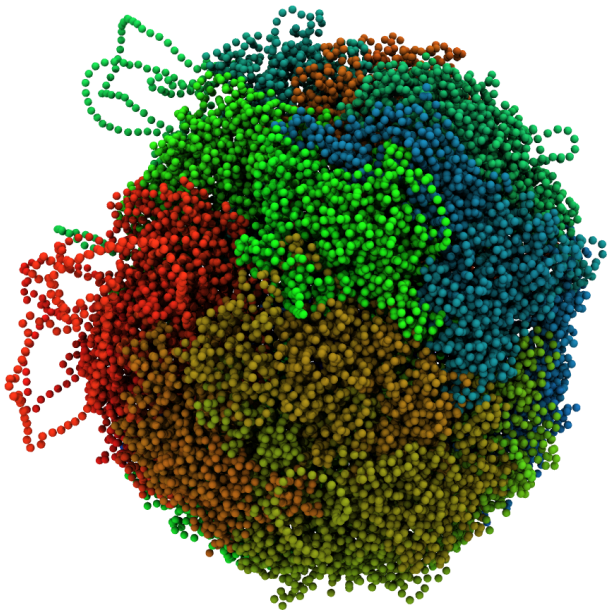

Figure C.9.
Cell 5, frame 104, scene 1

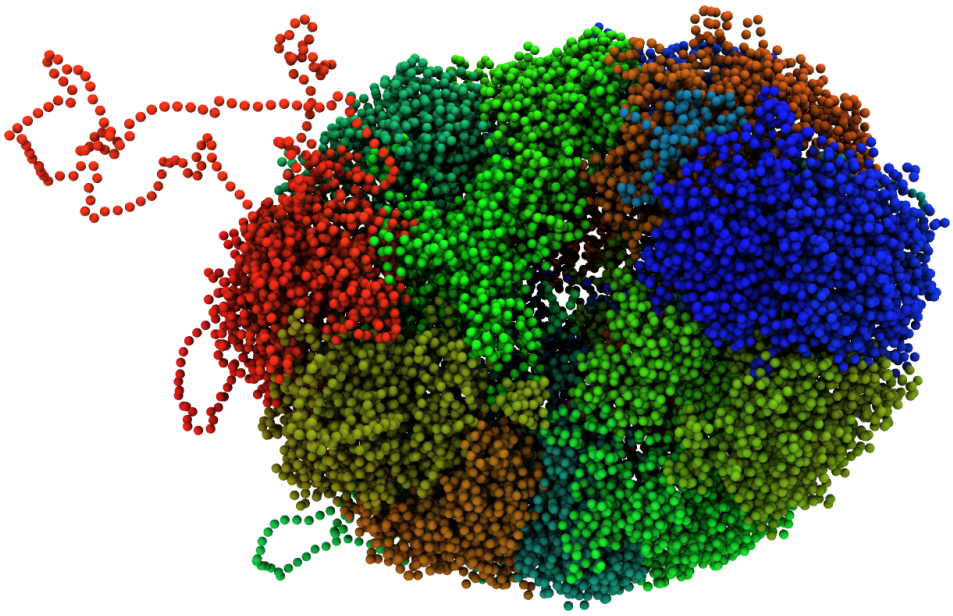

Figure C.10.
Cell 6, frame 104, scene 1

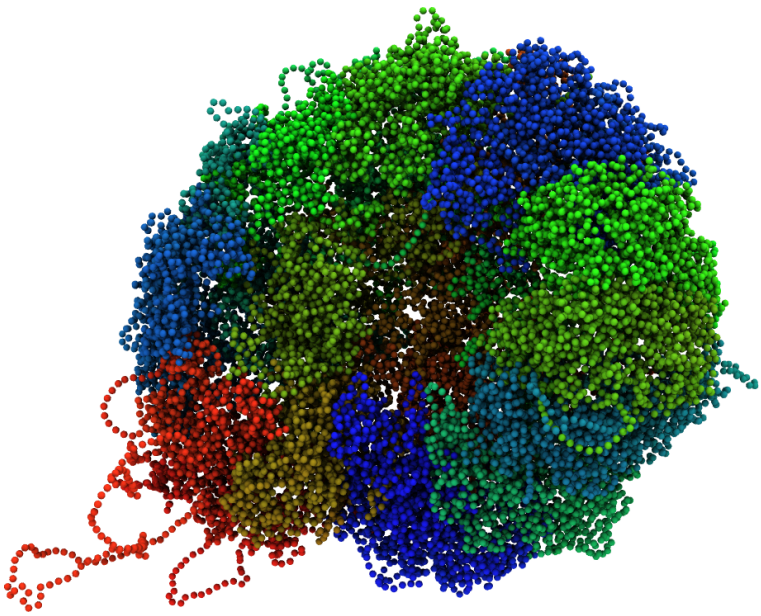

Figure C.11.
Cell 7, frame 104, scene 1

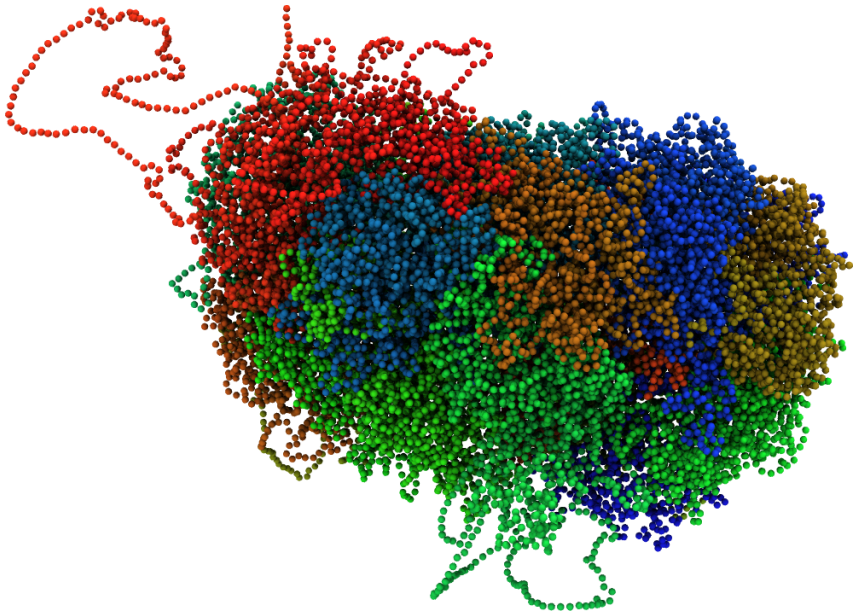

Figure C.12.
Cell 8, frame 104, scene 2