A Statistical Approach to infer 3D Chromatin Structure

C.Caudai, E.Salerno, M.Zoppè and A.Tonazzini

Abstract We propose a new algorithm to estimate the 3D configuration of a chromatin chain from the contact frequency data provided by HI-C experiments. Since the data originate from a population of cells, we rather aim at obtaining a set of structures that are compatible with both the data and our prior knowledge. Our method overcomes some drawbacks presented by other state-of-the-art methods, including the problems related to the translation of contact frequencies into Euclidean distances. Indeed, such a translation always produces a geometrically inconsistent distance set. Our multiscale chromatin model and our probabilistic solution approach allow us to partition the problem, thus speeding up the solution, to include suitable constraints, and to get multiple feasible structures. Moreover, the density function we use to sample the solution space does not require any translation from contact frequencies into distances.

1 Introduction

The nuclear DNAis arranged in a 30 nm fiber called chromatin, and in human cells has a length of about 2 m in total, folded in 46 chromosomes. Its spatial organization ensures the continuous accessibility of DNA to translation, replication regulation and repair machinery. Understanding how DNA is organized will help to discover its functional features and the epigenetic mechanisms involved. A first important step in describing the organization of DNA within the nucleus was done with the experiments of fluorescence in situ hybridization (FISH) [1], a technique used to detect and

C.Caudai, E.Salerno, A.Tonazzini
National Research Council of Italy, Institute of Information Science and Technologies,Via Moruzzi 1, 56124, Pisa, Italy, e-mail: claudia.caudai@isti.cnr.it

M.Zoppè
National Research Council of Italy, Institute of Clinical Physiology,Via Moruzzi 1, 56124, Pisa, Italy

localize specific DNA sequences. Recently, high resolution techniques have been developed, called Chromosome Conformation Capture (3C) [5], which provide contact frequencies between pairs of DNA fragments in the whole genome. The latest such technique, called HI-C [3], has a very high genomic resolution, reaching a few kbp, depending on the enzyme used in the procedure.

From HI-C information, it is possible to formulate hypotheses about the three-dimensional chromatin configurations. Many approaches have been proposed to address this problem. They can be divided into three main categories, each offering specific advantages and criticalities: constrained optimization, Bayesian inference, and polymer models. The new reconstruction method we propose in this chapter was conceived to exploit the benefits of the state-of-the-art methods while avoiding some of their drawbacks.

All the constrained optimization strategies proposed introduce a model for the solution, a set of constraints, and a cost function to be optimized against the available data. As mentioned, the 3C data available are contact frequencies evaluated over the whole population of cells in the experiment, typically many millions. The first attempts to translate these data into geometrical information assume that the chromatin configurations are not very different throughout the population, and that pairs of fragments often found in contact are closer than pairs with low contact frequencies. On this basis, most of the existing methods propose some formula to translate the contact frequencies into Euclidean distances, to be fitted by the reconstructed structures. Duan *et al.* [8] propose a three-dimensional model of yeast genome, in which chromatin is modeled as a bead chain, with partially impenetrable beads, forced to stay in a spherical nucleus of 1 μm. The objective function to be minimized exploits an inverse proportionality relationship between contacts and distances. The same deterministic law is also adopted by Fraser *et al.* [9] and Dekker *et al.* [5]. In Sect. 2, we show how this translation leads to severe geometric inconsistences. Baù *et al.* [2] translate the contact frequencies into harmonic forces, calibrating the distances between beads. The constrained optimization approach has the advantage of introducing geometric and biophysical constraints into the model, but has two big disadvantages: the high dimensionality of the systems and the absence of confidence intervals to evaluate the uncertainty of the solutions obtained.

The data are affected by errors and biases and, as mentioned, derive from experiments on millions of cells. This makes necessary the adoption of a probabilistic approach to sample the space of the feasible solutions. The first probabilistic approach has been published by Rousseau *et al.* [16], who use a Markov Chain Monte Carlo sampling on a Gaussian likelihood, built through an inverse-quadratic law between contacts and distances (MCMC5C). Hu *et al.* [10] use the same relationship, proposing an algorithm called BACH (Bayesian 3D Constructor for HI-C data), to build consensus 3D structures. The novelty of the cited Bayesian approaches is the possibility to introduce biases into the data model (as in BACH). Another important advantage is the possibility of sampling the solution space: this aspect is essential, since it is more meaningful to search for sets of possible solutions rather than a single consensus. The major drawbacks of BACH are its computational complexity,

due to the large number of parameters to be estimated, and the absence of suitable topological constraints.

Another interesting approach is the integration of polymer physics into the 3D chromatin structure model. This has the advantage of not requiring the translation from frequencies into distances, and permits the adoption of iterative adaptive methods. Meluzzi *et al.* [14] propose a coarse-grained bead-chain polymer model approximating the physical behavior of a 30 nm chromatin fiber; the system evolves adjusting iteratively the model parameters, until a match with contact frequency data is reached. This approach is highly reliable but very expensive computationally. For this reason, it cannot yet be applied to experimental data: a validation has only been performed against reference data sets obtained from simulations of systems with up to 45 beads.

An analysis of the different solutions mentioned above reveals a number of drawbacks that must be overcome to obtain more reliable results. Our main point is the questionable adequacy of the translation of contact frequencies into Euclidean distances. In Sect. 2, we show that this strategy produces a set of distances often severely incompatible with the Euclidean geometry. Then, in Sect. 3, we briefly describe our solution model, our cost function, which does not include an explicit contact-to-distance relationship, and the stochastic algorithm we used to sample the solution space. Sect. 4 concludes the chapter, with some reference to our first experimental results.

2 Geometrical Consistency of the Frequency-Distance Translation

The problem of the geometrical inconsistencies derived from translating contact frequencies into Euclidean distances has been overlooked by almost all groups that have worked with contact frequency data. An exception is the work of Duggal *et al.* [7], who propose a filtering technique to select subsets of interactions obeying to metric constraints. This method is very interesting, but has a high computational cost.

It is important to exert some caution with the extraction of topological information (measurements, distances) from interaction data, because contacts are discrete events (sums of dichotomous events) with causal and random components, whereas spatial distances are continuous quantities forced to undergo precise geometric laws. It is necessary to check whether the distances meet the basic geometrical consistency conditions, *e.g.* the triangular inequality. The non-violation of these conditions is a necessary but not sufficient condition for geometric consistency. If geometric consistency conditions are severely violated, the set of distances cannot be used as a target to achieve sensible geometric conformations of chromatin. However, the fact that these inequalities are not violated, or are violated slightly, does not ensure the geometrical consistency of the system. For example, if we have a set of equal distances (e.g. all equal to 1), the triangular inequalities would never be violated, but no

structure in the 3D Euclidean space can show such a distance set, unless it is made of no more than 4 points.

Let us consider a chromatin chain made of N elements, and any subsequence S of it, with M elements, identified by the index set $I = \{1,2,...,M\}$. Let us now consider a partition P of S, that is, any set of $L \leq M$ consecutive segments that sum up to S, identified by the set of index pairs $K = \{(1,k_2),(k_2,k_3),...,(k_L,M)\}$, with $1 < k_2 < k_3 < ... < k_L < M$. A necessary condition for the Euclidean distances between all the possible pairs in S to be consistent with the 3D Euclidean geometry is that, for any possible K:

$$d_{1,M} \leq \sum_{(i,j)\in K} d_{i,j} \tag{1}$$

where $d_{i,j}$ is the distance between the i-th and the j-th elements of S.

In our preliminary study, we considered two sets of experimental data made available in the literature, from the entire human genome in GM06690 [13] and GSE18199 cells [17], both with genomic resolution of 1 Mbp. Then, for both data sets, for any possible subsequence of all the chromosomes, and for 13 different frequency-to-distance relationships, we evaluated the number and the extent of the violations to condition (1). The results of this analysis are summarized in Table 1, whereas the contributions of each individual chromosome are plotted in Figs. 1 and 2. The number of violations and their weights rapidly decrease by applying the laws $1/\sqrt[n]{x}$, with $n \in \{1,2,..,5\}$. This does not mean that these laws are suitable to build a good target function, since they actually tend to produce a set of nearly equal distances, which normally lead to impossible structures.

Also considered from another viewpoint, the inversion process from contact frequencies into distances presents a heuristic gap, because the measured contact frequencies do not depend exclusively on geometric properties, but also on other factors, such as the presence of topological barriers, energy conditions, and random events. In summary, we think that assuming that pairs with many contacts are likely to be close to each other can be justified, whereas pairs with a few contacts are not warranted to be distant from each other. Our analysis demonstrates that experimental frequency data very often lead to distances that are more or less severely incompatible with real configurations in the 3D Euclidean space. For this reason, such distances cannot be used as rigid targets for structure estimation. Actually, any fixed distance system identifies a well defined structure in space, but our data do not come from a single structure, so a distance system obtained through whatever relationship is very likely to be geometrically inconsistent.

3 Our Approach

Each of the studies that proposed methods for 3D chromatin reconstruction from contact data presents problems and advantages, summarized in Table 2. As a contribution to the field, we propose a new algorithm that includes a list of desirable features:

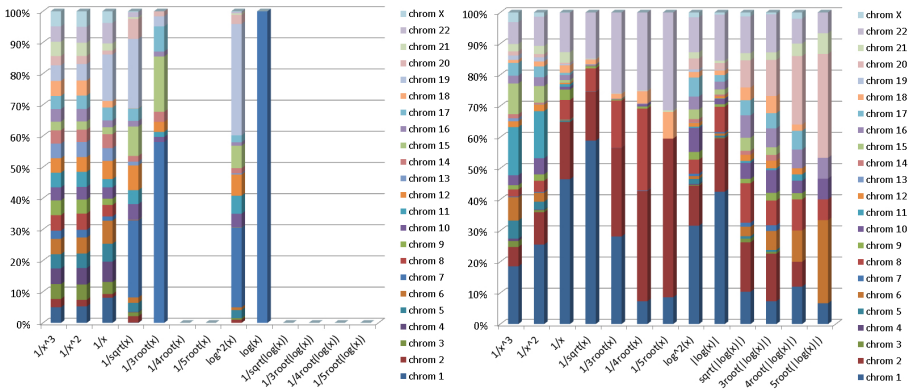

Fig. 1 Percent contributions of the different chromosomes to the total number of geometric violations, for the 13 transformation laws considered. Left: data from [13]. Right: data from [17]. For each column, the contributions of the chromosomes have always the same order: from chromosome 1 at the bottom to chromosomes 22 and X at the top of the column.

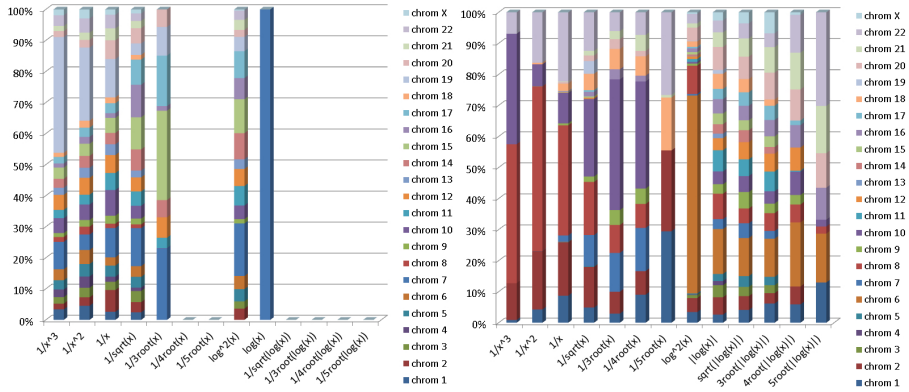

Fig. 2 Percent contributions of the different chromosomes to the average extent of the geometric violations, for the 13 transformation laws considered. Left: data from [13]. Right: data from [17]. For each column, the contributions of the chromosomes have always the same order: from chromosome 1 at the bottom to chromosomes 22 and X at the top of the column.

i) Possibility to enforce geometrical constraints on the solutions.
ii) Computational efficiency, including partitioning and parallel processing capabilities.
iii) No deterministic translation from contact frequencies to distances.
iv) Possibility to get multiple configurations compatible with the data.

To obtain features i) and ii), we rely on our chromatin model. If we model the chromatin fiber as a bead chain, we can first impose that it must remain connected, that is, that the beads must maintain their genomic locations, and then introduce con-

Table 1 Frequency-distance conversion laws for dataset available in [13, 17]. In the formulas x represents the contact frequency and d the Euclidean distance.

Lieberman-Aiden *et al.* [13]			Yaffe and Tanay [17]		
Transformation laws	Number of violations[a]	Average percentage violation	Transformation laws[b]	Number of violations[a]	Average percentage violation
$x \longrightarrow d = \frac{1}{x^3}$	28003.8	3458.5	$x \longrightarrow d = \frac{1}{x^3}$	2464.6	$4 \cdot 10^8$
$x \longrightarrow d = \frac{1}{x^2}$ [10, 16]	26502.8	424	$x \longrightarrow d = \frac{1}{x^2}$	1439.6	$6 \cdot 10^5$
$x \longrightarrow d = \frac{1}{x}$ [5, 8, 9]	8954,1	42.4	$x \longrightarrow d = \frac{1}{x}$	766.7	1501.8
$x \longrightarrow d = \frac{1}{\sqrt{x}}$	72,3	8.6	$x \longrightarrow d = \frac{1}{\sqrt{x}}$	604.9	99.4
$x \longrightarrow d = \frac{1}{\sqrt[3]{x}}$	2.7	1.5	$x \longrightarrow d = \frac{1}{\sqrt[3]{x}}$	287.9	32
$x \longrightarrow d = \frac{1}{\sqrt[4]{x}}$	0	0	$x \longrightarrow d = \frac{1}{\sqrt[4]{x}}$	55.9	16.3
$x \longrightarrow d = \frac{1}{\sqrt[5]{x}}$	0	0	$x \longrightarrow d = \frac{1}{\sqrt[5]{x}}$	9.1	4.6
$x \longrightarrow d = \frac{1}{\log^2(x)}$	65.4	8.7	$x \longrightarrow d = \log^2(x)$	1143	1095.5
$x \longrightarrow d = \frac{1}{\log(x)}$	0.3	0.3	$x \longrightarrow d = \lvert\log(x)\rvert$	566.7	72.8
$x \longrightarrow d = \frac{1}{\sqrt{\log(x)}}$	0	0	$x \longrightarrow d = \sqrt{\lvert\log(x)\rvert}$	34	28
$x \longrightarrow d = \frac{1}{\sqrt[3]{\log(x)}}$	0	0	$x \longrightarrow d = \sqrt[3]{\lvert\log(x)\rvert}$	7.1	18.5
$x \longrightarrow d = \frac{1}{\sqrt[4]{\log(x)}}$	0	0	$x \longrightarrow d = \sqrt[4]{\lvert\log(x)\rvert}$	2.2	8.5
$x \longrightarrow d = \frac{1}{\sqrt[5]{\log(x)}}$	0	0	$x \longrightarrow d = \sqrt[5]{\lvert\log(x)\rvert}$	0.7	5.2

[a] Averaged on chromosomes.
[b] Contact frequency values normalized to 1.

straints on the distances between adjacent beads and on the angles formed by any two consecutive bead pairs. This amounts to constrain the length of any subchain and its maximum curvature. Of course, the appropriate values for these constraints must be decided on the basis of the relevant biological knowledge. Partitioning the problem can enable us to speed up the estimation process. We reach this goal by taking into account the existence of chromatin segments, called *topological domains* [6], that have no important interactions with other genomic regions, and exploiting the multiscale capabilities of our chromatin model. The structure of each topological domain can be estimated from the data coming exclusively from the fragments belonging to it. The resulting structure is then considered as a bead in a lower resolution chain, whose contact frequencies are evaluated along with possible higher-level isolated domains. The structures of these new topological domains are reconstructed by the same strategy described above. This process can continue recursively, until a data set with a single domain is found. The full-resolution structure is then reconstructed by substituting, recursively, the lower-resolution beads with the subchains reconstructed at finer resolutions. Except for the finest resolution available, our beads are not spheres, but are equipped with the macroscopic properties of the subchains they represent, each being a non-deformable triplet identified by the centroid of the related subchain and its endpoints. Fig. 3 depicts an example of this model for two consecutive scales.

Requirement iii) is reached through our cost function. We first observe that, as mentioned in Sect. 2, fragment pairs characterized by high contact frequencies can reliably be considered in close proximity, but the converse does not need to be true: pairs with low contact frequencies do not need to be far apart. We thus avoid to consider the lowest frequencies in our cost function, which, anyway, can sufficiently determine the problem by exploiting the geometrical constraints. The resulting expression is:

$$\Phi(\mathscr{C}) = \sum_{i,j \in \mathscr{L}} n_{i,j} \cdot d_{i,j} \tag{2}$$

where $\mathscr{C}$ is the configuration of the subchain being estimated, $\mathscr{L}$ is the set of bead pairs that are likely to be close to each other, and $n_{i,j}$ is the contact frequency characterizing the (i,j)-th pair. Thus, no target distance is included in the formula: the contact frequency data are directly used to weight the contributions of the individual pairs in the summation. It is apparent that an unconstrained optimization of this cost function would find global minima in each configuration with $d_{i,j} = 0$ for all $(i,j) \in \mathscr{L}$. The constraints, however, make these solutions unfeasible.

Table 2 Chart of problems and advantages in the previous state of the art.

	Problems	**Advantages**
Constrained Optimization Dekker *et al.* [5] Fraser *et al.* [9] Duan *et al.* [8] Baù *et al.* [2]	Very high dimensionality. No confidence intervals can be computed to measure the uncertainty of the structure obtained.	First attempt of conversion of a set of noisy contact frequencies measurements into more interpretable data. Introduction of constraints based on the structure of the chromatin fiber.
Bayesian Inference Russeau *et al.* [16] (MCMC5C) Hu *et al.* [10] (BACH)	Any evaluation of structural variations of chromatin at different resolution scales. No geometrical constraints. Geometrical inconsistencies given by translation of contact frequencies into distances.	Bayesian approach to sample the whole space of solutions. Introduction of systematic biases into the data model (BACH).
Polymer Models Nagano *et al.* [15] Meluzzi *et al.* [14]	Complexity of the system.	Conversion from frequencies into distances not required. Integration of polymer physics into the 3D chromatin structure model.

Finally, requirement iv) is satisfied by our estimation algorithm. Although the configurations that are not compatible with the constraints are not feasible solu-

tions, it is expected that the cost function reaches minimum values for many different feasible configurations. To be able to sample the solution space, we treat the objective function as a negative log-density, and use a Monte Carlo approach to find high-probability configurations. In practice, we use a classical simulated annealing procedure [12], where the model updates are proposed through quaternion operators [11]. This choice allows us to maintain automatically the coherence of the reconstructed chain at each update, thus avoiding to check the fit to most of the constraints before continuing with the iteration. Indeed, the compatibility of the current solution with the constraints must only be checked against possible spatial interferences between pairs of beads. Since so many configurations fit well the data and the constraints, different runs of this stochastic procedure will produce different highly reliable results, whose structures should reproduce the variety of the configurations assumed by the chromatin chain in the experimental cell population. Our multiscale approach can also be exploited to generate different configurations of the subchains at any resolution, and then combine them to produce, recursively, different configurations of the overall chain.

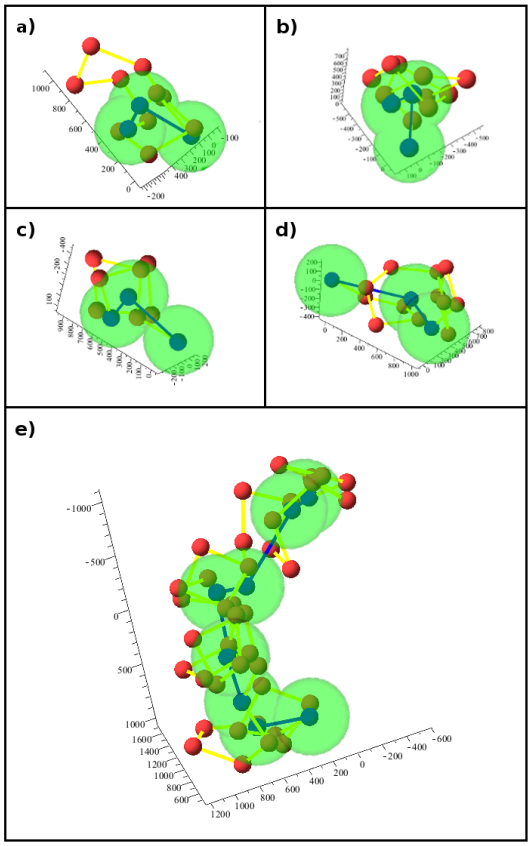

Fig. 3 a-d) Consecutive fragments of the chromatin fiber, represented as bead sequences (red balls linked by yellow segments), and as centroid-endpoints triples (blue balls linked by blue segments). The larger spheres represent the assumed sizes for the beads at the lower resolution. e) Lower-resolution chain composed by the fragments in a-d).

4 Conclusions

In this chapter, we propose a new approach for the estimation of chromatin configurations starting from HI-C contact frequency data. The main characteristics of our approach are:

- The data-fit function does not require the translation of frequencies into Euclidean distances.
- The multiscale bead-chain model can be equipped with biophysical constraints; any prior information available must be translated into geometrical constraints.
- The probabilistic procedure samples the solution space so that multiple configurations compatible with both the data and the constraints can be found.
- The model evolution during the iterations is obtained through quaternion operators.

Thanks to these features, our procedure avoids some of the drawbacks in the algorithms proposed so far in the literature. Also, our algorithm is conceptually simple, and amenable to be speeded up by exploiting several levels of parallelism. As a proof of principle, we have performed some tests on real HI-C data from human cells [4]. In these tests, we obtained a number of different structures characterized by similar values of the cost function but showing a few distinct spatial behaviors (two examples are shown in Fig. 4, from data related to the long arm of the human chromosome 1 [13]). The macroscopic appearance of these structures is compatible

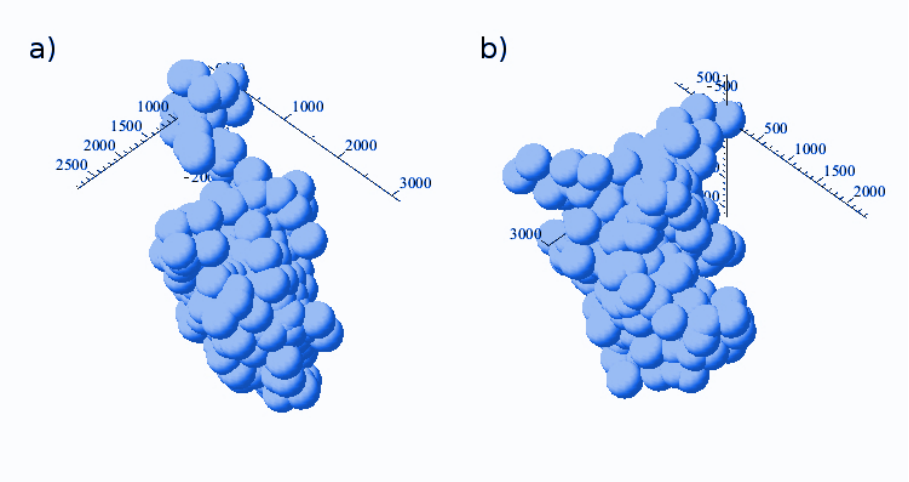

Fig. 4 Two typical configurations resulting from our experiments (measurements in nm).

with the expected shape of a portion of chromosome.

In conclusion, we have generated an algorithm that can substantially contribute to the elucidation of chromosomal structure, by producing families of structures

compatible with biological information. Our procedure is also innovative in the use of quaternions to evolve the model during the estimation process.

Acknowledgements This work has been funded by the Italian Ministry of Education, University and Research, and by the National Research Council of Italy, Flagship Project InterOmics, PB.P05. The authors are indebted to Luigi Bedini and Aurora Savino for helpful discussions.

References

1. Amann,R., Fuchs,B.M. (2008): Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques, Nature Reviews Microbiology 6: 339-348.
2. Baù,D., Marti-Renom,M.A. (2011): Structure determination of genomic domains by satisfaction of spatial restraints, Chromosome Research 19: 25-35.
3. van Berkum,N.L. et al. (2010): Hi-C: a method to study the three-dimensional architecture of genomes, J. Vis. Exp. 39: 1869-1875.
4. Caudai,C. et al. (2014): Reconstructing 3D Chromatin Structure from Chromosome Conformation Capture Data, InterOmics Flagship Project, Report cnr.isti/2014-PR-003, National Research Council of Italy - ISTI, Pisa.
5. Dekker,J. et al. (2002): Capturing chromosome conformation. Science 295: 1306-1311.
6. Dixon,J.R. et al. (2012): Topological domains in mammalian genomes identified by analysis ofchromatin interactions, Nature 485: 376-380.
7. Duggal,G. et al. (2013): Resolving spatial inconsistencies in chromosome conformation measurements, Algorithms for Molecular Biology 8, 8.
8. Duan,Z. et al. (2010): A three-dimensional model of the yeast genome, Nature 465: 363367.
9. Fraser,J. et al. (2009): Chromatin conformation signatures of cellular differentiation, Genome Biology 10, R37.
10. Hu,M. et al. (2013): Bayesian inference of Spatial organizations of chromosomes, PLOS Comp. Biol. 9, 1002-893.
11. Karney,C.F. (2007): Quaternions in molecular modeling, J. Mol. Graph. Model. 25: 595-604.
12. Kirkpatrick,S. et al. (1983): Optimization by Simulated Annealing, Science 229: 671-680.
13. Lieberman-Aiden,E. et al. (2009): Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome, Science 326: 289-293.
14. Meluzzi,D., Arya,G. (2013): Recovering ensembles of chromatin conformations from contact probabilities, Nucleic Acid Res. 41: 63-75.
15. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., Fraser, P.(2013): Single-cell hi-c reveals cell-to-cell variability in chromosome structure. Nature 502: 59-64
16. 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.
17. Yaffe,E., Tanay,A. (2011): Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture, Nature Genetics 43: 1059-1067.