Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Geometrical probability approach for analysis of 3D chromatin structure in interphase cell nuclei

E. Gladilin*, S. Goetze†, J. Mateos-Langerak†, R. van Driel†, K. Rohr* and R. Eils*
* German Cancer Research Center, Theoretical Bioinformatics,
Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany
Email: e.gladilin@dkfz-heidelberg.de
† Swammerdam Institute for Life Sciences, University of Amsterdam,
BioCentrum Amsterdam, Kruislaan 318, 1098 SM Amsterdam, The Netherlands

***Abstract*— Investigation of 3D chromatin structure in interphase cell nuclei is important for the understanding of genome function. For a reconstruction of the 3D architecture of the human genome, systematic fluorescent *in situ* hybridization in combination with 3D confocal laser scanning microscopy is applied. The position of two or three genomic loci plus the overall nuclear shape were simultaneously recorded, resulting in statistical series of pair and triple loci combinations probed along the human chromosome 1 q-arm. For interpretation of statistical distributions of geometrical features (e.g. distances, angles, etc.) resulting from finite point sampling experiments, a Monte-Carlo-based approach to numerical computation of geometrical probability density functions (PDFs) for arbitrarily-shaped confined spatial domains is developed. Simulated PDFs are used as bench marks for evaluation of experimental PDFs and quantitative analysis of dimension and shape of probed 3D chromatin regions. Preliminary results of our numerical simulations show that the proposed numerical model is capable to reproduce experimental observations, and support the assumption of confined random folding of 3D chromatin fiber in interphase cell nuclei.**

I. Motivation

The dynamic 3D folding of the chromatin fiber in the interphase nucleus is a key element in the epigenetic regulation of gene expression in eukaryotes [2], [7]. Chromatin structure can be studied by 3D confocal laser scanning microscopy (CLSM) after fluorescent in situ hybridization (FISH) labeling of specific sequence elements in the genome under conditions that preserve biological structure. Using Bacterial Artificial Chromosomes (BACs), FISH enables selective visualization of complete individual chromosomes or subchromosomal domains, or smaller genomic regions of only few hundreds kilobase pairs (Kb = 10^{-3} Mb). In the present study four CSLM imaging channels were used for the simultaneous visualization of three genomic sequence elements (using three spectrally differently labeled BACs) and the overall size and shape of the interphase nucleus (using DAPI labeling of all nuclear DNA). For probing large chromatin regions with a finite number of sampling points, multiple measurements in different cells have to be performed. However, reconstruction of 3D chromatin structure in a 'piece-by-piece' manner from such series of finite point samplings assumes the existence of a mechanically conservative 3D structure that exhibits only small topological cell-to-cell variations. In contrast, distance measurements in different cells show extensive cell-to-cell variations that are not due to measuring errors and therefore indicate a flexible and dynamic structure of interphase chromatin [5]. Repetitive measurements in many otherwise identical cells yield statistical series of simultaneously labeled pairs or triplets of BACs, i.e. coordinates of their mass center points, and the corresponding geometrical features (e.g. pairwise distances) that are used for quantitative analysis of probed chromatin regions. In a number of previous studies [4], [5], [8], statistical models of 3D chromatin folding were proposed. These models provide a qualitative description for some basic experimental observations, for example the $D(L) \approx \sqrt{L}$ relationship between 3D physical distance (D) and genomic length (L) for each two probes along the DNA on a small $L < 2$ Mb genomic scale. However, the saturation plateau of $D(L) \approx const$ observed on larger genomic scales $L > 2$ Mb is not yet satisfactory explained in the literature. The distribution of fluorescent probes for a random folded DNA fiber and probability densities for the corresponding geometrical features (e.g. BAC distances) are derived by authors on the basis of some general thermodynamic principles, e.g. equilibrium state, resulting in a general integral form for sought probability density function (PDF). Closed form solutions of such integrals can only be obtained for some particularly simple geometries of spatial domains, e.g. a spherical ball. However, for the validation of some structural hypotheses and comparative analysis of experimental data, PDFs for a wide range of geometrical features and domain shapes may be of interest. In this article, we present a general approach for numerical computation of arbitrary geometrical PDFs using Monte-Carlo simulations, which is applied for interpretation of statistical series resulting from 4 channel CLSM of human fibroblast cell nuclei. Exemplarily, we restrict our simulations to a natural assumption of uniform confined point distributions, although any arbitrary non-uniform point density function can be used instead. Our simulation results show that the proposed model is capable to reproduce experimental PDFs and provides numerical bench marks for quantitative analysis of probed spatial confinements in terms of dimension and shape. Furthermore, we present a confined random folding model of 3D chromatin fiber, which gives a qualitative explanation for biphasic behavior of the $D(L)$ relationship.

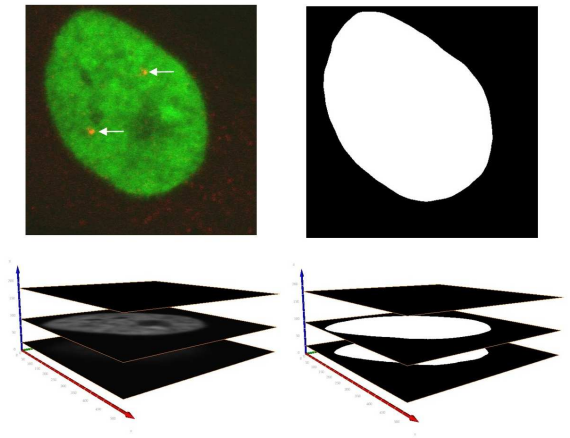

Fig. 1. Left column: 3D CLSM image of a DAPI-stained human fibroblast nucleus (bottom), single slice of 3D image with one BAC in red-color channel (top). Arrows point to two spots of the same BAC corresponding to two sister-chromosomes. Right column: 3D view (bottom) and single 2D slice (top) of the binarized image.

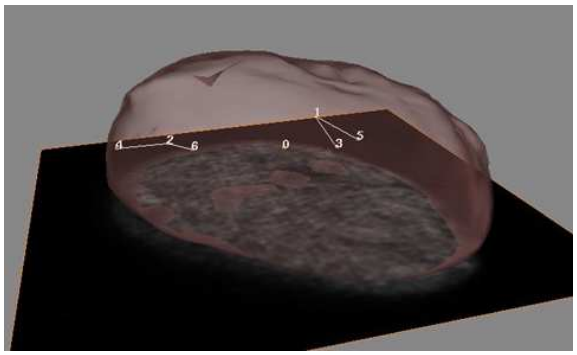

Fig. 2. Segmented 3D CLSM image. Numbered nodes indicate positions of mass centers of the whole nucleus (0) and three BACs in two sister-chromosomes (1-6), respectively.

II. Methods

A. Image Acquisition and Preprocessing

3D confocal laser scanning microscopy images of DAPI-stained human fibroblast nuclei are used for geometrical reconstruction of 3D chromatin structure in interphase nuclei of human fibroblasts, see Fig. 1. Besides overall nucleus shape, up to 3 BAC regions are simultaneously labeled via the FISH technique resulting in 4 channel 3D image of the nucleus. 3D images of all channels are consistently preprocessed using following basic steps [3]:

- Fourier band-pass image smoothing,
- threshold-based image segmentation.

After thresholding and segmentation, target structures are represented in the binarized images by clearly bounded white regions, see Fig. 1 (right column). Pointwise representation of segmented nuclear and BAC regions is obtained by computing their mass centers:

$$x_i^{mc} = \frac{1}{N}\sum_{j=1}^{N} x_i^j, \tag{1}$$

where x_i^j is the i-th coordinate of the j-th of totally N voxels of one segmented region (e.g. particular BAC or entire nucleus domain). Since each BAC label produces two signals corresponding to two sister-chromosomes, totally 7 mass center points are localized after preprocessing 4 channel 3D images of the nucleus, see Fig. 2.

B. Strategies of finite point sampling of 3D chromatin fiber

Due to the limited number of independent color-channels, only few gene loci along the chromatin fiber can simultaneously be sampled in one single cell. For continuous dense sampling of larger chromosome regions and assessment of statistical cell-to-cell variations in chromatin structure, each combination of BACs is probed in 50 different cells resulting in statistical series of coordinates of mass center points. On the basis of these statistical series, probability distributions for geometrical features (distances, angles, etc.) are calculated. Two strategies for placement of sampling probes along the DNA fiber and assessment of statistical series are applied:

1) series of BAC pairs with increasing genomic distances,
2) series of BAC triplets.

a) Series of BAC pairs: with increasing genomic distance are used for assessment of the $D(L)$ relationship between physical D and genomic L distances for each BACs at small $L = 0.1 - 3$ Mb and large $L = 3 - 28$ Mb genomic scales, see Fig. 3 (left).

b) Series of BAC triplets: are measured for assessment of geometrical features of 3-point combinations (e.g. triangle angles, etc.), which indicate structural variability of probed 3D domains along the DNA fiber. Stochastical analysis of statistical distributions of geometrical features of BAC triplets is applied for the estimation of cross section dimension and shape of sampled chromosome regions. For dense sampling of larger genetic regions in a piece-by-piece manner, a *moving mask* technique is applied. That is three BACs are placed along the DNA in a way that each next triplet has an overlap with the previous one by two common BACs, see Fig. 3 (right). This approach is applied for 3D visualization of finite point sets using multidimensional scaling of cross-distance matrices, see Section III-D.

C. Geometrical features of BAC combinations

Spatial distribution of BAC pairs and triplets is analyzed using following geometrical features:

- pairwise distances between BACs,
- radial distances of BACs w.r.t. the nucleus center,
- angles of triangles spanned by BAC triplets,
- spatial orientation of BAC triplets w.r.t. the nucleus center.

Fig. 4 gives an overview over geometrical features of BAC triplets used for numerical computation of PDFs via the Monte-Carlo approach as described in Section II-E.

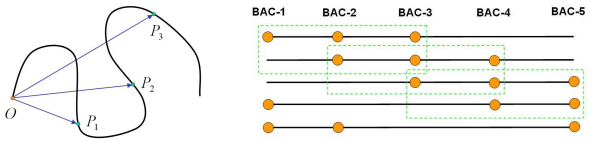

Fig. 3. Left: the relationship $D(L)$ between physical distances $D_i = |OP_i|$ and genomic lengths $L_i = \int_O^{P_i} dl$ for a finite number of sampling points ($i = 1...N$) sequentially placed along the DNA provides insights into 3D chromatin folding. Right: schematic view of the *moving mask* approach in case of 5 BAC triplets, green frames show overlapping pairs of BACs.

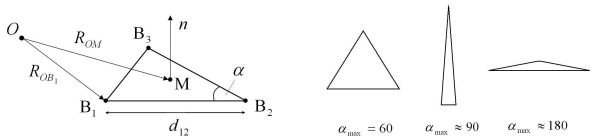

Fig. 4. Left: geometrical features of BAC triplets B_i: pairwise distances $d_{ij} = |B_iB_j|$, radial distances w.r.t. a fixed point $\mathbf{R}_{OB_1}$, triangle angles α, orientation of BAC triangle w.r.t. a fixed point $(\mathbf{n} \cdot \mathbf{R}_{OM})$. Right: maximum triangle angle $\alpha_{\max}$ as a feature of triangle shape.

c) Pairwise distances: between BACs are computed as Euclidean distances for all 3 pairs of BACs in each triplet $d_{ij} = |B_iB_j| \in [0, S_{\max}]$, $\forall i = 1..3, j > i$, where $S_{\max}$ is the maximum cross section of the cell nucleus.

d) Radial distances: of BACs w.r.t. the mass center of the nucleus are computed as Euclidean distances between mass centers of the nucleus O and each BAC region B_i, i.e. $R_{OB_i} = |OB_i| \in [0, \frac{S_{\max}}{2}]$.

e) Maximum angle: of BAC triangle $\alpha_{\max}$ serves as a feature of triangle shape, see Fig. 4 (right). The PDF of $\alpha_{\max} \in [60, 180]$ provides insights in the overall shape of the probed domain.

f) Spatial orientation: of BAC triplets w.r.t. the nucleus center is characterized by a scalar product $s = (\mathbf{n} \cdot \mathbf{R}_{OM}) \in [-1, 1]$, where $\mathbf{n}$ is the triangle normal and $\mathbf{R}_{OM}$ is the vector pointing from center of the nucleus to the middle of BAC triangle. $s = 0$ means that $\mathbf{R}_{OM}$ lies in the triangle plain, while $|s| = 1$ indicates that $\mathbf{R}_{OM}$ is perpendicular to the triangle plain.

D. Geometrical probability density function

Series of finite point samplings of 3D chromatin fiber in different cells yield statistical distributions of invariant geometrical features (e.g. pairwise distances, angles etc.), which are analyzed using geometrical probability techniques. Stochastical analysis of such statistical series is aimed at

- investigation of the order of randomness and
- quantification of dimension and shape of probed chromatin regions,

and is based on construction of geometrical PDF.

Formally, the PDF of a probability distribution is defined as a non-negative function $p(x) > 0$ of a statistically distributed variable x such that the integral

$$P(A \le x \le B) = \int_A^B p(x)dx \le 1, \tag{2}$$

gives the probability $P(A \le x \le B)$ for the variable x being found in the interval $A \le x \le B$. From (2) it immediately follows that

$$\int_{-\infty}^{+\infty} p(x)dx = 1. \tag{3}$$

For a discrete distribution of x_i ranging in the interval $x_i \in [A, B]$, $\forall i = 1...N$, the PDF can be constructed using the histogram function $h_j(x_i)$, which is defined as an array of tabulated frequencies of x_i being found within the j-th of totally n intervals $\frac{j}{n}(B-A) \le x_i \le \frac{(j+1)}{n}(B-A)$:

$$p_j(x_i) = C\,\frac{h_j(x_i)}{N}, \tag{4}$$

where C is the normalization constant resulting from the condition (3)

$$C = \left(\sum_{j=1}^{n} \frac{h_j(x_i)}{N} \frac{(B-A)}{n} \right)^{-1}. \tag{5}$$

For a uniform random distribution of N sampling points $\mathbf{r}_i$ with Cartesian coordinates x_j in a spherical k-dimensional ball $\boldsymbol{B}_k = \{\mathbf{r}_i(x_j) : \sum_{j=1}^{k} x_j^2 \le R^2\}$, where R denotes the radius of $\boldsymbol{B}_k$, the PDFs $p_k(d)$ for pairwise distances $d = ||\mathbf{r}_{p=1..N} - \mathbf{r}_{q=(p+1)..N}||$ between the points can be obtained in a closed form [6]:

$$\begin{aligned} \boldsymbol{B}_1: &\quad p_1(d) = \tfrac{1}{R}(1 - \tfrac{d}{2R}) \\ \boldsymbol{B}_2: &\quad p_2(d) = \tfrac{2d}{R^2} - \tfrac{d^2}{\pi R^4}\sqrt{4R^2 - d^2} - \tfrac{4d}{\pi R^2}\arcsin\left(\tfrac{d}{2R}\right) \\ \boldsymbol{B}_3: &\quad p_3(d) = \tfrac{3d^2}{R^3} - \tfrac{9d^3}{4R^4} + \tfrac{3d^5}{16R^6} \end{aligned} \tag{6}$$

Plots of $p_k(d)$ in case of unit 1D/2D/3D-balls (i.e. $2R = 1$) are shown in Fig. 5. As one can see, PDFs for pairwise distances essentially depend on the spatial dimension k of $\boldsymbol{B}_k$. These PDFs can be seen as characteristic signatures of random point distributions for 1D, 2D and 3D isotropic spherical confinements, respectively. Interestingly, the distance with the highest probability $(d_m \in [0, 1] : \max(p(d)) = p(d_m))$ for a unit spherical ball of dimension higher than 1 is not the smallest-possible, but some intermediate one:

$$\begin{aligned} \boldsymbol{B}_1: &\quad d_m = 0 \\ \boldsymbol{B}_2: &\quad d_m = 0.42 \\ \boldsymbol{B}_3: &\quad d_m = 0.52 \end{aligned} \tag{7}$$

These key-values together with further standard PDF features

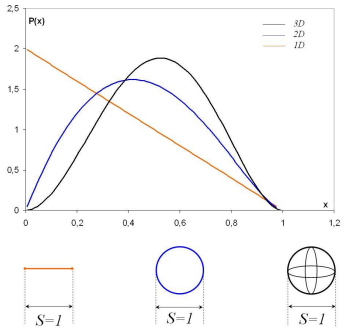

Fig. 5. Plots of theoretical probability density functions $p_k(d)$ for pairwise distances between randomly distributed points in case of spherical k-dimensional balls ($k = 1, 2, 3$) with a unit cross section dimension ($S = 2R = 1$).

such as first four statistical moments M_i:

$$\begin{aligned} M_1 &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} p(d_i) \\ M_2 &= \sqrt{\tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(p(d_i) - M_1)^2} \\ M_3 &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(\tfrac{p(d_i)-M_1}{M_2}\right)^3 \\ M_4 &= \left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(\tfrac{p(d_i)-M_1}{M_2}\right)^4\right) - 3 \end{aligned} \tag{8}$$

can be used for quantification of the order of randomness of BAC distributions on the basis of experimental PDFs of pairwise distances. If, for instance, we would observe in our experiment a PDF for pairwise distances, which behaves very much the same as $p_3(d)$ in (6), we would have a strong evidence for a completely random distribution of measured points within a probed spatial domain. And the other way round: if we assumed that the underlying 3D point distribution is confined, random and isotropic, statistical features of the experimental PDF $\bar{p}(d)$ such as (7) gave us an estimate for the upper bound of the cross section dimension $\bar{S}$ of the probed spherical domain, namely

$$\bar{S} = \left(\frac{S}{\bar{d}_m}\right)\bar{d}_m \approx 1.92\,\bar{d}_m, \tag{9}$$

where $S = 1$ is the cross section dimension of the reference unit ball and $\bar{d}_m : \max(\bar{p}(d)) = \bar{p}(\bar{d}_m)$.

E. Numerical computation of PDFs for an arbitrary domain

The approach for bench marking experimentally observed PDFs of pairwise distances vs theoretically predicted PDFs can be extended to the case of an arbitrary geometrical feature and arbitrarily-shaped k-dimensional confinement $\Omega \subset \mathbb{R}^k$ with the boundary $\Gamma \subset \Omega$. For more complex geometries, the PDF for pairwise point distances or any other geometrical feature can not be derived in a closed form. However, it can be computed numerically, for example, via a Monte-Carlo simulation. The basic steps of our simulation scheme are as follows

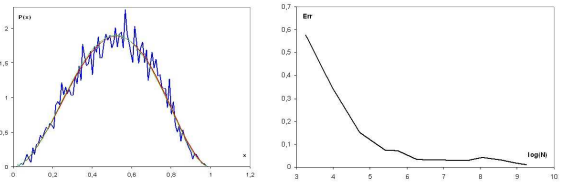

Fig. 6. Left: results of numerical computation of PDFs of pairwise distances for a unit 3D spherical domain with $N = 270$ (blue) and $N = 5219$ (green) random points vs theoretical solution (red). Right: L_2 error norm of numerically computed PDFs w.r.t. theoretical solution as a function of $\log(N)$, where N is the number of random points in the Monte-Carlo simulation.

1) specify the domain Ω in a suitable parametric form (e.g. surface or volumetric meshes, point clouds, etc.),
2) generate sufficiently large number of random points $\mathbf{r}_i$ within the bounding box of Ω and select only the points lying inside the domain $\mathbf{r}_i \in \Omega$
3) compute geometrical features (distances, angles, etc.) for all pairs and triplets of $\mathbf{r}_i \in \Omega$,
4) compute corresponding histogram and PDF for simulated statistical series of geometrical features.

At the end, the PDF for a geometrical feature and spatial confinement is given by an $(n-1)$-array of tabulated values corresponding to n intervals of the histogram function. Further details on computation of PDFs for some special cases of 3D domain geometry are in Section III.

F. Confined random folding model

For the interpretation of the $D(L)$ relationship between physical and genomic distances (see Fig. 3 (left)), a confined random folding model of 3D chromatin fiber is proposed. We consider 3D chromatin fiber or one of its fragments to be randomly folded within a confined spatial domain $\Omega \subset \mathbb{R}^3$: $\sum_{i=1}^{3}(x_i - x_i^{mc})^2 \leq R^2, \forall x_i \in \Omega$, where x_i are coordinates of points, x_i^{mc} is the mass center of Ω, and R is a characteristic dimension of the confinement Ω. The simulation of a randomly folded 3D fiber begins with the generation of a confined random point distribution for Ω as described in Section II-E. These points are understood as vertex nodes of a 3D fiber randomly folded in Ω, see Fig. 11. An algorithm is developed to perform the reconstruction of a non-closed, loop-free 3D fiber connecting all points of Ω pairwise using the closest neighborhood connectivity. Details on the validation of the confined random folding model are in Section III-E.

III. Experimental Results

In this section, we present the results of stochastical simulations of confined random point distributions and 3D chromatin fiber folding carried out for the interpretation of experimental observations.

A. Numerical simulation vs theoretical solution

First, numerical algorithms for computation of geometrical PDFs are validated by a direct comparison with closed form solutions (6). Fig. 6 (left) shows the results of numerical

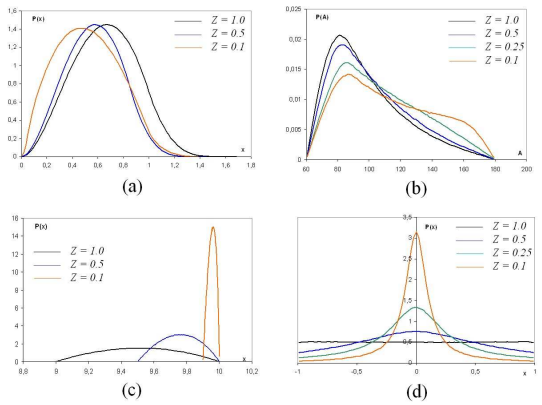

Fig. 7. Simulated PDFs of geometrical features for isotropic ($Z = 1.0$) an Z-scaled spatial domains: (a) pairwise distances, (b) maximum angles of point triplets, (c) radial distances w.r.t. a fixed point, (d) triangle orientation ratio.

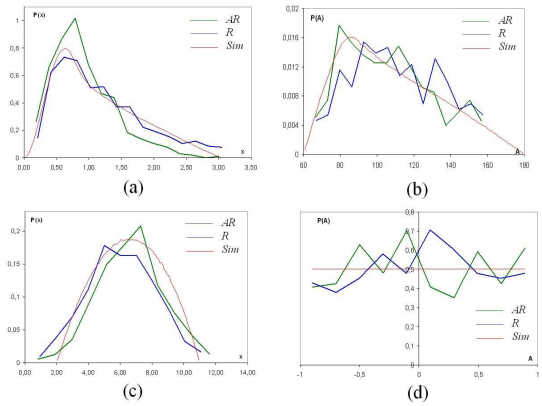

Fig. 8. Geometrical PDFs of statistical series of probing five ridge (blue curves) and antiridge (green curves) regions of human chromosome 1: (a) pairwise BAC distances, (b) maximum angles of BAC triplets, (c) radial distances of BACs w.r.t. the nucleus center, (d) orientation ratio of BAC triangles. Red curves denote simulated PDFs.

computations of the PDF of pairwise distances $p_3(d)$ for $N = 113$ and $N = 5219$ sampling points of Monte-Carlo simulations. Plot 6 (right) illustrates the reduction of the numerical error vs theoretical solution with increasing number of sampling points. From numerical point of view, $N = 350$ sampling points is an acceptable lower bound for accurate computation of smooth PDFs. However, one should take into account that $N = 350$ sampling points correspond to $\frac{N^2-N}{2} = 61750$ pairwise distances, i.e. single distance measurements! One can also reversely calculate the number N of virtual sampling points corresponding to N_d distance measurements: $N \approx \sqrt{2N_d}$. This means that $N_d = 500$ distance measurements correspond to only $N = 31$ virtual sampling points, and, in order to simulate $N = 100$ sampling points, $N_d = 9900$ distance measurements are required.

B. Impact of domain geometry

In order to investigate impact of domain geometry on the PDF pattern, numerical simulations are carried out Fig. 7 show the simulation results for PDFs of different geometrical features discussed above and several values of Z-scaling factor. As one can see, domain geometry has a strong impact on PDF patterns, which can be exploited for the interpretation of experimental PDF curves.

C. Geometrical PDFs of experimental series

Our measurements did focus on two regions of the q-arm of human chromosome 1 in G1-phase human primary fibroblasts. The human transcriptome map [1] shows the presence of a gene-dense region of highly expressed genes (named a region of increased gene expression (ridge)) and a gene-sparse region (named antiridge), each comprising several Mb. One ridge (R) and one antiridge (AR) region along the q-arm of human chromosome 1 were probed on a scale 0.7 - 3.3 Mb, see Table Table I. Average distances and standard deviations for these statistical series are summarized in Table II. Distances for each

TABLE I

GENOMIC DISTANCES (IN MB) FOR FIVE AR/R BACS.

AR	1	8	19	28	35
1	0	0.81	1.61	2.49	3.26
8		0	0.80	1.68	2.45
19			0	0.88	1.65
28				0	0.77
35					0

R	5	12	21	28	39
5	0	0.84	1.58	2.39	3.20
12		0	0.74	1.55	2.36
21			0	0.81	1.62
28				0	0.81
39					0

TABLE II

AVERAGE PHYSICAL DISTANCES (IN μM) FOR FIVE AR/R BACS.

AR	1	8	19	28	35
1	0	0.65±0.36	0.9±0.52	1.30±0.64	1.37±0.74
8		0	0.61±0.42	1.21±0.60	1.27±0.45
19			0	0.74±0.41	1.08±0.49
28				0	0.76±0.32
35					0

R	5	12	21	28	39
5	0	0.72±0.33	1.13±0.44	1.32±0.95	2.12±1.10
12		0	0.52±0.27	1.23±0.52	2.18±1.09
21			0	0.92±0.28	2.51±0.76
28				0	2.57±0.77
39					0

BAC triplet were measured in about 50 clonally grown fibroblast cells and for each of these measurements all geometrical features described in Section 3.2 were computed. Altogether, R/AR chromatin regions sampled by five BAC triplets were totally probed $N_t = 200 - 272$ times, which corresponds to $N_d = 3N_t = 600 - 816$ pairwise distances R_{ij}/AR_{ij} and, recalling the discussion in Section III-A, $\sqrt{2N_d} \approx 35 - 40$ virtual sampling points, respectively. Experimentally assessed geometrical PDFs for entire R/AR series of five BAC probes vs simulated PDFs are plotted in Fig 8. Quantitative comparison between experimental p_i^{ex} and simulated p_i^{sim} PDF patterns carried out on the basis of least square norms $||p_i^{ex} - p_i^{sim}||$ and statistical moments M_3 (skew) and M_4 (kurtosis) indicates a random distribution of BACs within anisotropically shaped 3D confinements, whereas some significant differences in spatial structure of R and AR regions were observed:

- AR domain probed by five BACs is more compact and smaller in size ($1.45 \times 0.42 \times 0.28$ μm) than R domain ($3.16 \times 1.62 \times 0.26$ μm),
- AR BACs have in average larger radial distances ($\bar{r}_{AR} = 6.62 \pm 1.98$ μm) w.r.t. the nucleus center compared to R BACs ($\bar{r}_R = 5.98 \pm 2.23$ μm)

D. 3D visualization of average cross-distance matrix

Ridge and antiridge BAC triplets have been placed according to the scheme shown in Fig. 3 (right). This BAC placement strategy was introduced for consistent 3D visualization of finite point probes of chromatin regions using a multidimensional scaling (MDS) approach [9], which requires a matrix of cross-distances d_{ij} between all BACs, see Table II. The result of 3D visualization of average R/AR distance matrices after decomposing d_{ij} via the MDS is shown in Fig. 9 (top). In view of large statistical deviations in cell-to-cell distance measurements, the probed chromatin regions can not be regarded as rigid objects with a same constant shape. Thus, 3D loops depicted in Fig. 9 represent an *average* shape of probed R/AR regions resulting from a statistical series of distance measurements. Fig. 9 (middle, bottom) show statistical uncertainty spheres whose radii are equal to standard deviations of the corresponding BAC coordinates $r_i = \sigma_i$

E. Validation of confined random folding model

A pure geometrical approach has been applied for qualitative validation of a confined random folding model of 3D chromatin fiber. A synthetic 3D fiber is constructed using the basic computational steps described in Section II-F. The initial random point cloud consisting of $N = 528$ vertices and the corresponding 3D fiber are shown in Fig. 10. The $D(L)$ relationship between Euclidean distance D and genomic length L of all vertices w.r.t. a fixed starting point of the fiber is shown in Fig. 11 (bottom, red curve). This is a typical behavior of $D(L)$ function with large periodic oscillations appearing after the $\sqrt{L}$ regime for small L. At this point, we want to draw attention to the fact that the result of numerical computation of $D(L)$ essentially depends on the choice of the starting point for the construction of 3D fiber. For example, if the starting

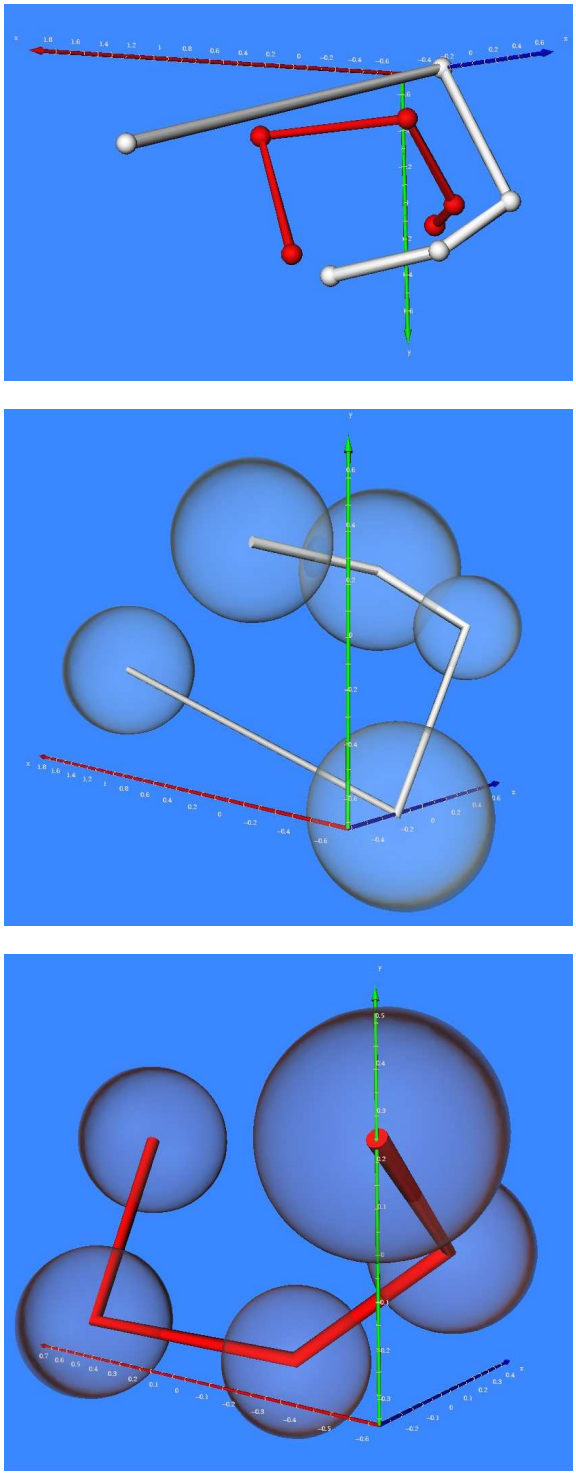

Fig. 9. From top to bottom: 3D visualization of average cross-distance matrices for five ridge (white polygon) and antiridge BACs (red polygon) probed along human chromosome 1 on a genomic scale $0.7 - 3.3$ Mb, and their statistical uncertainty spheres ($r_i = \sigma_i$).

point is selected on the boundary of a spherical confinement with the radius R, the maximum possible Euclidean distance is $D_{\max} = 2R$, whereas if the starting point is selected in the center of the sphere, the maximum value is $D_{\max} = R$. Also, periodic patterns of $D(L)$ for different starting points are shifted w.r.t. each other with a random phase. From experimental point of view it has following consequences: since experimental $D(L)$ curves are constructed from sequential measurements of the same genetic region in different cells, the physical position of the starting BAC, as well as overall 3D folding of the probed genomic region are varying from cell to cell. This means that an experimentally obtained $D(L)$ relationship is, in fact, the result of averaging N single $D(L)$ curves, where N is the number of measurements in different cells. The result of numerical simulation of such average $D(L)$ is shown in Fig. 11 (bottom, black curve). As one can see it exhibits a biphasic behavior very much similar to experimentally assessed $D(L)$ curve in Fig. 11 (top). Obviously, the saturation plateau of experimentally assessed $D(L)$ relationships results from statistical smoothing of single $D(L)$ curves due to measurements in different cells. However, one can still recognize remaining quasi-periodic oscillations of single random $D(L)$ curves in the average $D(L)$ on large genomic scale $L > 2$ Mb.

IV. Conclusion

In this article, we have presented a novel approach for stochastical analysis and visualization of finite point sampling of 3D chromatin in interphase cell nuclei. The core idea of our approach consists in application of geometrical probability techniques for interpretation of statistical series of finite point sampling of chromatin regions. Numerically computed probability density functions (PDFs) serve as bench marks for the validation of experimentally observed statistical distributions of canonic geometrical features of two- and three-point combinations, e.g. pairwise and radial distances, angles, etc. We have introduced a general Monte-Carlo-based simulation scheme for computation of PDFs of geometrical features of random point distributions for arbitrarily-shaped confined 3D domains, and derived numerical criterions for the estimation of the order of randomness of observed statistical distribution as well as dimension and shape of probed chromosome regions. Preliminary experimental results of sampling human chromosome 1 in primary human fibroblasts in G1 cell cycle phase by five overlapping ridge and antiridge BAC triplets on a genomic scale $0.7 - 3.3$ Mb support the assumption of confined random folding of 3D DNA fiber in interphase cell nuclei. We have proposed a geometrical model of confined random chromatin folding, which is capable to reproduce experimentally observed relationship between physical and genomic distances on a large genomic scale. Further sampling experiments with 4 simultaneous BAC labels and larger statistical series ($\approx 10^4$ distance measurements) are required to provide a more consistent source of geometrical information for distinctive analysis of 3D chromatin structure.

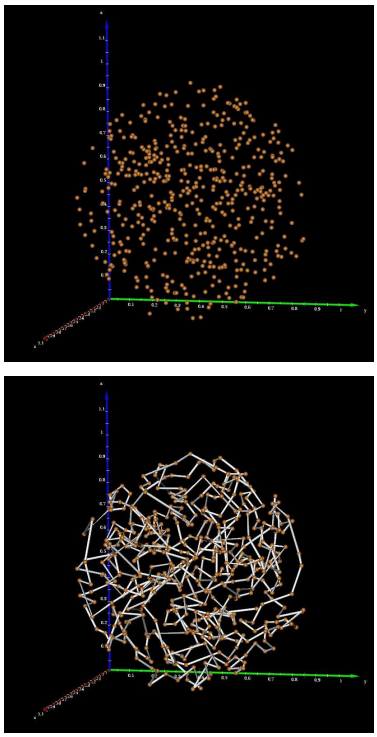

Fig. 10. Top: initial random distribution of $N = 528$ points for a unit spherical confinement. Bottom: randomly folded 3D fiber computed on the basis of the initial point cloud.

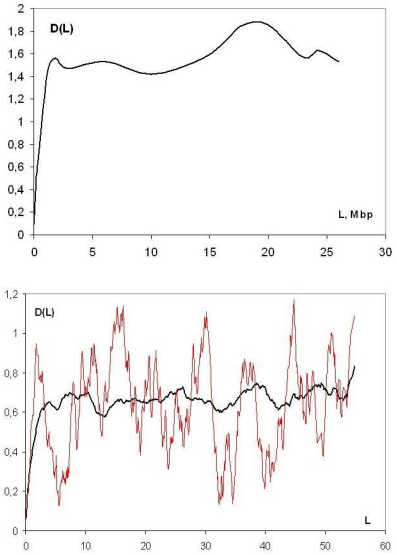

Fig. 11. Top: experimentally assessed relationship $D(L)$ between physical and genomic distances on a scale $0.1 - 28$ Mb for human chromosome 1. Bottom: numerically computed $D(L)$ for a synthetic fiber randomly folded within a unit 3D spherical confinement (cf. Fig. 10): red curve corresponds to $D(L)$ for a single (starting point dependent) simulation, black curve shows an average $D(L)$ for $N/2 = 264$ simulation runs with varying starting points.

ACKNOWLEDGEMENTS

This work is part of the 3DGENOME program, which is supported by the European Commission: contract LSHG-CT-2003-503441.

REFERENCES

[1] H. Caron et al. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, 291:1289–1292, 2001.

[2] T. Cremer et al. Higher order chromatin architecture in the cell nucleus: on the way from structure to function. *Biol. Cell.*, 96:555–567, 2004.

[3] E. Gladilin et al. Topological analysis of 3D cell nuclei using finite element template-based spherical mapping. In *Proc. SPIE MI*, volume 6144, pages 1557–1566, 2006.

[4] P. Hahnfeldt et al. Polymer models for interphase chromosomes. *Proc. Natl. Acad. Sci. USA, Biophysics*, 90:7854–7858, 1993.

[5] R. K. Sachs et al. A random-walk/giant-loop model for interphase chromosomes. *Proc. Natl. Acad. Sci. USA, Biochemistry*, 92:2710–2714, 1995.

[6] S. J. Tu et al. Random distance distribution of spherical objects: general theory and applications to n-dimensional physics. *Math. Phys.*, 35(31):6557–6570, 2002.

[7] R. van Driel et al. The eucariotic genome: a system regulated at different hierarchical levels. *J Cell. Sci. 116*, pages 4067–4075, 2003.

[8] H. Yokota et al. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. *J Cell. Biol.*, 130(6):1239–1249, 1995.

[9] F. M. Young et al. *Theory and Application of Multidimensional Scaling*. Eribaum Associates, Hillsdale, NJ, 1994.