Estimation of the Spatial Chromatin Structure Based on a Multiresolution Bead-Chain Model

Claudia Caudai, Emanuele Salerno, *Senior Member, IEEE,* Monica Zoppè, and Anna Tonazzini

Abstract—We present a method to infer 3D chromatin configurations from Chromosome Conformation Capture data. Quite a few methods have been proposed to estimate the structure of the nuclear DNA in homogeneous populations of cells from this kind of data. Many of them transform contact frequencies into Euclidean distances between pairs of chromatin fragments, and then reconstruct the structure by solving a distance-to-geometry problem. To avoid inconsistencies, our method is based on a score function that does not require any frequency-to-distance translation. We propose a multiscale chromatin model where the chromatin fibre is suitably partitioned at each scale. The partial structures are estimated independently, and connected to rebuild the whole fibre. Our score function consists in a data-fit part and a penalty part, balanced automatically at each scale and each subchain. The penalty part enforces *soft* geometric constraints. As many different structures can fit the data, our sampling strategy produces a set of solutions with similar scores. The procedure contains a few parameters, independent of both the scale and the genomic segment treated. The partition of the fibre, along with intrinsically parallel parts, make this method computationally efficient. Results from human genome data support the biological plausibility of our solutions.

Index Terms—Chromosome conformation capture, Chromatin configuration, Bayesian estimation.

1 Introduction

The mechanisms underlying DNA packing are not entirely described or understood, although it is increasingly apparent that chromosomal organization is one of the factors involved in regulation of gene function. Indeed, several processes during all the phases of cell life are regulated through epigenetic changes, that is, without altering the DNA sequence. At present, it is generally accepted that some factors governing epigenetic changes are related to the spatial organization of chromatin, which is the complex formed by the DNA double helix plus the macromolecules that mediate its packing. As stated in [1], "The genome forms extensive and dynamic physical interactions in the form of chromatin loops and bridges, which bring distal elements of the chromosome into close physical proximity, with potential consequences for gene expression and/or propagation of the genome." For the dynamic nature of these processes, a study of the three-dimensional structure of chromatin cannot rely on a static view of the cell.

A step towards an understanding of the chromatin geometry has come with the fluorescence *in-situ* hybridization technique (FISH, [2], [3]). A major boost in chromatin studies, however, comes with the next-generation sequencing techniques, which enabled a number of methods of the type *Chromosome Conformation Capture* (3C, 4C, 5C, Hi-C, see [4], [5], [6], [7], [8]). These techniques count the contacts between all the possible pairs of chromatin fragments in the nuclei of a homogeneous population of cells. Depending on the restriction enzymes used, their genomic resolution can be very high (typically, a few kbps). Resulting from aggregate counts on millions of cells, however, the genome-wide contact frequency matrices thus obtained are subject to sparseness and poor repeatability. This is why they are normally binned to lower resolutions (typically, 100 kbp). An experimental protocol called Single-Cell Hi-C [9], [10], [11], [12], [13], [14], where the counts are not aggregated, has confirmed the results of aggregated Hi-C. Through this technique, it was also found that the intra-chromosomal structures are stable across different cells, whereas inter-chromosomal interactions have a larger variability.

The aggregated Hi-C data tell us which chromatin fragments are often close to each other in a population of cells. Conversely, single-cell Hi-C tells us which fragments are in mutual contact in a particular cell. Both these types of data are cues to derive the 3D structure of the chromatin. As the contacts cannot be related directly to geometry, however, this is not an easy task. A number of solutions to this problem have been proposed, from either aggregated or single-cell data. A possible classification of these methods is based on their outputs: finding either a unique "consensus" configuration, or a set of configurations compatible with the data. In our view, a unique output configuration is significant when based on single-cell data [13], [14], [15], [16], [17]. It becomes less significant with aggregated data, as these support a variety of individual configurations. Finding a set of consistent solutions, on the other hand, allows us to study the most frequent configurations assumed during the lifecycle of a specified type of cell. The algorithms proposed differentiate for the data model, the solution model and the computational strategy adopted.

For the data model, a popular approach assumes a close relationship between the contact frequencies and the Euclidean distances between pairs of fragments (see, $e.g.$,

- *C. Caudai, E. Salerno and A. Tonazzini are with the National Research Council of Italy, Institute of Information Science and Technologies, Pisa, Italy.*
  E-mail: name.surname@isti.cnr.it
- *M. Zoppè is with the National Research Council of Italy, Institute of Clinical Physiology, Pisa, Italy.*
  E-mail: mzoppe@ifc.cnr.it

Manuscript received , ; revised .

[18]), assuming the latter as the input data. Pairs of fragments that are frequently in contact are likely to be spatially close, whereas pairs of fragments with a few contacts are assumed to be farther apart. This allows the problem to be treated as a classical distance-to-geometry problem. Some of the frequency-to-distance transformations proposed are deterministic and derived from a mix of empirical and theoretical considerations [4], [19], [20], whereas others are probabilistic [21]. Some approaches [22], [23], [24], [25] do not assume frequency-distance relationships, or only use physical distances to restrain the solutions.

The solution models proposed include: bead-chain [20]; piecewise-linear-curve [19]; molecular dynamics or polymer models [9], [20], [24], [26], [27], [28]. Common constraints derive from known geometrical and topological properties of the chromatin fibre [20], polymer physics [26], or flexible target distances obtained from Hi-C data [9].

The first computational strategies proposed are based on constrained optimization, that is, they search the solution that best fits the data, subject to the model constraints [4], [19], [20], [29]. The well-known drawbacks of constrained optimization in high-dimensional spaces motivated a number of probabilistic approaches, ranging from Markov Chain Monte Carlo sampling [30] to fully Bayesian approaches [27], [31], [32]. In particular, the method proposed in [32] is able to accept multiple contact matrices establishing a generalized linear model with Poissonian statistics for the data, and solves the problem by minimizing a score function penalized by the Euclidean distances. Another approach consists in adjusting the parameters of a polymeric model to fit the contact data (see, *e.g.* [24]). A more comprehensive and reasoned account of all these aspects can be drawn, for example, from [33], [34], [35], [36], [37], [38], [39], [40], [41].

Studying this problem, we chose to use population-based Hi-C data, with the aim at conceiving a method to obtain a rich set of compatible configurations. We posed two basic requirements: 1) Avoid any translation from contact frequencies to Euclidean distances; 2) Produce an efficient and scalable algorithm, allowing for flexible geometrical constraints and data partitioning. The first requirement addresses the appropriateness of frequency-distance translation: despite the intuitive meaning of this choice, we observe that two fragments that are seldom in contact can well be spatially close, and that the distances derived from real data through popular rules are often severely incompatible with the Euclidean geometry [42], [43]. The second requirement is motivated by the potentially enormous amount of possible pairs to be processed, and the need to narrow the space of the feasible solutions.

The resulting approach is characterized by a multiresolution, recursive, modified bead-chain model for the chromatin fibre. The multiresolution choice is motivated by the existence of genomic regions with many internal contacts and poor interactions with the rest of the genome. This feature is found practically at all the scales [34], [44]: at 1-10 Mbp scale, the *genomic compartments* show this property; at smaller scales (100 kbp and below), structures that behave similarly are referred to as *topological association domains* (or TADs). Since the related chromatin segments interact very weakly with the rest of the genome, their structures can be estimated irrespective of the rest of the data. All the segments thus estimated become individual elements of a lower-resolution chain, and the procedure is repeated recursively until no further partition is possible. Once computed, all the elements of the lowest-resolution chain have a higher-resolution counterpart that can be used to build a more detailed estimate and, recursively, the entire structure at its highest resolution. To exploit this possibility within a bead-chain chromatin model, we devised a bead structure that permits to align properly the different subchains in the final solution. The beads in this model are not simple spheres, but complex elements characterized by endpoints and centroids. This model is then evolved using a score function that assumes its lowest values for the most likely configurations of the bead chain under study. Our strategy is different from the ones based on molecular dynamics or polymer models, since the requirements we put in the score function are purely geometric: no physical driving force is involved in the model evolution. The score function we proposed in [45], only included the fit to the data, measured through the closeness of the most frequently interacting pairs of beads. The other beads were only subject to *a posteriori* constraints imposed "rigidly" to avoid unfeasible solutions: any such solution was simply discarded from the final configuration set. The results obtained were promising, but the algorithm presented a number of drawbacks in enforcing the constraints and setting the sampling rules at different scales.

To improve our results, we propose here a method characterized by a) A modified bead-chain model applicable directly to all the scales, equipped with an approximated size for each individual bead; b) A solution space generated by a score function including "soft" constraints that penalize gradually the unlikely configurations rather than rejecting them. The latter is sampled by an annealing scheme [46] that is not intended to find a global optimum but to produce a number of results that are compatible with both the data and the constraints. The candidate solutions are evolved through quaternion operators [47], which offer computational advantages over the classical rotation matrices using Euler angles [45]. The code implementing this method is written in Python and is called ChromStruct. The current version (3.2) is available here as supplemental material. The free parameters to be set are very few, and this allows the solutions to be controlled easily and effectively.

In Section 2, we give the details of our solution model, score function and recursive procedure. In Section 3, we analyze the performance of the method, the features of some results obtained from real data, and their correlation with known biological properties, along with some comparisons with the methods proposed in [45] and [48]. Final remarks on the current results and possible future directions are given in the conclusion.

2 Methods

2.1 Chromatin model

As mentioned, our bead-chain model partitions the chromatin fibre in weakly-interacting domains. Following a method suggested in [34], our algorithm identifies the corresponding diagonal blocks in the contact matrix through the

relative minima of the moving average on suitably-sized off-diagonal triangles. From each extracted block, the structure of a chromatin segment is estimated. This corresponds to a subchain in our overall model, and is modeled as a single bead in a coarser-scale chain. Its internal contacts sum up to the total contacts in the corresponding block, and become a single diagonal entry of the data matrix at the successive scale.

At each scale, a bead is modeled through three fixed points equipped with an approximate size (the blue dots and the diameters of the green spheres, respectively, in Fig. 1). The three points are connected by two segments, identifying a centroid and two endpoints. The centroid coordinates are computed by averaging the coordinates of the centroids in the finer-scale subchain (red dots in Fig. 1), and the endpoints coincide with the centroids of the first and the last beads in the finer-scale subchain. The approximate size is estimated by a principal component analysis of the centroid coordinates in the finer-scale subchain. To motivate this strategy, let us consider the covariance matrix of a point set in the 3D space. The square root of its maximum eigenvalue is proportional to the extent of the point set along its first principal component (if all the coordinates are uniformly scaled, the eigenvalues are scaled by the square of the scale factor). The real-world subchain is far from being an impenetrable spheroid: our approximate size is only intended to prevent the chain from being too packed. Thus, we set it as a fraction of the square-rooted largest eigenvalue, and find the appropriate fraction by trial and error, checking the properties (*e.g.* the overall size) of the output chain. In our experience, this is not difficult. By effect of our score function, the approximate bead sizes act gradually as repulsive forces between pairs of beads. At the finest scale, that is, at the first recursion level, the bead structures are not known, so the first-level beads can only be modeled as spheres, whose radii can be approximated from the number of their internal contacts. Intuitively, indeed, many internal contacts mean that a fragment is packed tightly, whereas a few contacts mean that it is relatively stretched out. Even at the finest scale, however, each bead is modeled by the usual triple. The difference from the higher levels is that the three points are collinear, and the distances between the centroid and the two endpoints are both equal to the estimated radius.

The beads are linked respecting their genomic order, and the second endpoint of each bead coincides with the first endpoint of the successive bead. Figure 1 illustrates how four consecutive subchains are modeled as modified beads and then connected to form a chain at a lower resolution. The lengths of the segments joining the endpoints with the centroid, and the angle they form, are not changed during the evolution of the model. Conversely, the planar and dihedral angles defining the position of each bead with respect to the adjacent ones are perturbed at each iteration, subject to constraints that establish chain flexibility and mutual distance ranges. Our score function allows us to avoid special constraints on angles, as the bead sizes alone provide a good control.

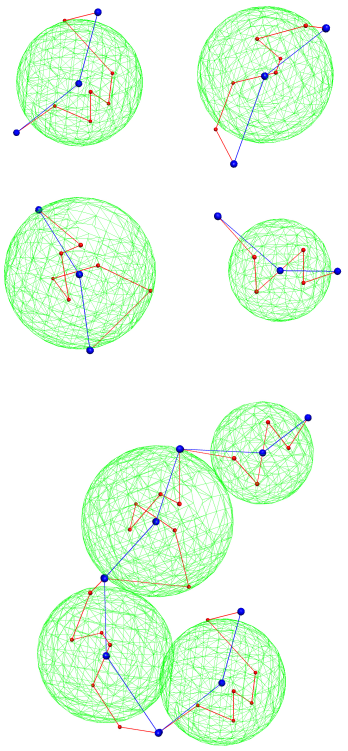

Fig. 1. **Illustrative picture of the modified bead-chain model.** Top: Four chromatin fragments, represented as bead sequences (small red dots), and as centroid-endpoints triples (bigger blue dots). The green sphere wireframes represent the assumed sizes for the beads at the immediately coarser scale. Bottom: Connected chain composed by the fragments above, properly located and rotated. The sequence of spherical wireframes represents the coarser-scale chain.

2.2 Score function

Following a classical approach in inverse problems [49], the solution space for each subchain to be estimated is generated by a score function with the following form:

$$\Xi(\mathcal{C}) = \Phi(\mathcal{C}) + \lambda\Psi(\mathcal{C}) \tag{1}$$

where $\mathcal{C}$ is the chain configuration, Φ and Ψ are the data-fit and the constraint terms, respectively, and λ is a parameter that balances their influence. The two terms are sums of positive contributions, φ_{ij} and ψ_{ij}, penalizing unlikely interactions between the beads indexed by i and j.

For the data fit term, we make φ_{ij} proportional to the squared distance between the i-th and the j-th beads; the summation range, however, is limited to a suitably chosen subset $\mathcal{L}$ of highly interacting pairs in $\mathcal{C}$:

$$\Phi(\mathcal{C}) = \sum_{i,j\in\mathcal{L}} \varphi_{ij} = \sum_{i,j\in\mathcal{L}} n_{ij}[d_{ij} - (r_i + r_j)]^2 \tag{2}$$

In Eq. (2), n_{ij} is the contact frequency between the i-th and j-th beads ($i \neq j$), d_{ij} is the distance between their centroids, and r_i and r_j are their radii. To populate $\mathcal{L}$ for each subchain, we select the pairs exceeding a pre-defined percentile of the contact frequencies in the related block. This strategy can be advantageous in presence of biased data [36], since the smallest contact frequencies, percentually, are more affected by biases. The factors n_{ij} give different weights to the different pairs, thus letting the most frequent

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2791439, IEEE/ACM Transactions on Computational Biology and Bioinformatics

contacts to predominate. Function Φ becomes small when the centroid-centroid distances of the pairs in $\mathcal{L}$ nearly equal the sums of the radii of the related beads. Conversely, solutions with large distances between pairs in $\mathcal{L}$ are quadratically penalized. When any two beads in $\mathcal{L}$ interpenetrate, the corresponding term in brackets becomes negative. The maximum data-fit penalization of this situation occurs when $d_{ij} = 0$, and φ_{ij} assumes the finite and unmodifiable value $n_{ij}(r_i+r_j)^2$. The constraint term Ψ is needed to both control this penalization and extend it to all the pairs in $\mathcal{C}$. We let

$$\Psi(\mathcal{C}) = \sum_{i,j \in \mathcal{C}} \psi_{ij} = \\ = \sum_{i,j \in \mathcal{C}} \frac{r_i + r_j}{2d_{ij}} \left[1 - \frac{\{c[d_{ij} - (r_i + r_j)]\}^b}{1 + \{c|d_{ij} - (r_i + r_j)|\}^b} \right] \quad (3)$$

where c is a scale factor that makes the terms in braces dimensionless, and the exponent b is an odd natural. A plot of ψ_{ij} is shown in Fig. 2. Note that: a) for d_{ij} near zero, ψ_{ij} behaves as $(r_i+r_j)/d_{ij}$; b) in an interval around (r_i+r_j), ψ_{ij} behaves as $(r_i+r_j)/(2d_{ij})$; c) for d_{ij} sufficiently larger than $(r_i + r_j)$, ψ_{ij} vanishes rapidly. Factor c tunes the width of the intermediate interval: large values of c make it narrow. Exponent b, in turn, tunes the slope of ψ_{ij} in the transition zones: large values of b produce abrupt transitions. Function (3) is thus intended to prevent any two beads from interpenetrating more than some fraction of their sizes. Since the bead structures are always known approximately, imposing this requirement rigidly, as done in [45], could exclude good configurations from the feasible solutions, besides making more difficult to enforce the constraints coherently at the different scales. The position of each bead in a pair is penalized gradually as a function of its distance from the other bead. As moderate interpenetrations between adjacent beads are allowed, provided that c and b are chosen wisely, constraining the mutual angles between adjacent beads can become unnecessary. It is useful to note that this strategy is completely different from establishing target distances increasing for decreasing values of n_{ij}. Function Ξ does not enforce large distances between low-contact-frequency beads, but lets them unconstrained, compatibly with the actual genomic sequence and the interference with other beads in the chain.

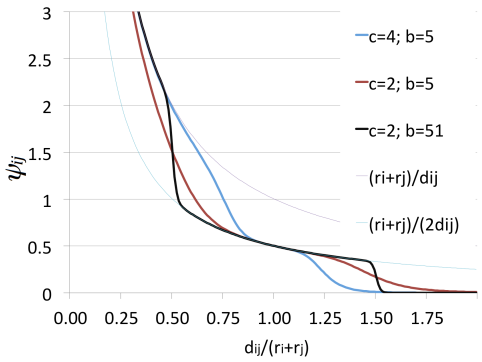

Fig. 2. **Function ψ_{ij} in (3), plotted for a few values of c and b.** The larger c, the narrower the moderate-slope interval around $d_{i,j} = r_i + r_j$; the larger b, the steeper the slopes in the two transition regions. The two thin lines mark the hyperbolas bounding the values of ψ_{ij}.

As said, the role of λ in (1) is to balance the two terms so to obtain solutions that are consistent with both the data and our prior knowledge. The classical strategies to choose its value (*e.g.* [50]) do not fit our type of problem, since no contact frequency can be computed from a single solution sample. This is why we set λ to obtain, statistically, a fixed balance between Φ and $\lambda\Psi$ throughout all the blocks and all the resolution levels. A dedicated random sampling is used to obtain this result: for each block to be estimated, we produce a fixed number of random configurations, computing the corresponding values of both Φ and Ψ. Then, λ is chosen to set $\lambda\Psi/\Phi$, in average, to the desired value. The appropriate value is decided easily and once for all, by analyzing the features of the solutions obtained in a set of preliminary runs.

2.3 Sampling strategy

The solution space generated by Ξ is sampled through an approximate simulated annealing [46]: a warm-up phase to determine the start temperature for each block treated is followed by a slow cooling to reach comparable final scores in different runs. At each step, a quaternion operator is used to generate a random configuration. This is then accepted or rejected probabilistically on the basis of the differential score between the current and the proposed configurations divided by the current temperature. When the final scores for a pool of consecutive accepted configurations fall within a given tolerance, the iteration is stopped. This procedure is included in the overall recursion described by this simplified pseudocode:

ChromStruct(`matrix`, `chain`, `scale`):

1) extract diagonal blocks from `matrix`
2) *initialize lower-resolution chain*
   `lo-res_chain` ← null
3) For all the blocks
   a) *populate set* $\mathcal{L}$;
   b) *set initial guess*: $\mathcal{C}_0$=`chain(block)`;
   c) *sample the penalty landscape*:
      $\mathcal{C}_{\texttt{block}}$=`annealing(block,`$\mathcal{L}$`,`$\mathcal{C}_0$`)`;
   d) *save* $\mathcal{C}_{\texttt{block}}$
   e) *compute the equivalent low-resolution bead*:
      $\texttt{Bead}_{\texttt{block}} \leftarrow$ *beadify*($\mathcal{C}_{\texttt{block}}$)
   f) *Append* $\texttt{Bead}_{\texttt{block}}$ *to* `lo-res_chain`
4) *if # of extracted blocks* $= 1$
   $\mathcal{C} \leftarrow$ *recursive composition of all saved* $\mathcal{C}_{\texttt{block}}$
   save $\mathcal{C}$
   leave
   else
   update `scale`
   bin `matrix` *to the new resolution*
   ChromStruct(`matrix`, `lo-res_chain`, `scale`)

Since, at each scale, each subchain is estimated independently of the others, steps 3(a)-(e) can be performed in

parallel for all the blocks. This is a computational advantage, along with the fact that sampling low-dimensional spaces is much less expensive than sampling the entire panorama at once. Indeed, simulated annealing has a theoretical complexity $\mathcal{O}\{\exp(n)\}$, where n is the number of unknowns, so the finer the partition of the chain the more is saved in computing time. This strategy produces *one* structure per run. Different runs produce different structures. We could also proceed by using all the stable subchain configurations up to some scale to produce the structures at the subsequent scales [45], thus further reducing the computing effort.

3 Results and Discussion

In this section, we report some experimental results to validate our method and to compare it with our first method sharing the multiscale model and with the structure estimation method in the popular library TADbit[1] [48].

A significant comparison should be against the ground truth, which for these structures is only partially available. All the methods proposed in the literature so far have been validated and compared against known biological properties that are independent of the variability in the 3D chromatin shape. In this paper, the comparisons are only made with respect to the fitness to data, through the synthetic contact frequency matrices built on more or less extensive solution sets. A comparison between two different methods in terms of efficiency is fully significant in presence of equal outputs. However, the various existing methods do not provide the same results on the same data. On the other hand, the output structures depend on "free" parameters, and this further complicates a comparison. This is why we present some general considerations on efficiency, potential parallel implementation and computational complexity, with no reference to actual computing times.

The results on a chromosome segment, Section 3.1, are presented in detail, as these were the basis for our validation and study of the typical output structures. Sections 3.2-3.4, conversely, only contain a summary of our results with whole chromosomes and comparisons with the other methods. Some details are reported in the Appendices, here enclosed as a supplemental file.

3.1 Experiments with a chromosome segment

To explore the features of our algorithm and its results, our procedure was first tested against Hi-C data from human lymphoblastoid cells GM06990, chromosome 1, q range [150.28 Mbp, 179.44 Mbp] [7], at a genomic resolution of 100 kbp. For these data, our algorithm identifies two scale levels: the first includes 292 beads, each spanning 100 kbp, and the second includes 23 beads, with genomic sizes (in Mbp, the minimum being fixed at 0.7) of 1.1, 2.2, 0.7, 2.1, 0.7, 1.8, 1.0, 0.7, 0.7, 1.1, 2., 1.6, 1.3, 0.7, 0.7, 1.8, 1.8, 1.7, 1.3, 0.8, 1.3, 0.9, and 1.2. The contact frequency matrices at both scales are visualized as heat maps in Fig. 3.

We performed a preliminary series of experiments aimed at establishing the most appropriate constraints to obtain physically plausible results. All the subsequent experiments were performed with this set of parameters. In a subsequent phase, we produced 200 different solutions, from which we first checked the reliability of the method. Next, we evaluated the relevant geometrical features of the estimated structures and looked for a possible validation from known biological properties. In the following subsections, we report and comment these results.

1. https://3dgenomes.github.io/TADbit/, last accessed 2017, November 7^th

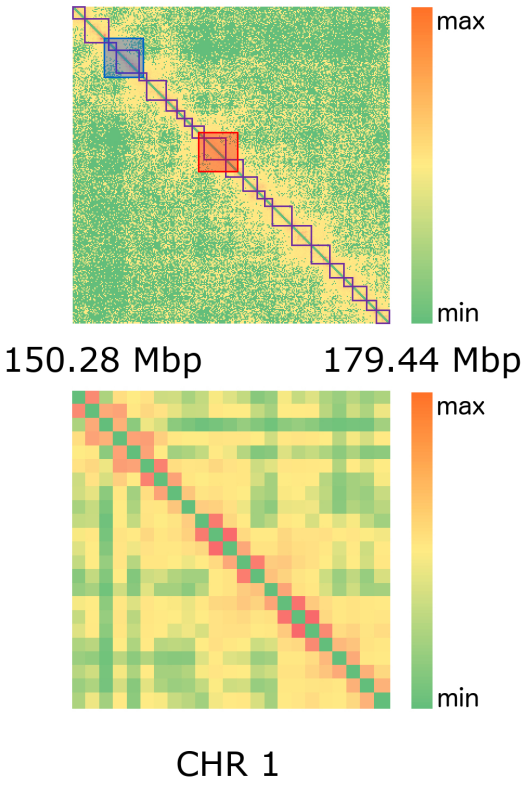

Fig. 3. **Contact frequency data.** Heat maps of the contact matrix for the long arm of chromosome 1, from a Hi-C experiment on human lymphoblastoid cells (GM06990, $150.28 \leq q \leq 179.44$ Mbp, main diagonal removed [7]). Left: Original 292×292 matrix at a resolution of 100 kbp. The 23 highlighted diagonal blocks mark the isolated domains detected. The two bigger blocks highlighted in blue (the leftmost one) and red, respectively, are related to a high-expression and a low-expression regions identified in the sequence at hand (see Section 3.1.2). Right: 23×23 matrix obtained by binning the original in accordance with the extracted blocks.

3.1.1 *Repeatability - Fit to the data*

The way we chose to check the repeatability of our solutions is to ensure that the associated scores lie in a limited range around some average value. This range obviously depends on the number of annealing cycles performed or, equivalently, on the tolerance of the stop criterion. To obtain a sufficiently rich set of configurations (see also [36]), we used two different tolerances: 10^{-5} and 10^{-2}. For the whole chain, we got Gaussian-distributed final scores, with standard deviations of, respectively, 11% and 30% of their mean values. The compatibility with the Gaussian distribution was checked by Shapiro-Francia tests [51]. The distributions we obtained for the individual subchains were much more peaked.

To produce a synthetic contact matrix for the entire population, we co-added all the single-configuration binary

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2791439, IEEE/ACM Transactions on Computational Biology and Bioinformatics

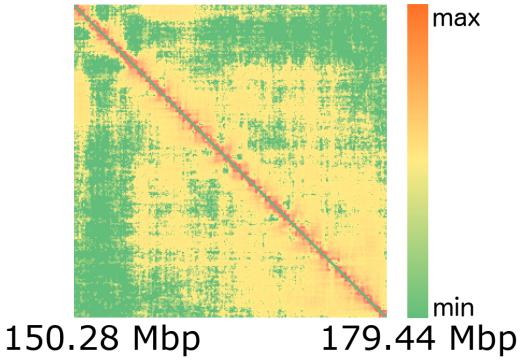

Fig. 4. **Synthetic contact matrix.** Heat map of the synthetic contact matrix built from 200 final configurations, whole chain at 100 kbp resolution, to be compared to Fig. 3, top panel.

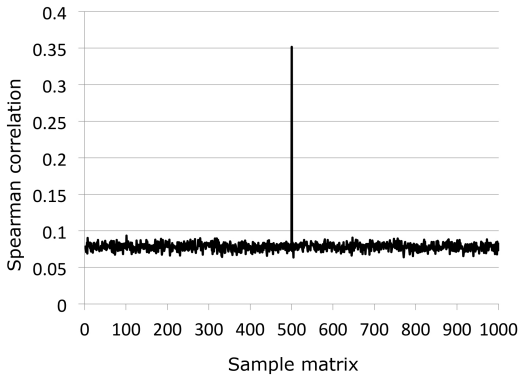

Fig. 5. **Fit to the data.** Spearman correlation between the original contact matrix, the synthetic contact matrix in Fig. 4 (at position 500 in the horizontal axis), and 1,000 random contact matrices.

matrices, obtained by considering in contact all the bead pairs whose mutual distance does not exceed 1.2 times the sum of their estimated radii. The result from our 200 outputs is shown in Fig. 4. The threshold distance assumed corresponds to the largest distance penalized by Ψ for the case $c = 4, b = 5$ (see Fig. 2). To check the data fit of our solutions, we compared the result to the input matrix in Fig. 3. It can be noted that most of the relevant structures in the original data are captured by our solutions. As a quantitative index of similarity, we used the Spearman correlation between the original and the reconstructed contact matrices, after suppressing their main diagonals and first subdiagonals. The Spearman correlation was then compared to the correlations between the original matrix and 1,000 random matrices with the same macroscopic features: the test matrices were symmetric by construction, and the values in each subdiagonal were uniformly distributed in the same range as in the original matrix. The results are summarized in the plot of Fig. 5. The seemingly modest value of about 0.35 results from the suppression of the dominant diagonals, and is one order of magnitude larger than the corresponding correlations obtained from the random matrices.

3.1.2 *Geometrical features - Biological plausibility*

All our results were compared using the mean-squared (M-S) Euclidean distance between pairs of beads as a function of their genomic distance [45], [52], indicative of the packing of each solution (see also [27]). The average plot of this function evaluated from all our results is reported in Fig. 6. Considering that a fully stretched chain would produce a quadratic behavior, this shows that most output configurations are tightly packed. To cluster the plots resulting from all the output configurations, we relied on hierarchical clustering on principal components (HCPC [53]). This is a hybrid technique combining principal components, hierarchical and partitional clustering, capable of detecting the number of relevant clusters from intra- and inter-cluster distance optimization. A summary of the results is reported in Fig. 7.

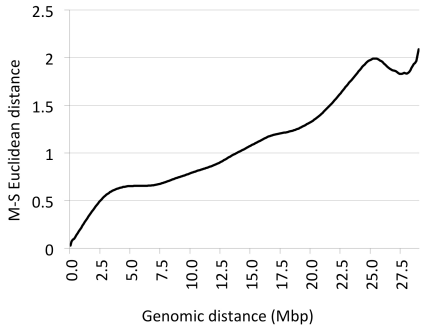

Fig. 6. **M-S Euclidean (μm^2) vs. Genomic distance plot.** For each possible genomic distance in our chain, the mean-square Euclidean distances are first computed for each output configuration, and then averaged over all our 200 configurations.

For biological plausibility, we first observe that the physical sizes of our solutions are compatible with the constraints posed by the nucleus size. The size of each estimated structure depends on both the sizes of the elementary beads and their packing. Since our solutions come directly with their physical dimensions and are not scaled *a posteriori*, this is a first index of reliability.

A second criterion derives from another property of DNA [54]: highly expressed or gene-rich domains are packed more loosely than the domains poor in genes or with low transcriptional activity. Our data belong to the lineage of immature B cells. From the related expression data,[2] we identified two stretches of about 3.5 Mbp as representative of, respectively, highly expressed and poorly expressed genomic domains: DNA from $q = 151.5$ Mbp to $q = 155.1$ Mbp, and DNA from $q = 162$ Mbp to $q = 165.5$ Mbp. These regions, highlighted by different colors in Fig. 7, are both modeled by 35-bead subchains, and are rich and poor in genes, respectively.[3] To check the packing of the corresponding segments in our output chains, we use again the M-S Euclidean distances as functions of the genomic distances between pairs of beads. In Fig. 8, we show the two boxplots for the highly and poorly expressed regions. A different behavior is apparent: the poorly expressed region occupies less space than the highly expressed one, although their genomic spans are nearly the same. We checked the statistical significance of these features: by Kolmogorov-Smirnov tests, we can reject

2. http://bioinfo.amc.uva.nl/HTMseq/controller, last accessed: 2016, June 15^{th}

3. http://www.ensembl.org/ Release 77, last accessed: 2017, November 13^{th}

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2791439, IEEE/ACM Transactions on Computational Biology and Bioinformatics

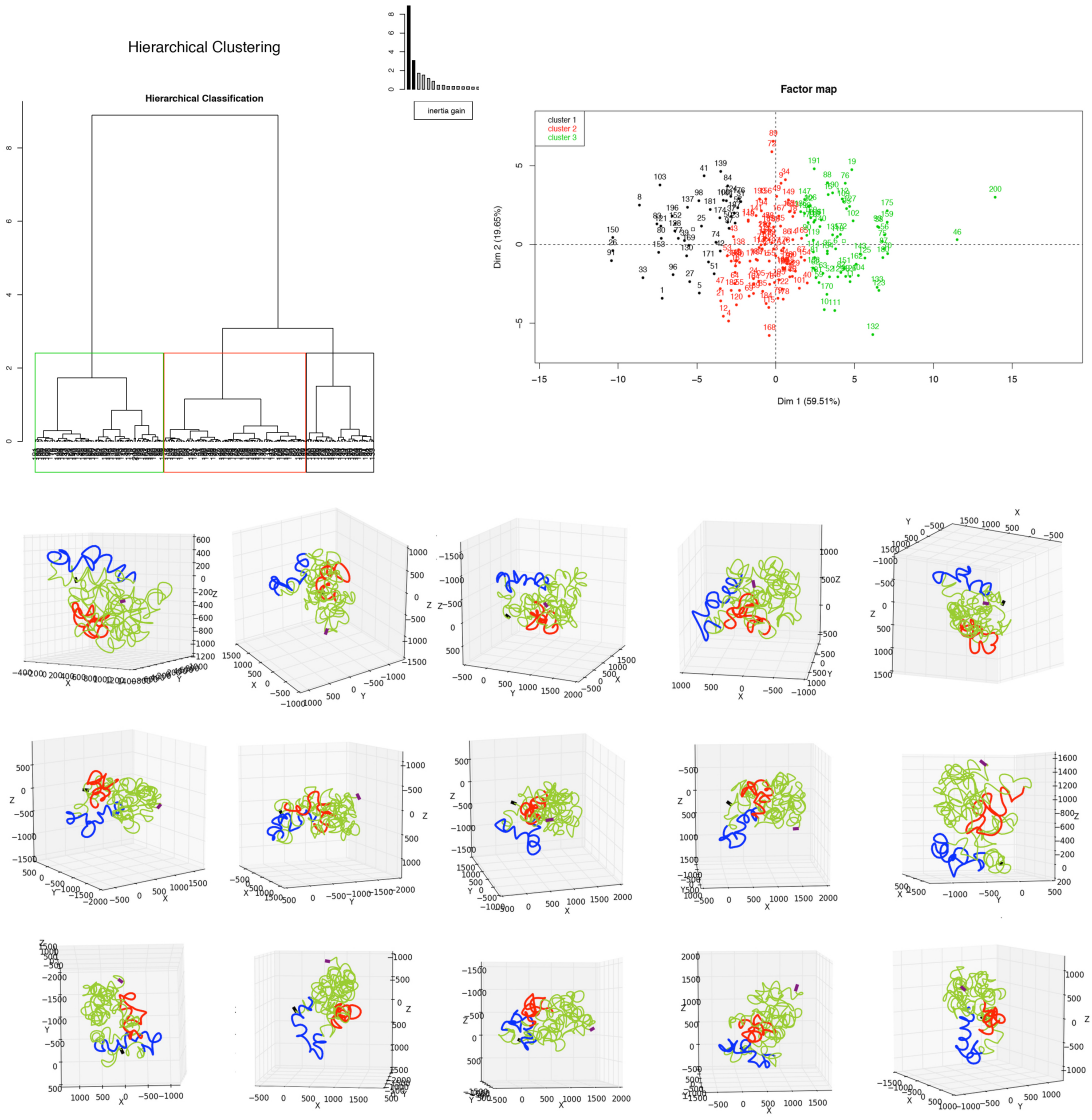

Fig. 7. **Summary of clustering results.** Top: the classification tree produced by HCPC, and the three clusters detected, projected onto the first two principal components. The subsequent rows show the five configurations closest to the centroid of each cluster. The black marks locate bead 1, and the purple marks locate bead 292. Measurements in nm. The subchains highlighted in blue and red are, respectively, the highly and poorly expressed regions identified to check biological plausibility.

the hypothesis that the values from the two regions, for each genomic distance, are drawn from the same distribution. Interestingly, for genomic distances larger than 1.1 Mbp, the distributions of the M-S Euclidean distances are always leptokurtic. Moreover, beyond a genomic distance of 2 Mbp, the kurtoses from the highly expressed region are always larger than the ones related to the poorly-expressed region. This means that a) statistically, the geometrical behavior of the two stretches is similar for limited genomic distances but, b) in its entirety, the poorly expressed DNA stretch is more folded than the highly expressed one. These features can also be seen by visual inspection of our solutions, including those shown in Fig. 7. We also observe that, besides being folded less tightly than the other, the highly expressed region very often interferes less with the rest of the chain, and is located at the periphery of the structure.

3.2 Whole chromosomes

To demonstrate the capabilities of our algorithm to treat whole chromosomes, we need to treat input matrices presenting large bands with missing data. Indeed, the centromere and telomere regions of all the chromosomes are characterized by many repeating base sequences, whose genomic locations cannot be assigned univocally. To face this difficulty, we assign a fixed shape and size to telomeres and centromeres, and include them as single beads in our chain model. This problem could also be solved by lowering the genomic resolution, thus including the missing and vanishing data in some nonzero entry of a binned contact matrix.

In Appendices A.1 and A.2, we show some results for the whole chromosomes 1 and 16, respectively, at different initial scales. Chromosome 1 has been reconstructed starting from 1Mbp and 100kbp scales (Figs. A1 and A2, respec-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2791439, IEEE/ACM Transactions on Computational Biology and Bioinformatics

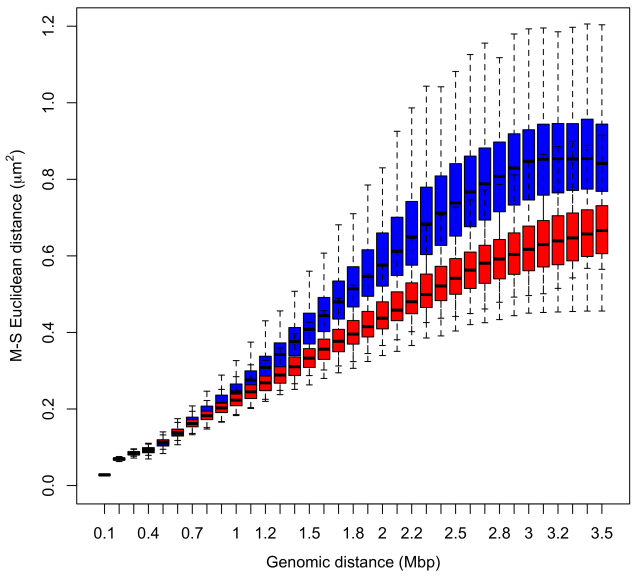

Fig. 8. **Boxplots of M-S Euclidean vs. Genomic distance.** Boxplots obtained from all our 200 solutions for the identified highly-expressed (blue) and poorly expressed (red) regions.

tively), and chromosome 16 starting from eight different scales, from 1Mbp to 5kbp (Figs. A3-A10). Of course, the figures at different initial resolutions do not need to be coherent to each other, since they are different samples from the solution space. We can observe, however, that the overall sizes of the different solutions are all coherent with the expected physical sizes, and this means that our strategy to assign the bead sizes brings to plausible results.

3.3 Comparison with *ChromStruct 1*

Appendix A.3 reports a comparison between the results obtained by CHROMSTRUCT 3.2 and the corresponding results of the method we proposed in [45], here called *ChromStruct 1*, on the data visualized in Fig. 3 (Fig. A11 in Appendix A.3). We already mentioned the drawbacks of the method implemented by *ChromStruct 1*, and the motivations that brought us to the new method and algorithm. As far as the results are concerned, Appendix A.3.1 summarizes the results in [45] by the synthetic contact frequency matrix, built as explained in Section 3.1 (see Fig. A13). In Appendix A.3.2, Fig. A15, we report the matrix obtained from the results of CHROMSTRUCT 3.2 on the same data set. The latter result is apparently more similar to the original than the one in Fig. A13, and this is also confirmed by a slightly larger Spearman coefficient: 0.39 versus 0.35. This could be explained by either the smaller solution set drawn from [45] or by the difficulty of sampling properly the solution space when forcing rigid constraints on the estimated structures.

3.4 Comparison with TADbit

In Appendix A.4, we report a comparison between our results and corresponding results produced by TADbit (by using its default parameters). In this case, the data matrix comes from the same segment of human chromosome 1 considered in Section 3.1 but, rather than being applied directly to the raw data, the algorithms are fed by the matrix normalized by the ICE method [56], shown in Fig. A18. Visually, the two synthetic matrices (Figs. A19 and A20) seem to be quite different from the original, but their patterns are similar. The matrix reconstructed by TADbit seems to saturate its entries towards the maximum values. The Spearman correlations are 0.44 for CHROMSTRUCT and 0.33 for TADbit. Of course, this is not indicative of actual performances, at least for TADbit, which has been run with no preliminary parameter tuning.

4 CONCLUSION

We propose a multiresolution chromatin structure estimation method based on a score function made of a balanced mix of data-fit and soft geometrical requirements. The recursive multiresolution setting enables us to exploit the presence of nearly isolated genomic domains typically recurring at all scales. The geometrical constraints introduced implicitly in the score function allow our prior knowledge to be exploited coherently and flexibly through a small set of tunable parameters. The solution space is sampled by simulated annealing, and the chromatin model is evolved through quaternion operators. The successive levels of genomic resolution, as well as the annealing parameters, are determined automatically.

Our results reproduce the main features of the original contact frequency matrices, avoiding an excessive bending of the chromatin and interpenetrations between beads. Besides being consistent with the data and physically plausible, our solutions show features that also suggest biological plausibility. In particular, the results demonstrate that two regions identified as highly and poorly expressed show the expected structural properties. As this feature was not specifically introduced in the model and no geometrical requirement introduced is site-specific, we consider this result significant for a biological validation.

In summary, this approach promises to be more efficient than the ones that do not consider the recursive behavior of the chromatin fibre through the different scales. It also offers consistency and controllability features that make it a good candidate to generate and analyze large populations of structures fitting the same data. As a method, it is suitable to estimate the structure of any set of adjacent genomic fragments as soon as enough contact data are available.

SUPPLEMENTAL MATERIAL

Some additional results are included in the appendices. Three Python codes and a Readme file are also provided:

1) File `APPENDICES.pdf`: Appendices A.1 to A.4;
2) File `ChromStruct_3.2_GUI.py`: the GUI version of CHROMSTRUCT 3.2;
3) File `Plot_Energy_Chain_2.0.py`: a separate command-line code to display the results;
4) File `README.pdf`: short usage notes.

The data used in Section 3.1 can be downloaded from `http://dx.doi.org/10.13140/RG.2.2.35785.13923`.

ACKNOWLEDGMENTS

This work was partially supported by the Italian Ministry of Education and by the National Research Council, Flagship Project InterOmics, PB.P05-WP1-ISTI. The authors thank Luciano Milanesi and Clelia Peano for useful discussions, and Ivan Merelli for his help with the TADbit results.

REFERENCES

[1] Cao, J *et al*. Three-dimensional regulation of transcription. Protein Cell. 2015; 6:241–253.
[2] Langer-Safer, P R *et al*. Immunological method for mapping genes on Drosophila polytene chromosomes. Proc. Natl. Acad. Sci. U.S.A. 1982; 79:4381–4385.
[3] Amann, R and Fuchs, B M. Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. Nat. Rev. Microb. 2008; 6:339–348.
[4] Dekker J *et al*. Capturing chromosome conformation. Science. 2002; 295:1306–1311.
[5] Zhao Z *et al*. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet. 2006. 38:1341–1347.
[6] Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat. Protoc. 2007; 2:988–1002.
[7] Lieberman-Aiden E *et al*. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009; 326:289–293.
[8] van Berkum NL *et al*. Hi-C: a method to study the three-dimensional architecture of genomes. J. Vis. Exp. 2011; 39:1869–1875.
[9] Nagano T *et al*. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013; 502:59–64.
[10] Nagano T *et al*. Comparison of Hi-C results using in-solution versus in-nucleus ligation. Genome Biol. 2015; 16:175.
[11] Nagano T *et al*. Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. Nat. Protoc. 2015; 10:1986–2003.
[12] Farlik M *et al*. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep. 2015; 10:1386–1397.
[13] Paulsen J *et al*. Manifold based optimization for single-cell 3D genome reconstruction. PLoS Comput. Biol. 2015; 11:e1004396.
[14] Le Dily F *et al*. 3D modeling of chromatin structure: is there a way to integrate and reconcile single cell and population experimental data?. WIREs Comput Mol Sci 2017, e1308.
[15] Cusanovich DA *et al*. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science, 2015; 348:910–914.
[16] Buenrostro JD *et al*. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015; 523:486–490.
[17] Rotem A *et al*. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat. Biotech. 2015; 33:1165–1172.
[18] Segal MR, Bengtsson HL. Reconstruction of 3D genome architecture via a two-stage algorithm. BMC Bioinf. 2015; 16:373.
[19] Fraser J *et al*. Chromatin conformation signatures of cellular differentiation. Gen. Biol. 2009; 10:R37.
[20] Duan Z *et al*. A three-dimensional model of the yeast genome. Nature. 2010; 465:363–367.
[21] Varoquaux N *et al*. A statistical approach for inferring the 3D structure of the genome. Bioinformatics. 2014; 30: i26–i33.
[22] Kahlor R *et al*. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotech. 2012; 30: 90–100.
[23] Gehlen LR *et al*. Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions. Nucleus. 2012; 3: 370–383.
[24] Meluzzi D, and Arya G. Recovering ensembles of chromatin conformations from contact probabilities. Nucl. Ac. Res. 2013; 41: 63–75.
[25] Nowotny J *et al*. Iterative reconstruction of three-dimensional models of human chromosomes from chromosomal contact data. BMC Bioinf. 2015; 16:338.
[26] Tokuda N *et al*. Dynamical modeling of three-dimensional genome organization in interphase budding yeast. Biophys. J. 2012; 102:296–304.
[27] Wang S *et al*. Inferential modeling of 3D chromatin structure. Nucl. Ac. Res. 2015; 43:54.
[28] Di Stefano M *et al*. Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. Scientific Reports 2016; 6:35985.
[29] Zhang ZZ *et al*. Inference of Spatial Organizations of Chromosomes Using Semi-definite Embedding Approach and Hi-C Data. in Deng M *et al*, Research in Computational Molecular Biology, Berlin, Springer-Verlag, 2013, pp. 317–332.
[30] Rousseau M *et al*. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC Bioinf. 2011; 12:414–429.
[31] Hu M *et al*. Bayesian inference of Spatial organizations of chromosomes. PLOS Comp. Biol. 2013; 9:e1002893.
[32] Zou C *et al*. HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. Gen. Biol. 2016; 17:40–53.
[33] Trussart M *et al*. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. Nucl. Ac. Res. 2015; 43:3465–3477.
[34] Lajoie BR *et al*. The hitchhiker's guide ti Hi-C analysis: Practical guidelines. Methods. 2015; 72:65–75.
[35] Junier I *et al*. On the demultiplexing of chromosome capture conformation data. FEBS Lett. 2015; 589:3005-3013.
[36] Imakaev M *et al*. Modeling chromosomes: Beyond pretty pictures. FEBS Lett. 2015; 589:3031-3036.
[37] Shavit Y *et al*. How computer science can help in understanding the 3D genome architecture. Brief. Bioinf. 2016; 17:733–744.
[38] Sekelja M *et al*. 4D nucleomes in single cells: what can computational modeling reveal about spatial chromatin conformation? Genome Biol. 2016; 17:54–63.
[39] Schmitt AD *et al*. Genome-wide mapping and analysis of chromosome architecture Nature Rev. Mol. Cell Biol. 2016; 17:743–755.
[40] Forcato M *et al*. Comparison of computational methods for Hi-C data analysis Nature Meth. 2017; 14:679–689.
[41] Yardimci GG and Noble WS, Software tools for visualizing Hi-C data Genome Biol. 2017; 18:26–34.
[42] Caudai C *et al*. A statistical approach to infer 3D chomatin structure. in Zazzu, V *et al*. Mathematical Models in Biology, Springer International Publishing Switzerland. 2015; 161–171.
[43] Duggal G *et al*. Resolving spatial inconsistencies in chromosome conformation measurements. Algorithms for Molecular Biology. 2013; 8:8.
[44] Dixon JR *et al*. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485:376–380.
[45] Caudai C *et al*. Inferring 3D chromatin structure using a multiscale approach based on quaternions. BMC Bioinformatics. 2015; 16:234–244.
[46] Kirkpatrick S *et al*. Optimization by Simulated Annealing. Science. 1983; 229:671–680.
[47] Vince JA. Geometric Algebra for Computer Graphics. Springer, Berlin. 2008.
[48] Baù D and Marti-Renom MA. Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. Methods 2012; 58:300–306.
[49] Tikhonov AN, and Arsenin VY. Solution of ill-posed problems Winston-Wiley, Washington. 1977.
[50] Wahba G. Practical approximate solutions to linear operator equations when the data are noisy. SIAM J. Numer. Anal. 1977; 14:651–667.
[51] Thode HC *et al*. Testing for Normality. Marcel Dekker; New York. 2002.
[52] Mateos-Langerak J *et al*. Spatially confined folding of chromatin in the interphase nucleus. PNAS. 2009; 106:3812–3817.
[53] Husson F *et al*. Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualising data? Applied Mathematics Department, Agrocampus. Rennes, France. 2010.
[54] Versteeg R *et al*. The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes. Genome Res. 2003; 13:1998–2004.

[55] Rao SSP *et al*. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 2014; 159:1665–1680.
[56] Imakaev M *et al*. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Meth. 2012; 9:999–1003.

Claudia Caudai received the MS degree in Mathematics in 2003 and the PhD in Biomedical Engineering in 2009, both from Pisa University. She has been working on modeling of ECG signals and neural and neuroglial messages in human brain, and on the evaluation of compliance and stiffness of biological and artificial muscles. From 2008 to 2011, she was with the scientific visualization unit at the CNR Institute of Clinical Physiology in Pisa, Italy, working on 3D visualization of biological processes. Currently, she is a post-doctoral fellow at Signals and Images Laboratory, Institute of Information Science and Technologies of National Research Council of Italy. Her research interests include mathematical modeling, bioinformatics, biology, genomics and proteomics. She participates in the Italian Flagship Project InterOmics.

Emanuele Salerno, an electronic engineer, joined the National Research Council of Italy in 1987. Currently, he is a senior researcher at the Institute of Information Science and Technologies in Pisa, Italy. He has been working in microwave tomography, monitoring of combustion processes, computer vision for robot guidance, and astrophysical imaging. His present scientific interests are in inverse problems with applications in SAR image processing, IT for cultural heritage, and computational biology. He has been teaching instrumentation and measurements and microwave techniques at the university of Pisa, and is a member of IEEE, Signal Processing Society, of the Italian federation of electrical and information engineering (AEIT), and a fellow of the Electromagnetics Academy.

Monica Zoppè graduated in biology at the University of Milan, Italy. Since then, she has been working at the Department of Biochemistry, University of Birmingham, UK, at the CNR Institute of Advanced Biomedical Technologies in Milan, at the Molecular Biology and Virology lab of the Salk Institute, La Jolla, CA, and at the lab of Gene and Molecular Therapy of the International Centre of Genetic Engineering and Biotechnology in Trieste, Italy. Since 2001, she is at the CNR Institute of Clinical Physiology in Pisa, Italy, where she founded and directs the Scientific Visualization Unit. Before switching to molecular visualization, she held several fellowships and awards for research on cytoskeletal proteins, gene therapy, adenoviral vectors and tobacco-related diseases. Her teaching experience includes the supervision of master and PhD theses, and several classes and seminars in biomedicine, biotechnology and gene therapy, and molecular graphics. She is a member of the International Society for Computational Biology, the Molecular Graphics and Modelling Society, and DonneScienza, the Italian society of women scientists.

Anna Tonazzini is a senior researcher at the Institute of Information Science and Technologies, National Research Council of Italy, in Pisa. She coordinated several Projects in Image Processing and Analysis, Neural Networks and Learning, Computational Biology and Document Analysis, and is co-author of over 100 published papers. She was also the ISTI responsible for the UE Project ISYREADET and the national Flagship Project InterOmics. She chaired the EUSIPCO2008 Special Session on restoration of degraded document images, and edited the special Issue on image and video processing for cultural heritage of the *Eurasip Journal on Image and Video Processing*. She supervised many theses in computer science, mathematics and information engineering, two Ercim fellowships, and various post-doctoral positions. She is an editor of *Digital Signal Processing*, a member of the IASTED Technical Committee on Image Processing, and a program committee member in several international conferences.