Consistency Tests for a recursive multi-scale 3D Chromatin Structure reconstruction Algorithm

Flagship Project InterOmics – WP1 CNR-ISTI

Claudia Caudai [1], Emanuele Salerno [1], Monica Zoppè [2], and Anna Tonazzini [1]

National Research Council of Italy
[1]Institute of Information Science and Technologies, and [2]Institute of Clinical Physiology, Pisa, Italy
claudia.caudai@isti.cnr.it

Abstract In this report, we test the consistency and coherence of an algorithm obtained as an extension of a technique we proposed in the past. This implements a recursive multi-scale reconstruction of the 3D chromatin structure from Chromosome Conformation Capture data. These data derive from millions of cells, so we cannot expect that they lead to a unique solution; for this reason, we adopt a statistic approach to sample the space of the solutions generated by a suitable objective function, in order to achieve configurations compatible with the input data and the known constraints. The consistency of the algorithm has been tested by producing a large number of results and evaluating the dispersion of the final values of the objective function. Using the same solutions, synthetic contact matrices have been produced and compared with the input matrix to test the coherence of our solutions with the initial data. Furthermore, we investigated the presence of typical structures in the solutions by hierarchical clustering.

1 Introduction

In our previous works [1, 2], we presented a method to reconstruct a set of plausible chromatin configurations starting from contact data obtained through Chromosome Conformation Capture techniques (HI-C, 3C, 4C, 5C) [3]. In our approach we do not look for a unique configuration because the experimental data derive from millions of cells. As opposed to most popular methods [5, 8], we do not translate contact frequencies into distances, since often not consistent with the Euclidean geometry [2]. Recent studies showed that chromatin is divided into segments called *Topological Association Domains* (TADs) [4], with few reciprocal interactions and a lot of inner contacts. Taking into account this topological feature, we adopted a multi-scale approach. Our algorithm automatically detects diagonal blocks in the input matrix, thus creating different, decreasing, resolution levels. At each level, the chromatin fiber is modeled as a chain of partially penetrable beads, whose features depend on the TAD structures computed at the immediately preceding level. Once the lowest resolution is reached, the full-resolution chain is reconstructed recursively from the intermediate results stored during the computation. Our algorithm samples the solution space generated by a specially designed objective function through a Monte Carlo method. The algorithm presented in [1] keeps the solutions in the feasible space by enforcing rigid geometrical constraints during the iteration. We are now trying to include *soft* constraints in the objective function in order to make easier the control of the solutions. The new objective function consists in two parts, the first concerning the fitness to the data, represented by the achievement of significant contacts, and the second concerning constraints, represented by penalties associated to interpenetrations between beads:

$$\Phi(\mathcal{C}) = \sum_{i,j\in\mathcal{L}} n_{i,j} \cdot d_{i,j} + \lambda \sum_{i,j\in\mathcal{C}} \frac{(r_i + r_j)}{2d_{ij}} \left[1 - \frac{\{c[d_{ij} - (r_i + r_j)]\}^b}{1 + \{c|d_{ij} - (r_i + r_j)|\}^b}\right] \qquad (1)$$

$\mathcal{C}$ is the 3D configuration of the chromatin segment, n_{ij} is the contact frequency between beads i and j, d_{ij} is their Euclidean distance, $\mathcal{L}$ is the set of significant bead pairs chosen from the contact matrix (the set of pairs for which n_{ij} exceeds a threshold), r_i and r_j are the radii of beads i and j, and λ is the regularization parameter, which represents the relative weight of the constraint component with respect of the fitness component. The term in the second summation in Equation (1) is referred to as the neighbor interaction function, ψ_{ij}. Parameter λ does not assume a fixed value, but is evaluated as the ratio between the values of the two summations in Equation 1 averaged on a fixed number of random chain configurations. Function ψ_{ij} is an equilateral hyperbola, modified to grow as $1/d_{ij}$ when d_{ij} tends to zero, and decrease quickly when d_{ij} goes well beyond a threshold distance $(r_i + r_j)$. c is a scale factor, whose presence introduces a moderate slope interval close to $(r_i + r_j)$, to allow partial interpenetration of beads; the exponent b, odd integer, regulates the slope of the transition intervals between the moderate slope region and the external intervals (see Figure 1). At each step of the Monte Carlo algorithm, the subchains are perturbed by using quaternions rather than Euler angles, in order to facilitate composition of rotations and to avoid singularities. Once the subchains are reconstructed, each segment can be treated as an element of a new chain, and the procedure can be repeated recursively at different scales (see Figure 2).
Our algorithm, implementing the flow diagram in Figure 2, contains a *Topological Association Domains* extraction algorithm, which automatically detects diagonal blocks of input contact matrix. The reconstruction method implements a Simulated Annealing. In Table 1, the parameters appearing in our code are listed and briefly described. These parameters can be set at the beginning of the experiments to take into account physical and biochemical information. The values appearing in the table are the ones we used to obtain the set of experimental results presented in this report. We have equipped our highest-resolution beads with geometric characteristics; the names in brackets refer to parameters in Table 1:

1. We assume that chromatin fiber has a diameter of about 30 nm (DIA);

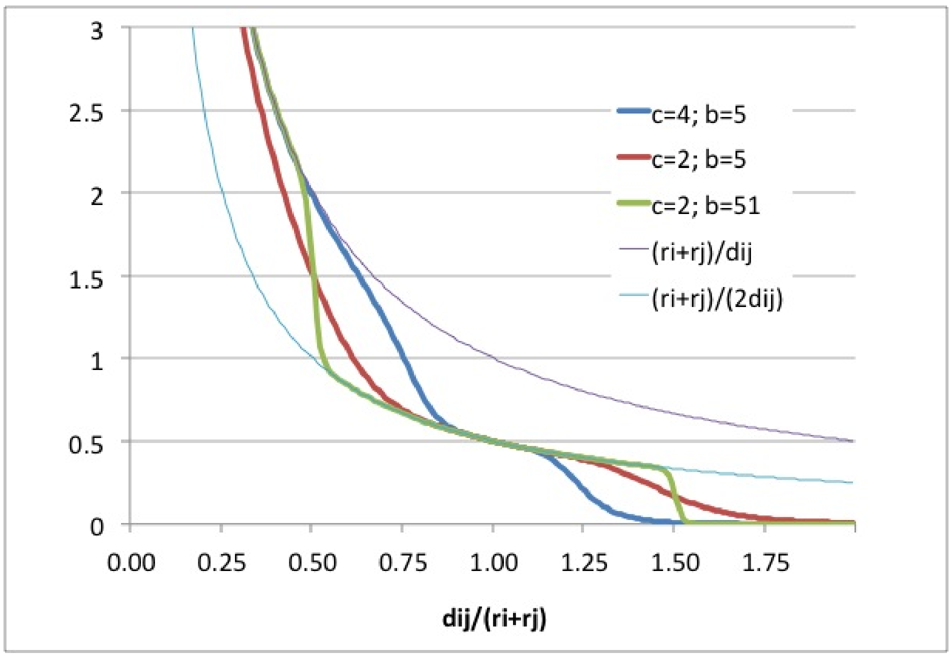

Figure1. Neighborhood interaction function ψ_{ij} with different values of c and b.

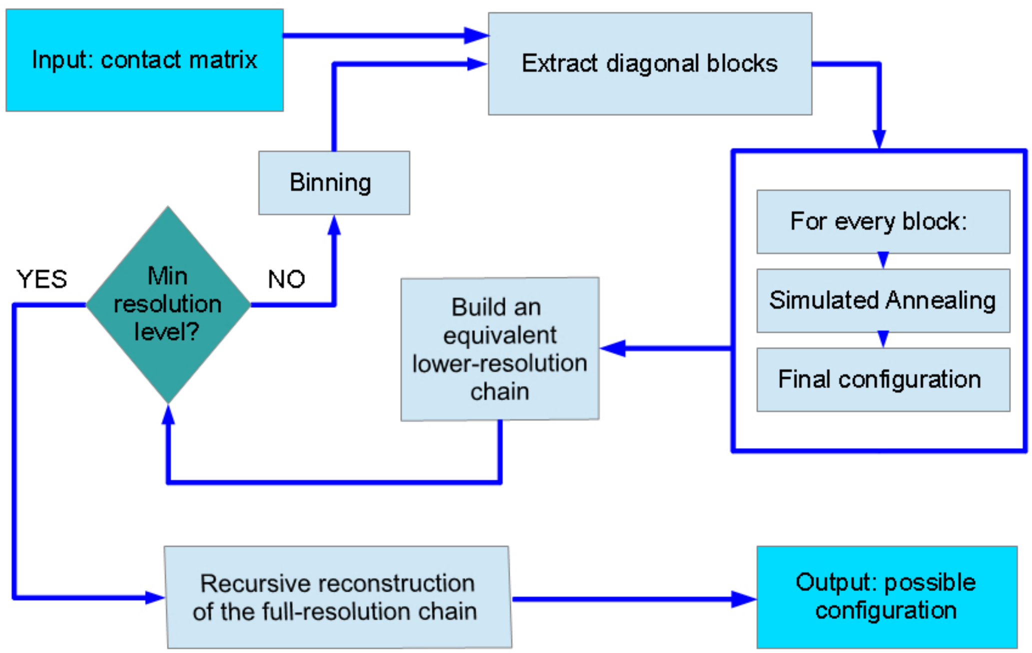

Figure2. Flow diagram of the algorithm.

2. we suppose that in a fragment of chromatin with length 30 nm (approximation of a chromatin *cube*) we can find 3 *kbp* (NB);
3. the number of *cubes* in a resolution unit is $RIS/NB = NC$, where RIS is the highest resolution of data into the contact matrix (in our experiments $RIS = 100kbp$);
4. we vary the diameter of each bead of the highest-resolution level linearly with the number of inner contacts (main diagonal of the contact matrix) within the range $[LMIN, LMAX]$. We suppose that many inner contacts imply a major compactness.
   We choose the maximum diameter of a bead as:
$$LMAX = \frac{NC \cdot DIA}{\pi} \tag{2}$$
   that is, the diameter of a circumference with length $NC \cdot DIA$.
   We choose the minimum diameter of a bead as:
$$LMIN = \sqrt[3]{NC} \cdot \sqrt{3} \cdot DIA \tag{3}$$
   We calculate the diameter of a bead at the maximum-resolution level with the formula:
$$L = \frac{LMAX - (LMAX - LMIN) \cdot ncount}{NMAX} \cdot DIA \tag{4}$$
   where $ncount$ (coincident with n_{ii} for the i-th bead) is the number of inner contacts of the bead, and $NMAX$ is the maximum number of contacts in the contact matrix;
5. we call EXT the radius of a bead. In the highest-resolution level, EXT is $L/2$. In the successive levels, each bead derives from the structure of a higher-resolution subchain. We compute the standard deviation of the coordinates of its beads along their first principal component, and assume a fixed fraction ($extrate$) of this quantity as the radius of the corresponding lower-resolution bead. Weonly take fraction of the standard deviation, because we want to model the lower-resolution levels as bead chains, so we need to equip subchains of a spherical envelope, which should contain much of the structure, not preventing reciprocal approaching, when necessary.

To test the consistency and the coherence of the algorithm we have selected HI-C data from the long arm of human chromosome 1 made available by [7], and run our code 100 times with the parameters reported in Table 1. The selected segment consists of 29.200 Mbp (292 segments of 100 *kbp*). The algorithm of authomatic detection of diagonal blocks identifies two resolution levels, the second made up of 25 beads, which reflects very well the division in TADs of Dixon [4], as shown in Figure 3. In Section 2, we discuss the repeatability of the results. In Section 3, we analyze the coherence of the results with input data. In Section 4, we illustrate some features of the solutions and in Section 5, we report our conclusions.

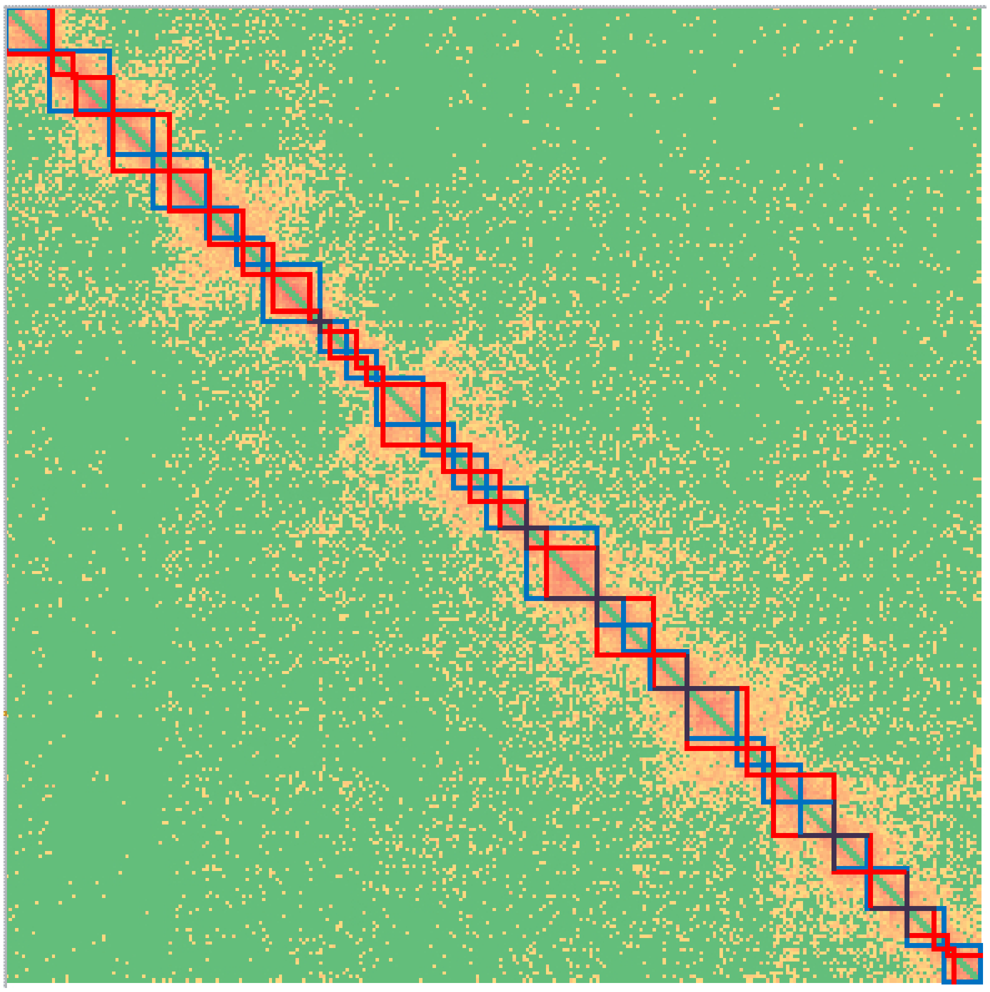

Figure3. Heat map of contact matrix of chromosome 1 (150.28 Mbp – 179.44 Mbp) [7]: red, TADs from Dixon [4]; blue, blocks detected by our algorithm [1, 2].

Table1. Set of parameters used in the 100 experiments

Parameters	Significance in the algorithm
ENERGY: FITNESS	
$diagneg = 1$	neglected diagonals in contact matrix
$datafact = 0.4$	fraction of the pairs per block used to build the penalty function (real)
ENERGY: CONSTRAINT	
$energy\_scale = 3.98$	tunes the extent of the moderate penalty range around the threshold distance
$energy\_exp = 5$	tunes the slopes of the penalty in the transitions between the moderate range and the external interval
TAD EXTRACTION	
$tol = 2$	row/column tolerance for approximate diagonal block extraction
$ThrDiv = 20$	amplitude tolerance for approximate diagonal block extraction (fraction of the off-diagonal maximum)
$minsize = 7$	minimum required diagonal block size
ANNEALING λ	
$regulenergy = 0.5$	target relative weight fitness/constraint
$avgenergy = 1000$	fixed number of averaging cycles to evaluate the fitness/constraint ratio
$percenergy = 90$	maximum accepted percentile in the energy sample sets (integer ≤ 100)
ANNEALING: WARM-UP	
$Tmax = 30$	Start temperature of warm-up phase
$itwarm = 50000$	Max number of warm-up cycles (positive integer)
$incrtemp = 1.2$	Fixed temperature increase coefficient at warm-up
$checkwarm = 500$	Fixed milestone on warm-up cycles to check the acceptance rate
$muwarm = 0.9$	Minimum acceptance rate to end warm-up
ANNEALING: SAMPLING	
$itmax = 50000$	Max number of annealing cycles (positive integer)
$itstop = 500$	consecutive cycles with energy variations within StopTolerance to stop annealing
$StopTolerance = 1 \cdot 10^{-5}$	Stop tolerance
$RANDPLA = 2 \cdot 0.05$	Max planar angle perturbation at each update (floating, radians, doubled for convenience)
$RANDDIE = 2 \cdot 0.05$	Max dihedral angle perturbation at each update (floating, radians, doubled for convenience)
$decrtemp = 0.998$	Fixed cooling coefficient at each annealing cicle
GEOMETRIC PARAMETERS	
$DIA = 30$	Chromatin fiber diameter (nm)
$RIS = 100$	Contact matrix (full) genomic resolution (kbp)
$NB = 3$	kilobase-pairs in a chromatin fragment with length DIA
$NC = RIS/NB$	fragments per genomic resolution unit
$LMAX = (NC \cdot DIA)/\pi$	hypothesized maximum size of a bead in the full-resolution model (nm)
$LMIN = \sqrt[3]{NC} \cdot \sqrt{3} \cdot DIA$	hypothesized minimum bead size in the full-resolution model (nm)
$extrate = 0.3$	Size tuning for beads at levels > 0

2 Repeatability of solutions

In this Section, we check our objective function against its capability of producing results compatible with both the data and the constraints. To do this, we verify the repeatability of solutions in terms of final energy, the final value of the objective function (1). As mentioned in Section 1, the solutions we are serching for cannot be unique. Thus, we ask that the solutions deriving from the same inputs, and obtained through the same parameters, are close in terms of final energy. We analyze the energies of the final configurations of the whole chain and all the subchains for our set of experiments. The randomness of these distributions is checked by Gaussianity tests, and their dispersion is evaluated by comparing the standard deviation of each group of results with the corresponding average. We test the normality of distributions of final energies of every subchain with a Shapiro-Francia test, using the function `sf.test` of R. In the Shapiro-Francia test the null hypothesis is: *the data follow a normal distribution*, if p-value > 0.05 we accept the null hypothesis, if p-value < 0.05 we remove the outlier with greatest deviation with the function `outlier` of R. We perform again the test for the distribution without the outlier. We report in Table 2 the number of outliers to be removed to obtain normality in every distribution, and the standard deviation expressed as percentage of the average, for every energy distribution.

We can observe that the standard deviation without outliers is always $< 5\%$ of the average, for every block. Considering the outliers, it is a little larger but still very small. This is a good result because it means that the final energies are not so scattered, and the final configurations fall in regions with similar energy values. In general, the higher the number of beads, the higher the standard deviation. This is reasonable, because a high number of beads leads to many degrees of freedom and a large variability in final configurations. For the whole chain, the standard deviation is 11% of the mean value.

As we can see by looking at the outliers, the distributions need at most the removal of a few outliers to be considered Gaussian. We notice that almost all outliers with greatest deviation correspond to low final energy values. This could mean that outlier configurations are often more compact than others (see Figure 4 and Figure 5). This apparently weird behavior is probably due to the great difficulty in achieving conformations with many simultaneous relevant contacts. When these conformations are reached without deep interpenetrations, the energy reaches low values. We have built the objective function so that the achievement of many contacts simultaneously, avoiding interpenetration of beads, is a good result.

3 Coherence of results with input data

To test the coherence of our results with the input data, we tried to reconstruct contact matrices of solutions, to be compared with the input contact matrix. Our experiments do not produce contact matrices, but, from each result, we are

Table2. Normality of distributions of final energies.

	nr of beads	nr of outliers	% st dev[a]	% st dev[b]
Block 0	13	1	4.11	3.93
Block 1	18	14	4.51	3.64
Block 2	13	0	3.95	3.95
Block 3	16	4	4.30	3.35
Block 4	9	1	4.00	3.59
Block 5	8	0	3.35	3.35
Block 6	17	1	3.76	3.36
Block 7	9	5	5.17	4.25
Block 8	8	1	3.69	3.39
Block 9	14	0	4.30	4.30
Block 10	9	5	3.81	2.76
Block 11	10	1	4.62	4.32
Block 12	12	0	3.98	3.98
Block 13	21	0	4.59	4.59
Block 14	8	0	4.11	4.11
Block 15	8	0	4.74	4.74
Block 16	11	2	4.16	3.42
Block 17	15	3	4.76	4.26
Block 18	8	5	4.10	3.21
Block 19	11	0	4.40	4.40
Block 20	10	1	5.02	4.65
Block 21	10	0	4.37	4.37
Block 22	12	0	4.44	4.44
Block 23	11	3	4.76	4.36
Block 24	11	4	4.55	3.79
Chain	292	0	11.63	11.63

[a] Considering outliers
[b] Without outliers

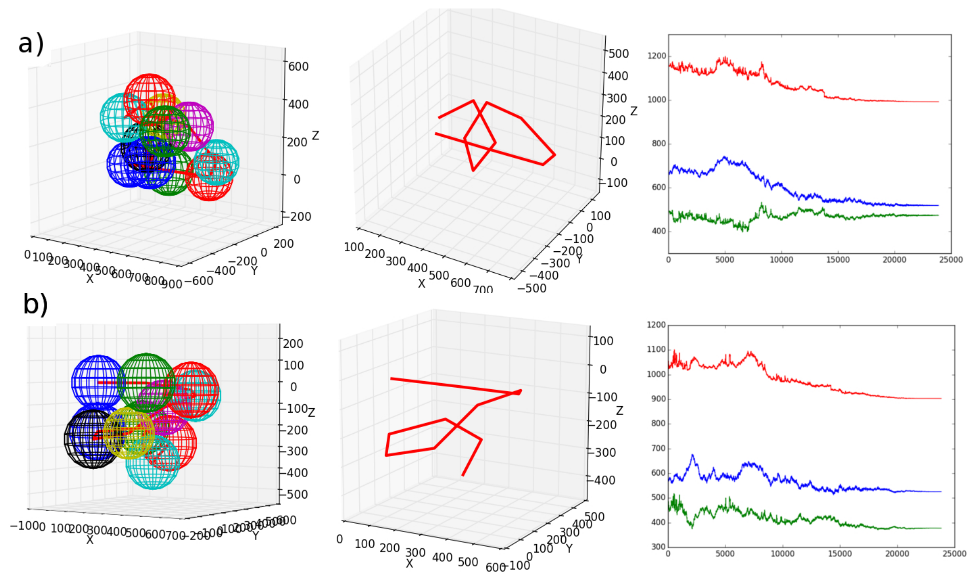

Figure4. Outliers of final energy distribution for block 16. a) First outlier b) second outlier. Plots of final conformations with and without space-beads, and corresponding plots of energy values assumed during annealing: blue, fitness component; green, constraint component; red, total energy.

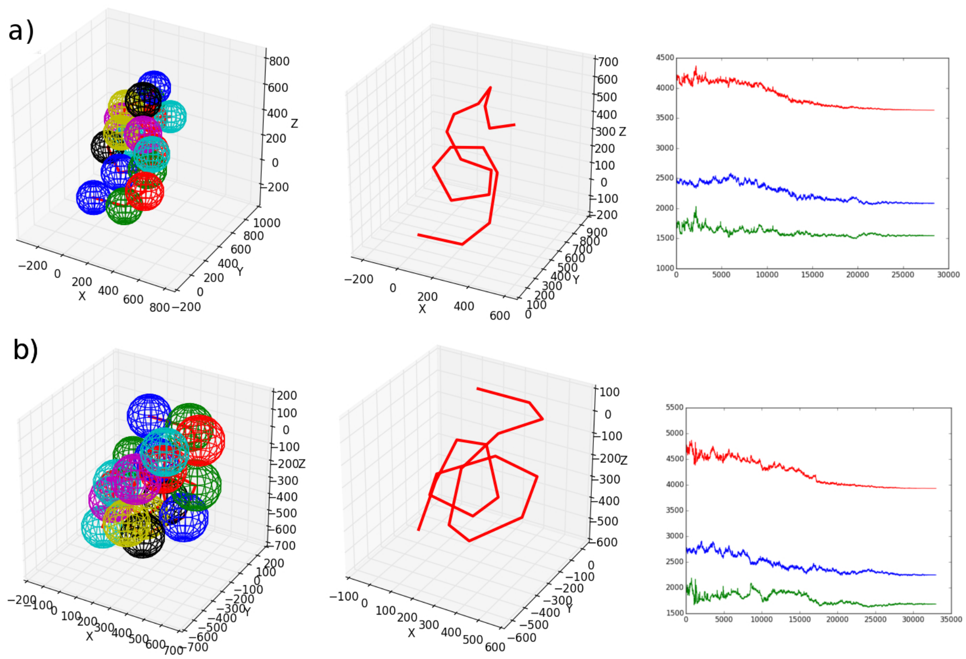

Figure5. Outliers of final energy distribution for block 1. a) First outlier b) second outlier. Plots of final conformations with and without space-beads, and corresponding plots of energy values assumed during annealing: blue, fitness component; green, constraint component; red, total energy.

able to compute the matrix of the distances between all the possible bead pairs. To derive a contact matrix from our 100 distance matrices, we need a criterion to decide when two beads are in contact. We base our criterion on the sum of the radii of the two beads, multiplied by a fixed factor k and taken as a threshold distance to assume that the two beads are touching each other:

$$d_{thresh} = (r_i + r_j) \cdot k \tag{5}$$

If the distance between the centroids of beads i and j is less than d_{thresh}, they are considered in contact; we can create different contact matrices by varying k and, for each subchain considered, by summing up all the binary contact matrices produced by all the distance matrices. We can thus compare the heat map of input contact matrix with the heat maps of our contact matrices. The main diagonal and the first diagonal are removed because their values are much larger than those in the rest of the matrices and their presence makes it difficut to appreciate the variability of the original and reconstructed contact frequencies. In Table 3, we report the values of k for the heat maps represented in Figure 6, 7, 8, 9 and 10.

Table3. Legend of values of k for heat maps of contact matrices.

input	$k = 1.5$	$k = 1.4$	$k = 1.3$
$k = 1.2$	$k = 1.1$	$k = 1.05$	$k = 1.04$
$k = 1.03$	$k = 1.02$	$k = 1.01$	$k = 1$
$k = 0.9$	$k = 0.7$	$k = 0.5$	$k = 0.4$
$k = 0.3$	$k = 0.2$	$k = 0.1$	

Here, we report the heat maps for blocks 13, 1, 16 and 5. We chose these blocks because they are representative of different sizes: block 13 is the biggest (21 beads), block 1 has 18 beads, block 16 has 11 beads and block 5 is one of the smallest, with 8 beads.

Looking at Figures 6, 7, 8, 9 we can appreciate some similarities between heat map of input and heat maps of contact matrices, expecially with $k \geq 1$. Red dots in last heat maps of all figures represent deep interpenetrations ($k \leq 0.3$). Their presence means that sometimes the constraint component of the objective function is overcome by the fitness component, expecially when the contact

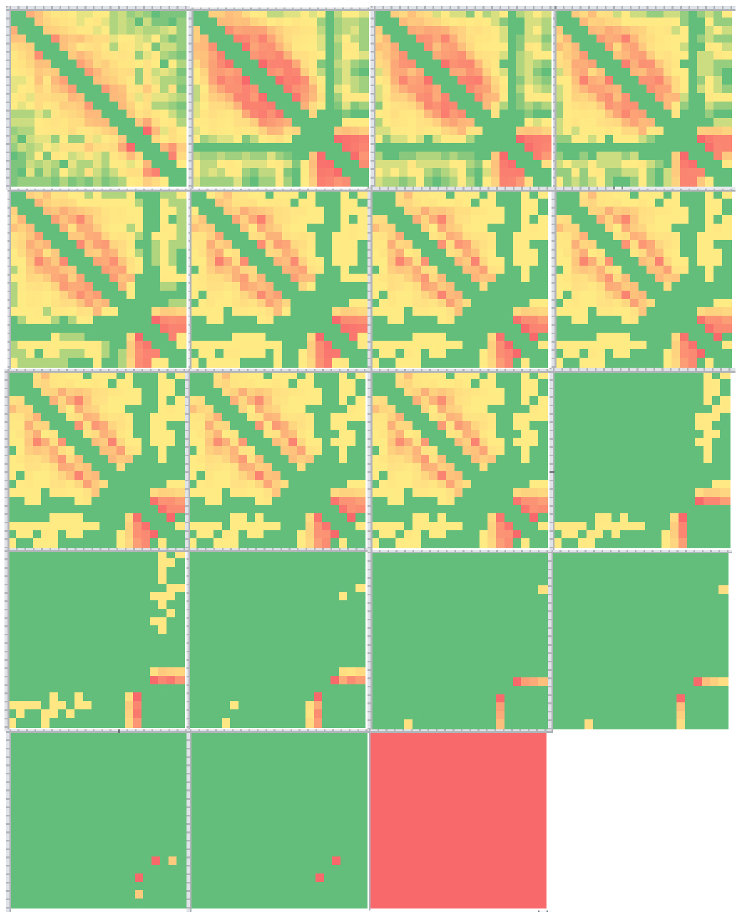

Figure6. Heat maps of contact matrices for block 13 with different values of d_{tresh} (see Table 3).

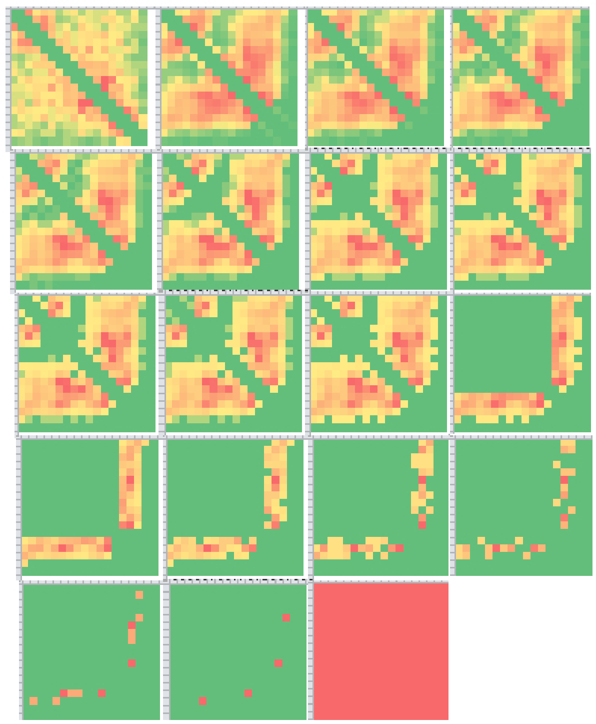

Figure7. Heat maps of contact matrices for block 1 with different values of d_{tresh} (see Table 3).

Figure8. Heat maps of contact matrices for block 16 with different values of d_{tresh} (see Table 3).

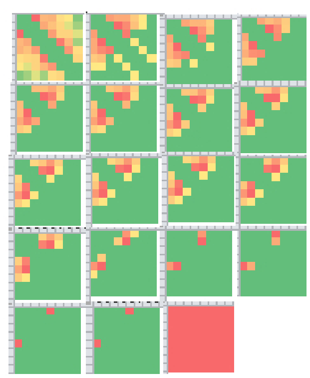

Figure9. Heat maps of contact matrices for block 5 with different values of d_{tresh} (see Table 3).

frequency n_{ij} is high. The number of final conformations with deep interpenetrations is very low ($< 5\%$) in every subchain and in the whole chain.
In Figure 10, we report the corresponding results for the whole chain. We can observe that the similarities between the input heat map and the heat maps of our contact matrices are greater when we choose $k = 1.1$ and $k = 1.2$. The reason could lay in the shape of the neighbor interaction function: note that the blue plot in Figure 1 corresponds to our choice for parameters c and b (see Table 1). The moderate slope interval ranges from 0.8 to 1.2. If the distance between centroids of beads i and j is within this range, they are close and produce a small penalty; this implies that the most reasonable threshold to collect all the contacts of the final conformations is just the one with $k = 1.2$.
From Figure 10, we notice that deeper interpenetrations, highlighted in last five heat maps ($k = 0.5$ to 0.1), are located just ouside blocks. This means that the most serious constraint violations affect the first and the last beads of every block. This is a consequence of the way we shape the higher-level chains: our choice for the physical bead sizes (see point 5 in Section 1) is such that most of the beads of any subchain are contained in the assumed size of the corresponding lower-resolution bead. However, some beads left out, and their presence does not increase the energy during the successive annealing phase, even though they interpenetrate. Looking at the boxplots of EXTs for the 25 low-resolution beads in Figure 11 and the plots in Figure 26, we see that the envelopes of the subchains at level 1 are very different: their magnitude is positively correlated with the genomic block sizes (reported in Table 2). Moreover, blocks with big sizes assume various final conformations and their EXTs can vary greatly from experiment to experiment. This subdivision in blocks with similar sizes would probably avoid variety in envelope sizes, but would disagree with the TAD's theory [4].
In summary, the final conformations derived from the same input assume similar energy values, with little variance with respect of the average, for every subchain and for the whole chain. The algorithm of automatic block detection works well, finding diagonal blocks similar to TADs. Moreover, the results of the simulations are satisfactorily consistent with the initial data: the contact matrices created through Equation 5 to verify the consistency of the algorithm assume patterns sufficiently similar to those of input contact matrix.

4 Features of the solutions

To analyze the geometric features of our solutions, we tried to cluster the 100 results obtained for each subchain and for the whole, full-resolution chain. We based the clustering on two different kinds of data frame: the distance matrices computed as in Section 3, and, for each chain, the mean-squared Euclidean distance between any pair of beads as a function of their genomic distance [1,9]. The latter is an index of compactness. We used the Hierarchical Clustering on Principal Components algorithm (function `HCPC` in R), a hybrid technique combining principal component methods, hierarchical clustering and partitional clustering [10]). As expected, we found that mean square distances are more efficient

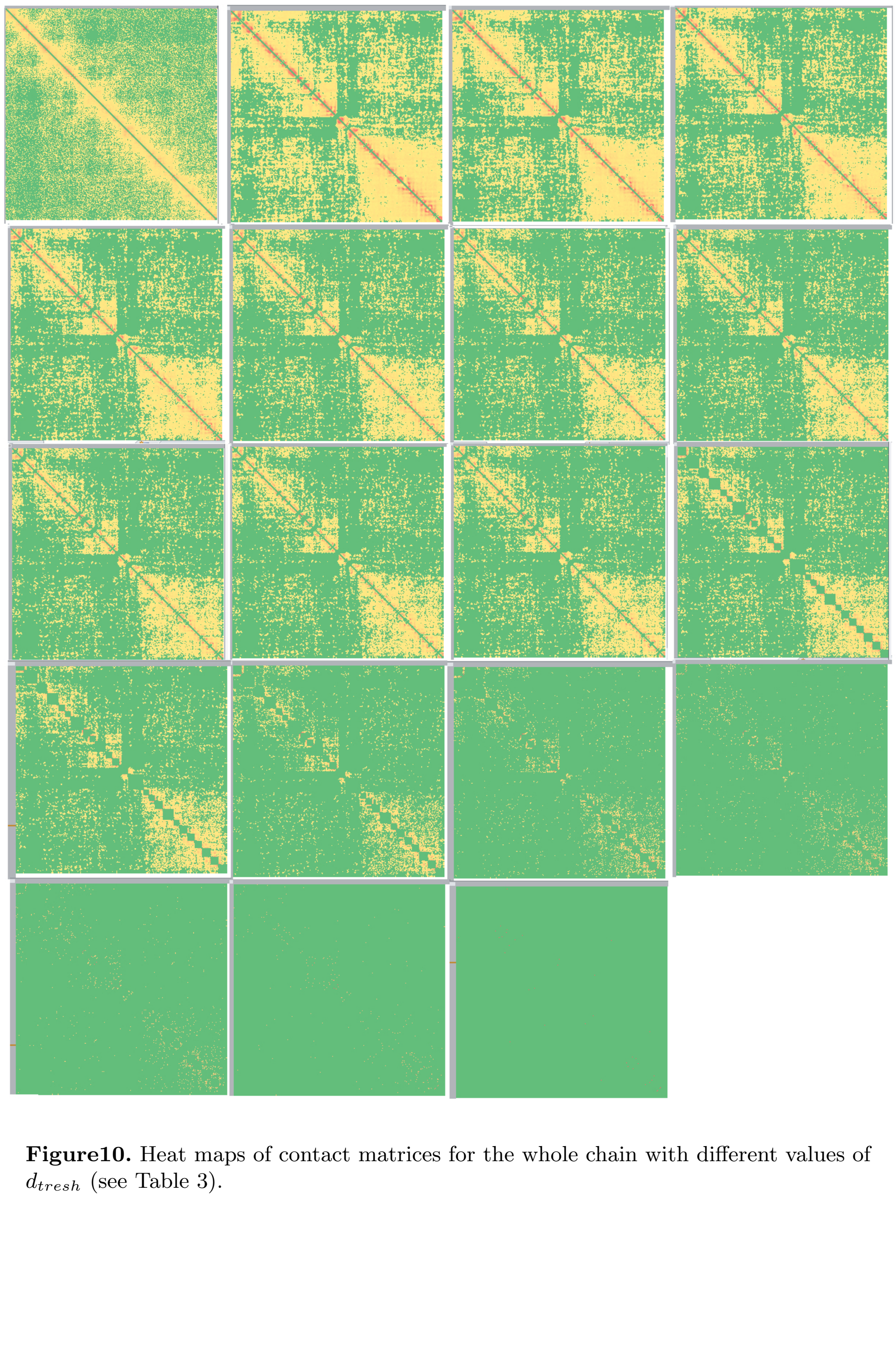

Figure10. Heat maps of contact matrices for the whole chain with different values of d_{tresh} (see Table 3).

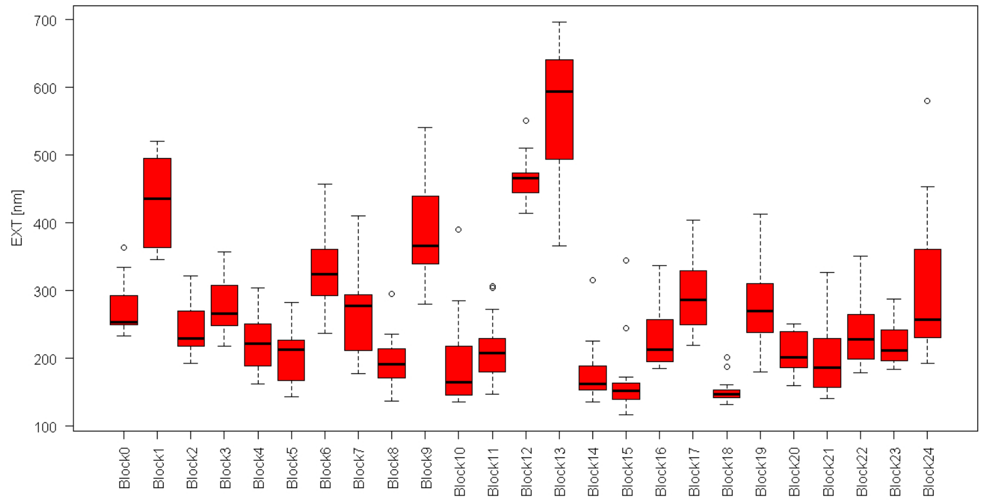

Figure11. Boxplots of EXTs of beads of level 1. Circles represent outliers, red rectangles make the interquartile range, and the central line is the median. The whiskers extend for at most 1.5 times the interquartile distance.

than distance matrices to our purposes, since they highlight topological differences in chromatin wrapping. For this reason, here we only report the results of clustering based on mean-square distances. `HCPC` automatically detects the number of clusters, minimizing the intra-cluster variance and maximizing the inter-cluster variance. In Figures 13, 16, 19, 22, 25, we report, for every cluster, the most representative element (the closest to centroid of the cluster) and the most dissimilar from other clusters (the farthest from centroids of other clusters). We have performed the boxplots of mean square Euclidean distances as a function of genomic distances and the Hierarchical Clustering on Principal Components for the whole chain and for every subchain.

4.1 Geometric features of subchains

In this Subsection, we list the results described above for blocks 13, 1, 16 and 5, chosen for the same reason as in Section 3. For each block, we report the pairs in set $\mathcal{L}$, the boxplots of mean square Euclidean distances, a representation of the clusters detected on the boxplot means, and some significant configurations for each cluster. Along with the final configurations, we also report the energy plots during the annealing, as already done in Figures 4 and 5.

Block 13: 21 beads
Relevant pairs: (0, 1), (0, 2), (0, 3), (0, 4), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (1,

7), (2, 3), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5), (4, 6), (5, 6), (5, 7), (6, 7), (6, 8), (6, 15), (7, 8), (7, 9), (7, 10), (7, 17), (8, 9), (8, 10), (8, 11), (8, 12), (8, 13), (8, 14), (8, 15), (8, 16), (8, 17), (8, 18), (8, 19), (8, 20), (9, 10), (9, 11), (9, 12), (9, 13), (9, 14), (9, 15), (9, 17), (9, 18), (9, 19), (9, 20), (10, 11), (10, 12), (10, 13), (10, 14), (10, 15), (10, 16), (10, 17), (10, 18), (10, 19), (10, 20), (11, 12), (11, 13), (11, 14), (11, 15), (11, 16), (11, 17), (11, 18), (12, 13), (12, 14), (12, 15), (12, 16), (12, 17), (13, 14), (13, 15), (13, 17), (14, 15), (14, 17), (14, 18), (14, 19), (15, 16), (15, 17), (15, 18), (15, 19), (16, 17), (17, 18), (18, 19), (18, 20), (19, 20).

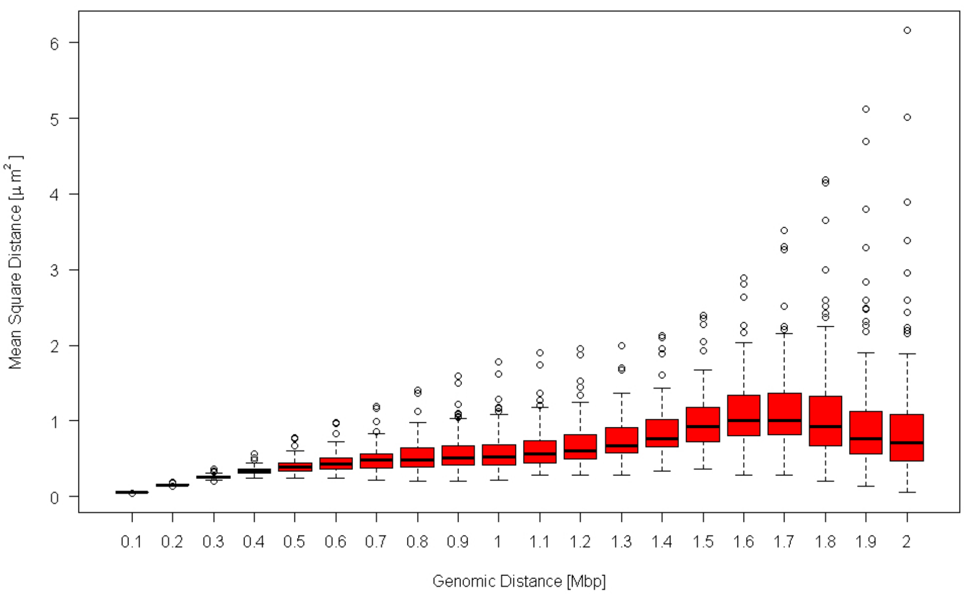

Figure12. Boxplots of mean square Euclidean distances as a function of genomic distances of bead pairs in Block 13.

Block 1: 18 beads

Relevant pairs: (0, 1), (0, 2), (0, 3), (1, 2), (1, 3), (2, 3), (2, 7), (3, 4), (3, 5), (3, 6), (3, 7), (3, 9), (4, 5), (4, 6), (4, 10), (5, 6), (5, 7), (5, 9), (5, 10), (5, 16), (6, 7), (6, 9), (6, 10), (6, 11), (7, 8), (7, 9), (7, 10), (7, 12), (8, 9), (8, 10), (8, 12), (9, 10), (9, 11), (9, 12), (9, 13), (9, 16), (10, 11), (10, 12), (10, 13), (10, 14), (10, 15), (10, 16), (10, 17), (11, 12), (11, 13), (11, 14), (11, 15), (11, 16), (11, 17), (12, 13), (12, 14), (12, 15), (12, 16), (12, 17), (13, 14), (13, 15), (13, 16), (13, 17), (14, 15), (14, 16), (14, 17), (15, 16), (15, 17), (16, 17).

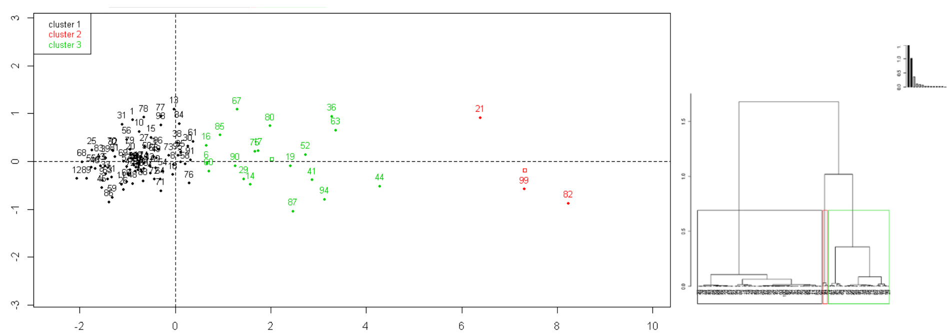

Figure13. HCPC by mean square Euclidean distances as function of genomic distances of final configurations of block 13. In the upper right the weight of principal components.

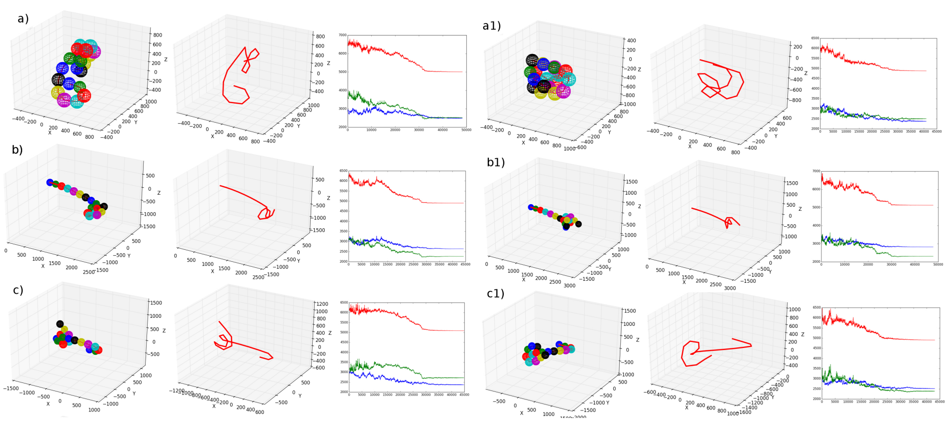

Figure14. Clustering by mean square Euclidean distances as function of genomic distances for block 13. a) The most representative element of cluster 1. a1) The most dissimilar element of cluster 1 from clusters 2 and 3. b) The most representative element of cluster 2. b1) The most dissimilar element of cluster 2 from clusters 1 and 3. c) The most representative element of cluster 3. c1) The most dissimilar element of cluster 3 from clusters 1 and 2.

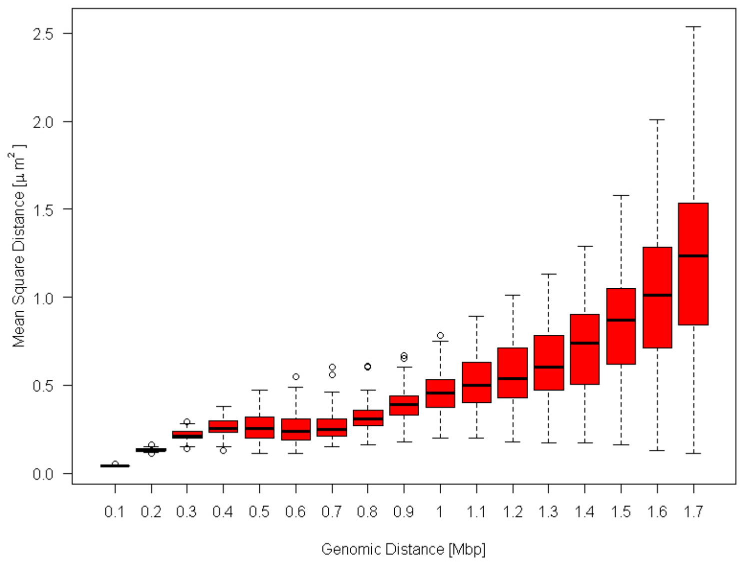

Figure15. Boxplots of mean square Euclidean distances as a function of genomic distances of bead pairs in Block 1.

Block 16: 11 beads
Relevant pairs: (0, 1), (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5), (4, 6), (4, 7), (5, 6), (5, 7), (5, 9), (5, 10), (6, 7), (6, 8), (7, 8), (7, 9), (7, 10), (8, 9), (8, 10), (9, 10).

Block 5: 8 beads
Relevant pairs: (0, 1), (0, 2), (1, 2), (2, 3), (3, 4), (3, 5), (3, 6), (4, 5), (4, 6), (5, 6), (5, 7), (6, 7).

4.2 Geometric features of the whole chain

In Figure 24, we show the boxplots of the mean square Euclidean distance between pairs of beads as a function of their genomic distance for the whole chain of 292 beads. `HCPC` finds 3 clusters (Figure 25). In Figure 26 and 27, we have plotted, for every cluster, the most representative configuration and the most dissimilar from the other clusters. In Figure 25 we have represented the final configurations with the low resolution representation (25 beads), in Figure 27

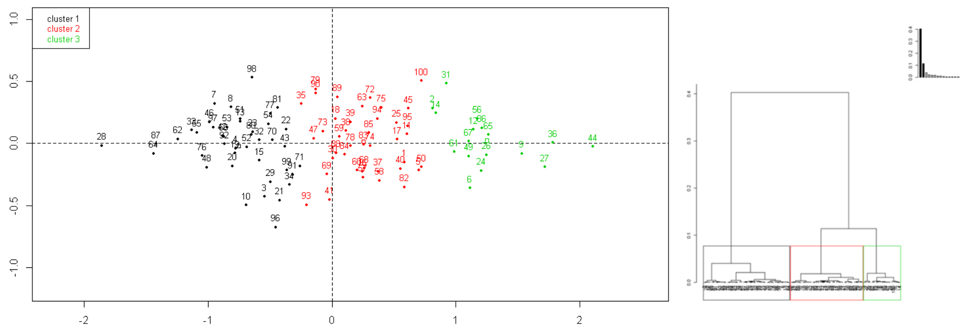

Figure16. `HCPC` by mean square Euclidean distances as function of genomic distances of final configurations of block 1. In the upper right the weight of principal components.

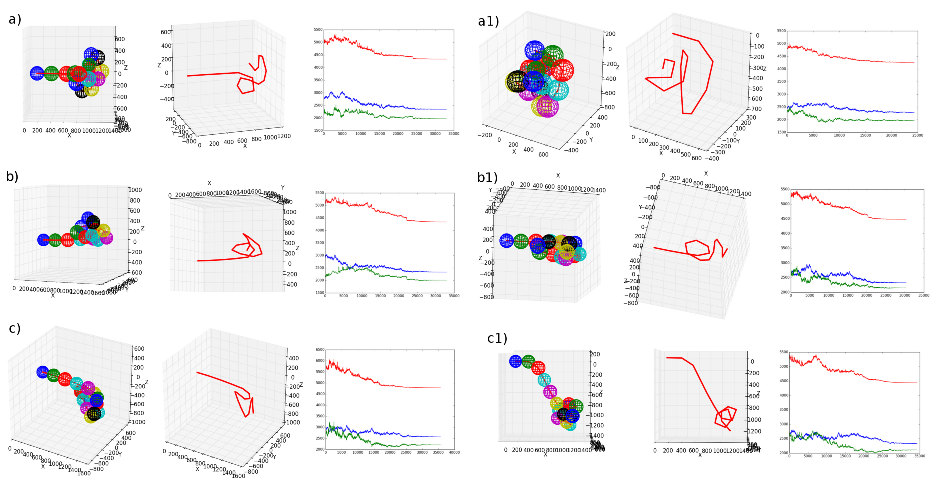

Figure17. Clustering by mean square Euclidean distances as function of genomic distances for block 1. a) The most representative element of cluster 1. a1) The most dissimilar element of cluster 1 from clusters 2 and 3. b) The most representative element of cluster 2. b1) The most dissimilar element of cluster 2 from clusters 1 and 3. c) The most representative element of cluster 3. c1) The most dissimilar element of cluster 3 from clusters 1 and 2.

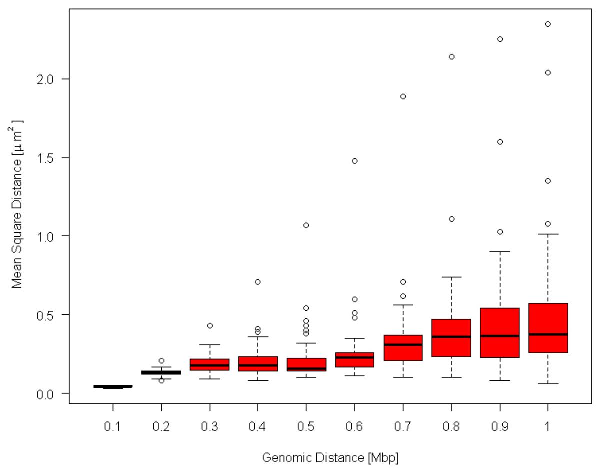

Figure18. Boxplots of mean square Euclidean distances as a function of genomic distances of bead pairs in Block 16.

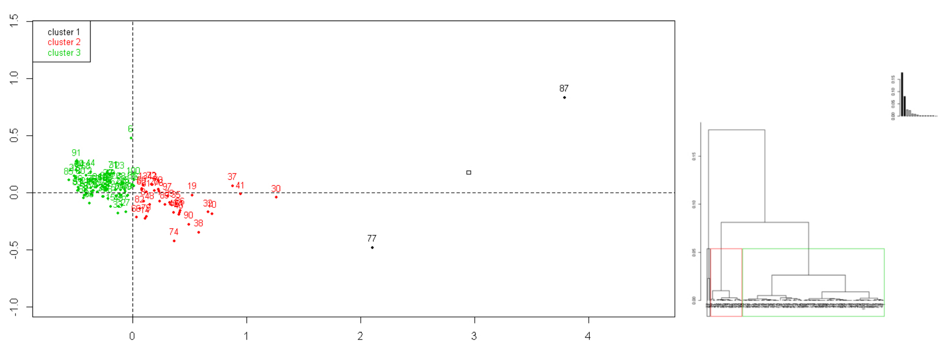

Figure19. `HCPC` by mean square Euclidean distances as function of genomic distances of final configurations of block 16. In the upper right the weight of principal components.

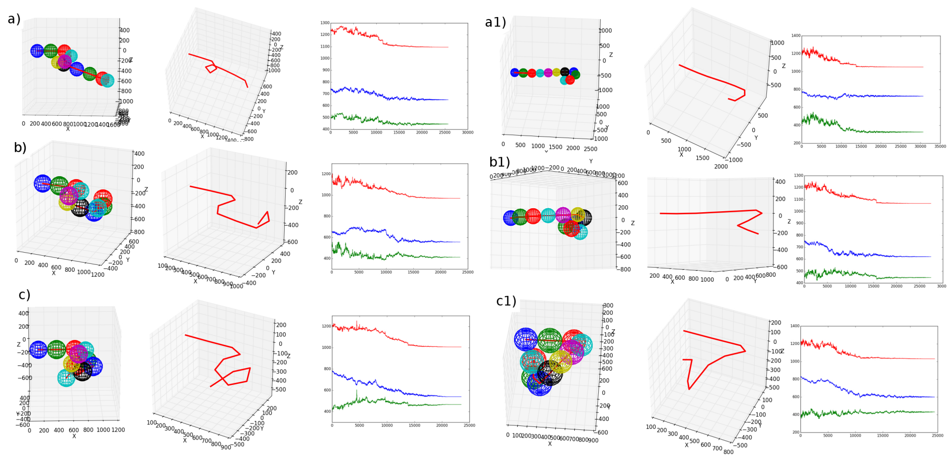

Figure20. Clustering by mean square Euclidean distances as function of genomic distances for block 16. a) The most representative element of cluster 1. a1) The most dissimilar element of cluster 1 from clusters 2 and 3. b) The most representative element of cluster 2. b1) The most dissimilar element of cluster 2 from clusters 1 and 3. c) The most representative element of cluster 3. c1) The most dissimilar element of cluster 3 from clusters 1 and 2.

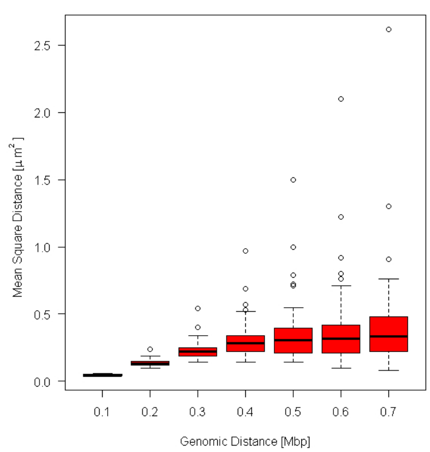

Figure21. Box plots of mean square Euclidean distances as a function of genomic distances of bead pairs in Block 5.

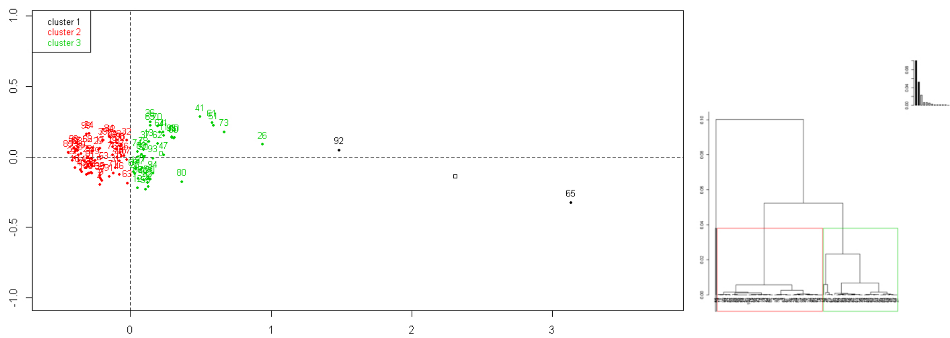

Figure22. `HCPC` by mean square Euclidean distances as function of genomic distances of final configurations of block 5. In the upper right the weight of principal components.

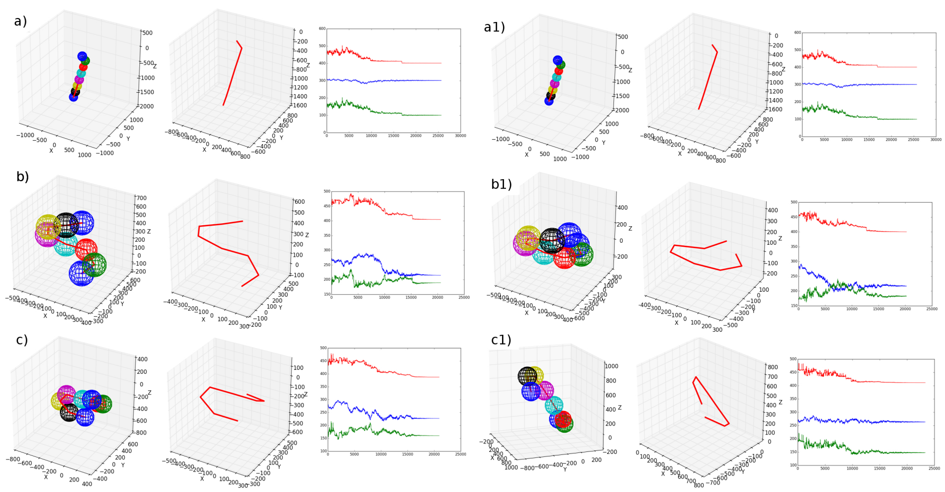

Figure23. Clustering by mean square Euclidean distances as function of genomic distances for block 5. a) The most representative element of cluster 1. a1) The most dissimilar element of cluster 1 from clusters 2 and 3. b) The most representative element of cluster 2. b1) The most dissimilar element of cluster 2 from clusters 1 and 3. c) The most representative element of cluster 3. c1) The most dissimilar element of cluster 3 from clusters 1 and 2.

the same configurations are represented reconstructing the subchains at higher resolution (292 beads).

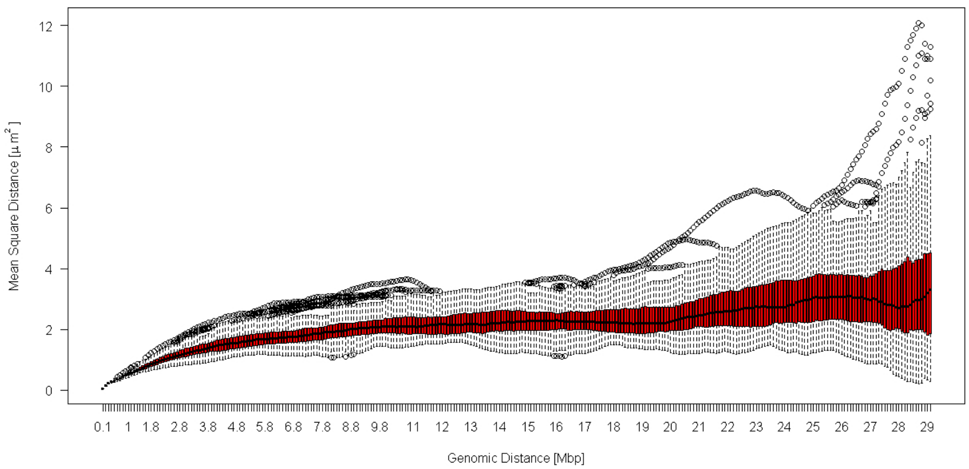

Figure24. Boxplots of mean square Euclidean distances as a function of genomic distances of bead pairs in the whole chain.

5 Conclusions

The 3D configurations obtained through energy function (1) and the current parameters appear biologically compatible with what we know on chromosome configuration: presence of supercoils, helices, TADs, and envelopes compatible with the radius of the nucleus [4, 7, 8]. We have created contact matrices whose heatmaps present empty regions. This can depend on both the relatively small number of solutions and our objective function, which neglects the low contact frequencies. Moreover, an analysis of the structures corresponding to intermediate penalty values has not been done. Some of these structures could contribute consistently to the final contact matrix. A problem is that sometimes consecutive beads interpenetrate; this behavior probably depends on the objective function, since the data fit part has a trivial minimum where all the centroids in $\mathcal{L}$ coincide. To overcome this drawback, choosing the contact of surfaces as contact condition could be sufficient. Another critical point of the present version of the algorithm is the evaluation of λ. The procedure $ANNEALING\lambda$ (see Table 1 in Section 1) only applies on unconstrained configurations, which have very scattered values of energy. In the 1000 steps dedicated to calculation of λ, the chain starts from the completely stretched configuration and arrives at a very

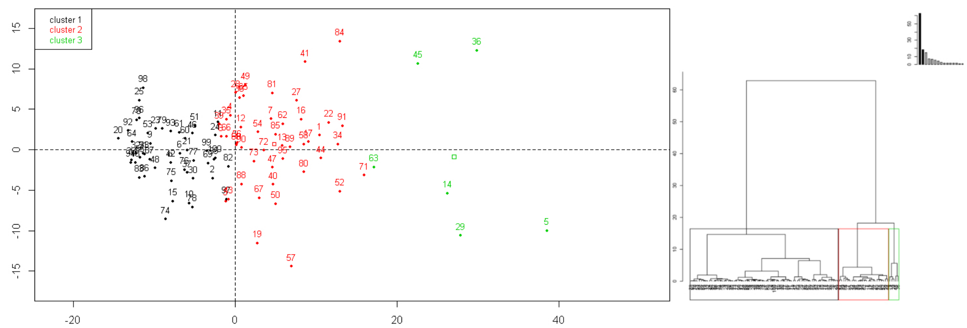

Figure25. HCPC by mean square Euclidean distances as function of genomic distances of final configurations of the whole chain. In the upper right the weight of principal components.

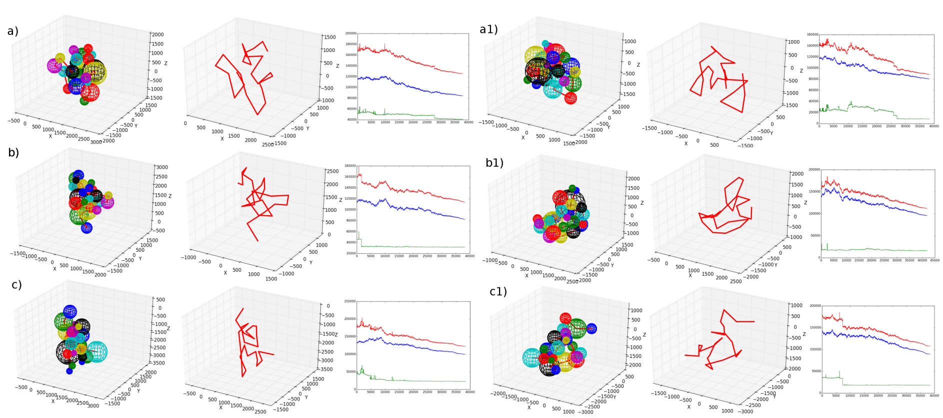

Figure26. Clustering by mean square Euclidean distances as function of genomic distances for the whole chain at lower-resolution level (25 beads). a) The most representative element of cluster 1. a1) The most dissimilar element of cluster 1 from clusters 2 and 3. b) The most representative element of cluster 2. b1) The most dissimilar element of cluster 2 from clusters 1 and 3. c) The most representative element of cluster 3. c1) The most dissimilar element of cluster 3 from clusters 1 and 2.

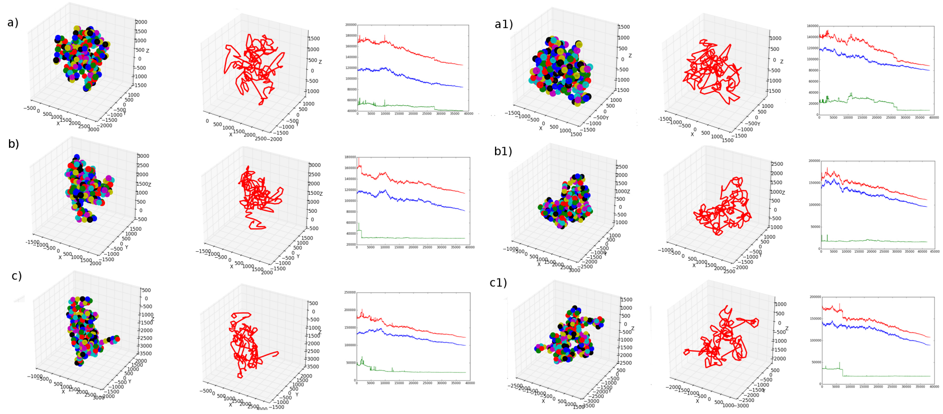

Figure27. Clustering by mean square Euclidean distances as function of genomic distances for the whole chain at higher-resolution level (292 beads). a) The most representative element of cluster 1. a1) The most dissimilar element of cluster 1 from clusters 2 and 3. b) The most representative element of cluster 2. b1) The most dissimilar element of cluster 2 from clusters 1 and 3. c) The most representative element of cluster 3. c1) The most dissimilar element of cluster 3 from clusters 1 and 2.

knotted one, with a lot of interpenetrations. In the experiments presented here, λ was calculated on the energy values within the 80-th percentile, and the result seems to be very inaccurate. We should choose a lower percentile and start the warm-up phase after resetting the chain configuration.

References

1. Caudai,C. et al. (2015): Inferring 3D chromatin structure using a multiscale approach based on quaternions, BMC Bioinformatics, 16: 234.
2. Caudai, C. et al. (2015): A Statistical Approach to Infer 3D Chromatin Structure, Mathematical Models in Biology, Springer, 161-171.
3. Dekker,J. et al. (2002): Capturing chromosome conformation. Science 295: 1306-1311.
4. Dixon,J.R. et al. (2012): Topological domains in mammalian genomes identified by analysis ofchromatin interactions, Nature 485: 376-380.
5. Hu,M. et al. (2013): Bayesian inference of Spatial organizations of chromosomes, PLOS Comp. Biol. 9, 1002-893.
6. Husson, F. et al. (2010): Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data? ,Tech. Report Agrocampus Ouest, Rennes.
7. Lieberman-Aiden,E. et al. (2009): Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome, Science 326: 289-293.
8. Rousseau,M. et al. (2011): Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling, BMC Bioinformatics 12: 414-429.
9. Mateos-Langerak, J. et al. (2009): Spatially confined folding of chromatin in the interphase nucleus, PNAS 106: 3812-3817.
10. Husson, F. et al. (2010): Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data? ,Tech. Report, Agrocampus Ouest, Rennes.