Physical Origin of the Contact Frequency in Chromosome Conformation Capture Data

Seungsoo Hahn and Dongsup Kim*
Department of Bio and Brain Engineering, Korea Advanced Institute of Science & Technology, Daejeon, South Korea

ABSTRACT Physical proximity between each pair of genomic loci in a nucleus is measured as a form of contact frequency in chromosome conformation capture-based methods. Complexity of chromosome structure in interphase can be characterized by measuring a statistical property of physical distance between genomic loci according to genomic separation along single chromatids. To find a relationship between the physical distance and the contact frequency, we propose a polymer model derived from the Langevin equation. The model is derived by considering a structure of a chromosome as a trajectory of a particle, where each consecutive segment in the chromosome corresponds to a transient position in the trajectory over time. Using chromosome conformation capture data, we demonstrate the functional relationship between the two quantities. The physical distances derived from the mean contact frequencies by the model show a good correlation with those from experimental data. From the model, we present that the mean contact frequency curve can be divided into three components that arise from different physical origins and show that the contact frequency is proportional to the contact surface area, not to the volume of segments suggested by the fractal globule model. The model explains both a decaying pattern of the contact frequency and the biphasic relationship between the physical distance and the genomic length.

INTRODUCTION

Genomic organization is closely related to functional processes occurring in the nucleus (1–6). Chromatin structures have been studied by various experimental techniques such as light microscopy, electron microscopy, cryo-electron microscopy, x-ray scattering, and x-ray crystallography (7–10). Although these techniques are useful in unraveling molecular structure of chromatin or overall shape of nucleus, they are not applicable to solve the three-dimensional structure of genome or long-range interaction between pairs of genomic loci on a genomewide scale. To investigate the complex, genomewide chromosomal structure, various techniques based on the chromosome conformation capture (3C) method have been developed (11–21). These 3C-based techniques have been applied to examine long-range interactions between genomic loci in many organisms (11–14). However, the resulting experimental data contain not only signals but also various systematic errors and noise. To extract information despite these error sources, various statistical approaches and theoretical models have been developed (14–16,22).

Characteristics of chromosome conformation have been described by two quantities: the contact frequency and the physical distance between genomic loci. Both quantities can be measured by experimental methods; 3C-based experiments have been performed to measure the contact frequencies (16), and fluorescent in situ hybridization (FISH) experiments have been performed to measure the physical distance (14,23). The functional dependence of the quantities on genomic length has been used to build various polymer models. Hahnfeldt et al. (24) suggested a random-walk model under a hard spherical boundary. Mirny (25) proposed the fractal globule model to explain the functional dependence of the two quantities on the genomic distance. The random loop model was suggested to explain an asymptotic behavior of the physical distance on a large genomic separation (23). The random-walk/giant loop model was suggested based on the biphasic relationship between the mean-square end-to-end distance and genomic length, where the biphasic relationship means that the physical distance increases like the random-walk model in the region of a short genomic length and it follows the fractal globule model in the region of a large genomic length (26). The multiloop subcompartment (MLS) model proposed that several consecutive loops form a subcompartment, which is a structural chromosomal unit (27). Various computer simulations have been performed to unravel the detailed structure of chromosomes in interphase by integrating all experimental observations based on the polymer models (12,14,15,28).

Several methods of interpreting 3C-based data have adopted a segment-based approach (12–15). In this approach, a target genome is divided into many labeled segments with a certain size, where each segment covers a specific continuous genomic region. Contact frequencies between the segments are determined by analyzing experimental data. This segment-based approach reduces computational requirements and experimental errors (22). Additionally, proper normalization of contact frequencies is necessary to reduce various systematic errors and signal noise (29,30). For further analysis, the contact frequencies must be converted into physical distances between the segments. Dekker et al. (16) assumed that the contact frequency is proportional to the local chromosome concentration around a target

Submitted May 15, 2013, and accepted for publication August 28, 2013.

*Correspondence: kds@kaist.ac.kr

Editor: Laura Finzi.

0006-3495/13/10/1786/10 $2.00

http://dx.doi.org/10.1016/j.bpj.2013.08.043

genomic locus. Duan et al. (12) used a mean contact frequency curve as a standard curve to convert contact frequencies into physical distances. Tanizawa et al. (14) directly measured physical distances between several genomic regions using FISH experiments and created a standard curve based on their experimental results to translate overall contact frequencies. Although it is clear that the contact frequency is inversely proportional to the physical distance, the exact functional relationship remains unknown.

It is obvious that a reasonable interpretation of the contact frequencies obtained from 3C-based experiments is the first step toward constructing the three-dimensional structure of a genome inside a nucleus. Here, we suggest what we believe to be a new approach to understand the physical origin of contact frequencies by establishing a relationship between the genomic length of a segment and its total number of contacts with surrounding genomic parts. We also demonstrate a method for converting the obtained relationship into the contact frequencies, even for the case in which the genomic segments have different genomic lengths each other. From the model, we explain the relationship between the physical distance and the contact frequency. Finally, we suggest possible mechanisms that may dictate the contact frequencies and discuss their physical meaning.

THEORY AND METHODS

For polymer models, a chromosome in a nucleus is considered to be composed of consecutive spheres linked by a string in which each sphere corresponds to a genomic segment in the chromosome (25). To understand the properties of a chromosome structure, we develop a model by considering the consecutive spheres as the trajectory of a single sphere obtained by tracing the position of the sphere inside a nucleus as displayed in Fig. 1 *A*. The motion of the sphere is assumed to follow the Brownian motion, wherein a particle moves randomly in a fluid as a result of collisions with other molecules. The stochastic nature of genomic locations indicates that individual genomic loci have different local conformations and variable spatial positions from cell to cell (21,31–34).

Segment activity

Each genomic segment in a chromosome is surrounded by other chromosomes or other portions of the chromosome. We define segment activity as the number of contacts between a segment and its surroundings measured by 3C-based experiments. Generally, a larger segment has a larger segment activity. Let us consider a large segment formed by a combination of n consecutive unit segments. We assume that the segment activity of the large segment, $F(n)$, is proportional to the mean-square end-to-end distance as

$$F(n) = k_1\langle|\mathbf{r}(n)|^2\rangle = k_1R(n)^2 \qquad (1)$$

where $\mathbf{r}(n)$ and $R(n)$ denote the end-to-end distance vector and the physical distance between the unit segments at both ends, respectively; the angle brackets denote an ensemble average over all genomic fragments in many cells; and k_1 is a proportional constant that depends on various factors such as the efficiency of the restriction enzyme, the number of paired sequences, the distribution of restriction sites, surface conformation, and other details of the experimental method. This assumption can be rationalized because the segment activity is proportional to the surface area of a segment.

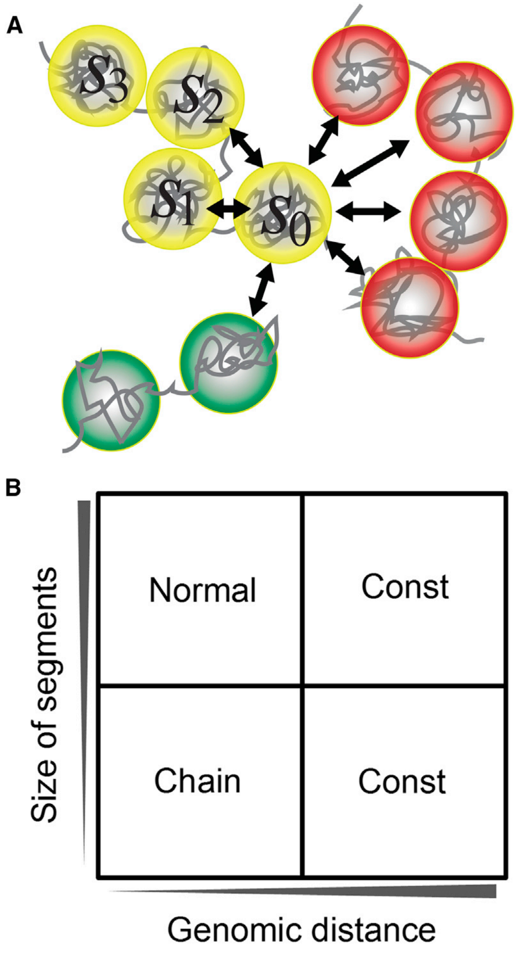

FIGURE 1 Schematic view of a subregion of the nuclear structure and three components involved in contact frequency. (*A*) Chromosomes are composed of equally divided genomic segments. Each segment is represented by a circle with a different color according to the chromosome. The segment s_0 contacts many surrounding segments (*double-sided arrows*). Similar to Brownian motion, consecutive segments (*yellow circles*) are regarded as a trajectory of the segment s_0 with time. (*B*) The contact frequency and the segment activity are composed of three components: normal, constriction, and chain-persistence components (denoted as *normal*, *const*, and *chain*, respectively). In contrast to the segment activity, which is mainly determined by the normal component, the contact frequency depends on the three components. For the contact frequency, the normal component contributes dominantly between the closely located segments with a large segment size, the chain-persistence component is dominant between the closely located segments with a small segment size, and the constriction component is dominant between the segments, which are separated. To see this figure in color, go online.

Mean-square end-to-end distance

Brownian motion can be described by the Langevin equation: a particle is decelerated by a frictional force in the direction opposite to its travel and is accelerated by a random force with a random direction (35). Similar to the Langevin equation, we assume that the position of a sphere in a chromosome adheres to the differential equation as

$$\frac{d^2}{dn^2}\mathbf{r}(n) = -\xi\frac{d}{dn}\mathbf{r}(n) + \mathbf{f}(n) \qquad (2)$$

where ξ and $\mathbf{r}(n)$ denote a frictional constant and the position of the nth sphere relative to its initial position, respectively. The second derivative of the position along the direction of the chain is proportional to the first derivative with a negative slope and to an external force, $\mathbf{f}(n)$, on the sphere. By multiplying each side of Eq. 2 by $\mathbf{r}(n)$ and by applying our knowledge of basic differential equations, we obtain

$$\frac{d^2}{dn^2}|\mathbf{r}(n)|^2 + \xi\frac{d}{dn}|\mathbf{r}(n)|^2 = 2\left|\frac{d}{dn}\mathbf{r}(n)\right|^2 + 2\mathbf{r}(n)\cdot\mathbf{f}(n) \quad (3)$$

By taking a statistical average, Eq. 3 becomes

$$\frac{d^2}{dn^2}R(n)^2 + \xi\frac{d}{dn}R(n)^2 = 2b^2 + 2\langle\mathbf{r}(n)\cdot\mathbf{f}(n)\rangle \quad \text{with } b = \left\langle\left|\frac{d}{dn}\mathbf{r}(n)\right|^2\right\rangle^{1/2} \quad (4)$$

where b denotes the average distance between nearest-neighbor spheres. From the definition of $\mathbf{r}(n)$, there are two boundary conditions: the mean-square end-to-end distance and its first derivative are equal to zero at the initial position (35). By applying the two boundary conditions and an additional assumption that the external force is randomly exerted, a solution to Eq. 4 can be obtained:

$$R(n)^2 = \frac{2b^2}{\xi}\left(n - \frac{1}{\xi}\left(1 - e^{-\xi n}\right)\right) \quad (5)$$

By comparing this result with the wormlike chain model, we find that the frictional constant is inversely proportional to the persistence length P as $\xi = b/P$, which coincides with the meaning of persistence length (36).

In the interphase nucleus, the end-to-end distance between spheres is reduced because chromosomes are compartmentalized and form chromosome territories (2,37). We include the effect of compartmentalization in the Langevin equation by introducing a first-order correction term of the external force function as

$$\langle\mathbf{r}(n)\cdot\mathbf{f}(n)\rangle = -\alpha n. \quad (6)$$

This equation indicates that a sphere experiences a large opposite force with a magnitude proportional to the contour length of the trajectory from the initial position. To fit the experimental data, it was sufficient to include only the first correction term. By applying Eq. 6 and the boundary condition in which the mean-square, end-to-end distance is equal to zero at the initial position, Eq. 4 becomes

$$R(n)^2 = \frac{1}{\xi}\left(2b^2 + \frac{2\alpha}{\xi}\right)n - \frac{\alpha}{\xi}n^2 + \frac{c}{\xi}\left(1 - e^{-\xi n}\right) \quad (7)$$

where c is an arbitrary constant that is not fixed because the slope of the distance is not equal to zero at the initial position. This equation shows that the mean-square end-to-end distance depends on three components:

1. A normal component, which is linearly proportional to the genomic length, describes the normal dependence of the end-to-end distance of the polymer chain length similar to the distance pattern appeared in the random-walk motion;
2. A constriction component, which is proportional to the square of the genomic length, originates from the compartmentalization effect and determines the long tail in the contact frequency curve; and
3. A chain-persistence component that relaxes the directionality in determining the next position.

Langevin equation for the chromosome conformation model

Based on the random motion of a sphere under the spatial constraint of compartmentalization, Eq. 7 was derived. This ideal model ignores several real properties of chromosomes such as the entropic repulsion between spheres and various chromatin-looping mechanisms. Such real properties have been considered by introducing a scaling exponent (23,36,38). Here, we also adjust the model by adding the scaling exponent v to Eq. 7 in an ad hoc manner, and we next apply the result to Eq. 1 to obtain the segment activity as a function of segment size:

$$F(n) = c_1 n^{2v} - c_2 n^{2v+1} + c_3\left(1 - e^{-\xi n}\right) \quad \text{with } c_1 = \frac{2k_1b^2 + 2k_1\alpha/\xi}{\xi},\ c_2 = \frac{k_1\alpha}{\xi},\ \text{and } c_3 = \frac{k_1 c}{\xi}. \quad (8)$$

This equation returns to Eq. 7 when a chromosome conformation follows the random-walk motion, where the scaling exponent is equal to 0.5. This model describes the relationship between the segment activity and the segment size by assuming that the segment activity is proportional to the segment's surface area. Due to the use of the scaling exponent, the explicit relationship between the proportional constants in Eq. 8 may change.

Contact frequency from segment activity function

In general, segment activity increases with increasing segment size. An equation is derived to convert the segment activity function into the contact frequency. Let us assume that there are three consecutive segments, s_i, s_j, and s_k, with sizes n_i, n_j, and n_k, respectively. The contact frequency $C(n_j;s_i,s_k)$ between two segments s_i and s_k that are separated by a genomic distance n_j can be derived from the segment activity function by the following methods.

The contact frequency between two neighboring segments s_i and s_j can be described as

$$F\left(n_i + n_j\right) = F(n_i) + F\left(n_j\right) - 2C\left(0; s_i, s_j\right) \quad (9)$$

where the contact frequency between two individual segments reduces the segment activity of a combined segment, $F(n_i + n_j)$, compared to a sum of individual segment activities. This equation can be easily extended to evaluate the contact frequency with a certain genomic gap:

$$F\left(n_i + n_j + n_k\right) = F(n_i) + F\left(n_j\right) + F(n_k) - 2C\left(0; s_i, s_j\right) - 2C\left(0; s_j, s_k\right) - 2C\left(n_j; s_i, s_k\right) \quad (10)$$

By combining Eqs. 9 and 10, the contact frequency for two segments with a certain genomic gap is derived as

$$C\left(n_j; s_i, s_k\right) = \frac{F\left(n_i + n_j\right) + F\left(n_j + n_k\right) - F\left(n_j\right) - F\left(n_i + n_j + n_k\right)}{2} \quad (11)$$

This conversion from a segment activity to a contact frequency function can be further clarified by substituting n_j with the average genomic distance between two segments s_i and s_k as

$$\overline{C}(g; s_i, s_k) = \frac{\left[F\left(g-\frac{n_i-n_k}{2}\right)+F\left(g+\frac{n_i-n_k}{2}\right)-F\left(g-\frac{n_i+n_k}{2}\right)-F\left(g+\frac{n_i+n_k}{2}\right)\right]}{2} \cong -\frac{n_i n_k}{2}F''(g), \ \ with \ \ g = n_j+\frac{(n_i+n_k)}{2}, \tag{12}$$

where g denotes the average genomic distance between two segments. From Eq. 12, we find that the contact frequency approaches to a quantity including a second derivative of the segment activity function as the genomic distance increases. This equation shows that the contact frequency is determined not only by the magnitude but also by the curvature of the segment activity curve.

Chromosome conformation capture

The chromosome conformation capture data for the human and fission yeast genomes were obtained from other studies (14,15). Kalhor et al. (15) reported 3C-based data for the human genome by applying an improved technique and suggested a population-based modeling method to build the genomewide chromosome conformation. Tanizawa et al. (14) studied the fission yeast genome and found long-range associations between genomic loci using a 3C-based technique.

The 3C-based data used in this article are available from the National Center for Biotechnology Information (NCBI, National Institutes of Health, Bethesda, MD) Sequence Read Archive (SRA) under submission No. SRA025848 for the human genome and No. SRA020835 for the fission yeast genome. There are ~49,000,000 paired reads for the human genome and 79,000,000 paired reads for the yeast genome.

In this work, the paired-end reads of the human genome were filtered in the following steps:

1. For each paired read, sequences from both sides were separately mapped to the genome (hg19) using the routine BOWTIE (39) with the option to allow fewer than three mismatches; the paired read was discarded if the genomic distance between two locations was <30 kb.
2. There are 825,083 *Hin*dIII restriction sites in the reference human genome if we regard two restriction sites as a single site when they are closer than 20 bp. The genomic locations obtained from the previous step were translated into new coordinates composed of the closest *Hin*dIII restriction site and the offset from that site. We discarded paired reads for which the offset was >100 bp.
3. The human genome was divided into 57,083 segments; each segment covered 50 kb of the genomic length. Because each restriction site belongs to a particular segment, each paired read gives contact information between two segments.

The fission yeast genome has 6932 *Hin*dIII restriction sites. The genome was divided into 608 segments, with each segment covering 20 kb. The paired reads for the fission yeast genome were filtered by following the above-described method for the human genome with a few differences.

Paired reads were discarded in the following cases:

1. The sequence of a paired read was aligned to multiple positions in the genome;
2. The offset from the restriction site was not equal to zero; or
3. The genomic distance between two sequences was <20 kb.

After applying the filtering method, 5,800,000 and 1,800,000 paired reads remained for the human and fission yeast genomes, respectively.

Contact frequency and segment activity

The genomes were divided into many labeled segments for error reduction and computational efficiency. The contact frequencies between segments were evaluated by counting the paired reads of sequences from 3C-based experimental data; a higher contact frequency value indicates that two segments are closer in the nucleus.

Additionally, we suggest the term "segment activity" to denote the total number of contacts between a genomic segment and all of its surroundings. The segment activity and contact frequency are closely related to each other.

Data fitting

Mean segment activities and mean contact frequencies were obtained from the experimental data. The mean segment activities were fitted using the NLFit module in the software ORIGIN (OriginLab, Northampton, MA) with the first two components in Eq. 8. The chain-persistence contribution was fitted using the mean contact frequencies because its contribution to the mean segment activity was very small whereas the contribution was large for the mean contact frequencies.

To fit the chain-persistence component, we used the following method:

1. The normal and constriction components in the mean segment activities were converted into contact frequencies using Eq. 12 and were subtracted from the mean contact frequencies.
2. The remainder after subtraction was fitted using the functional form of the chain-persistence component.
3. The contribution of the chain-persistence component to the mean contact frequencies was dominant for smaller segment sizes; thus, we used a segment size of 50 kb for the human genome and 20 kb for the fission yeast genome to fit the parameters for the chain-persistence contribution.
4. The mean contact frequency between nearest-neighbor segments on the DNA chain was dominated by the screening criterion for paired reads; thus, the contact frequency between nearest-neighbor segments was not used to fit the mean contact frequency curve.
5. When fitting the mean activity curves, we found a discontinuous function that could not be explained by the three components. The discontinuous function was linear with respect to genomic length, was small in magnitude, and did not contribute to the mean contact frequency curve; we therefore ignored the function during further analysis.

Physical distance from contact frequency

To verify the relationship between the segment activity and physical distance, we evaluated physical distances directly from the contact frequency map of the fission yeast genome using the Langevin equation for the chromosome conformation (LECC) model and compared the evaluated distances with experimental data. A contact frequency map is obtained from the yeast genome based on 20-kb segment size, and then normalized by the iterative correction method to adjust the map to have the equal segment activity for each genomic segment (30). Physical distances between several genomic loci in interphase measured by FISH experiment were obtained from the Tanizawa et al. (14). To obtain a contact frequency between each pair of genomic loci, we assign a block of genomic segments used in the frequency map into each genomic locus, where the segments contain or partly contain the genomic locus. For each pair of genomic loci, we evaluated an average contact frequency over all possible pairs between the blocks of genomic segments. The obtained contact frequencies were translated into physical distances by following steps:

1. The contact frequencies are converted into genomic distances using Eq. 12;
2. The genomic distances are converted into segment activities using Eq. 8; and
3. The segment activities are converted into physical distances by using

$$R(n) = R(H)\sqrt{\frac{F(n)}{F(H)}}, \tag{13}$$

where H denotes a unit separation between genomic segments.

Here, we use 20 kb as the unit separation, which corresponds to the physical distance between the nearest genomic segments along single chromatids.

RESULTS

Mean segment activity

The human and fission yeast genomes are sectored into many sequentially labeled segments for analysis. Each segment is in contact with its surroundings, including other segments, nucleoli, the nuclear membrane, etc. Segment activities are evaluated by counting the total number of contacts of a segment with surrounding genomic segments. Thus, a segment with a larger size has a larger segment activity on average. In Fig. 2, the mean segment activities are plotted according to the size of the segment. The segment activity curve shows a simple and smooth increase with the size.

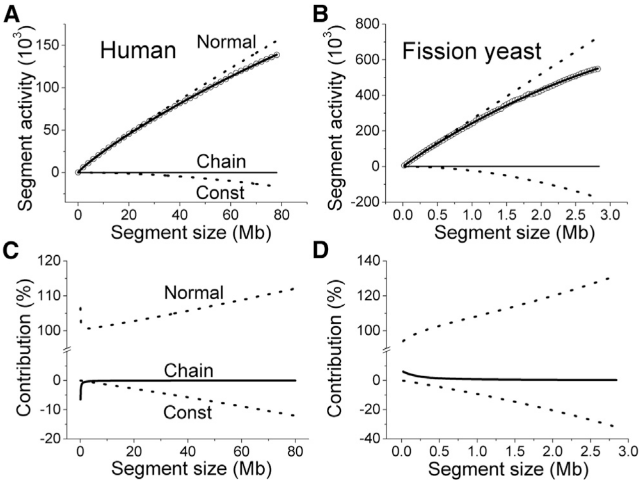

FIGURE 2 Mean segment activity curve and its components. Mean segment activities from experimental data are plotted (*open circles*), and the fitting curves are plotted (*solid lines*) for the human (*A*) and fission yeast (*B*) genomes. The three components of the mean segment activity curve are plotted separately according to segment size for the human (*C*) and fission yeast (*D*) genomes. The contribution shown is the percent contribution of each component to the mean segment activity for a certain segment size. Experimental data are obtained from Tanizawa et al. (14) and Kalhor et al. (15).

The mean segment activity curves were fitted with a coefficient of determination (R^2) of 1.000 for both the human genome and the fission yeast genome. The magnitude of the segment activity is mainly determined by the normal component. The constriction component reduces the segment activity, and the magnitude of the component increases as the segment size increases. Overall, segment activity increases with segment size, but its second derivative is negative; this observation coincides with the fact that the physical distance between segments becomes saturated as the genomic distance increases (23,38). This negative value for the second derivative is related to the compartmentalization: a segment with a larger size has a greater probability of experiencing self-contact and contact with the nuclear periphery, reducing segment activity (41).

The scaling exponent in Eq. 8 is a parameter to describe the folding state of the chromosomes. A more compact structure has a smaller scaling exponent. For example, the scaling exponent is 0.333 for a globular state and 0.5 for a random-walk polymer. We obtained scaling exponents of 0.443 and 0.483 for the human genome and fission yeast genome, respectively. From these values, we find that both genomes form a more compact structure in comparison to the random-walk structure, and the degree of compaction is larger in the human genome than in the fission yeast genome.

Mean contact frequency

The contact frequency between two segments decreases as their genomic distance increases. Mean contact frequencies are obtained from the experimental data and plotted as circles in Fig. 3, where a monotonically decreasing pattern is shown, as expected. The mean contact frequency curves are derived analytically from mean segment activity curves using Eq. 12. Pearson's correlation coefficients are 0.9996 for 50-kb segment size of the human genome and 0.9994 for 20-kb segment size of the yeast genome. Because the segment activity curve consists of three components, the mean contact frequency curve is also composed of three components: normal, constriction, and chain-persistence. As the genomic distance increases, both the normal and the chain-persistence components decrease whereas the constriction component increases, as illustrated in Fig. 1 *B*.

In Table 1, the sign of the proportional parameter c_3 is positive for the fission yeast genome whereas it is negative for the human genome, indicating that the chain-persistence component reduces the contact frequency for the human genome, although it increases the contact frequency for the yeast genome, as shown in Fig. 3. The number of monomers per persistence length is obtained by fitting the chain-persistence portion in the mean contact frequency curves. The fission yeast genome has a persistence length approximately twofold longer than that of the human genome.

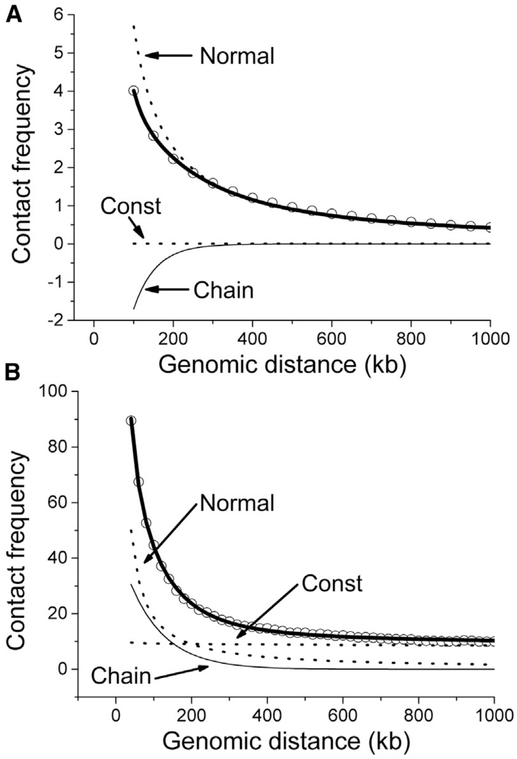

FIGURE 3 Mean contact frequency curve and its components. Mean contact frequencies from experimental data are plotted (*open circles*) for the human (*A*) and fission yeast (*B*) genomes. The segment sizes are 50 kb for the human genome and 20 kb for the fission yeast genome. Mean contact frequency curves are evaluated from mean segment activity curves and are plotted (*thick solid lines*). Similar to the mean segment activity curve, a mean contact frequency curve is composed of three components. (*Thin solid lines*) Chain-persistence components; (*dotted lines*) the other components. Here, we did not plot the nearest-neighbor points because they included a large error that originated from experimental limitation.

Physical distance from contact frequency

In the LECC model, we assume that a segment activity is proportional to the mean-square end-to-end distance and suggest an equation for the relationship between the segment activity and the contact frequency. To confirm the LECC model, physical distances from experimental data are compared to the corresponding distances derived from contact frequencies. In Fig. 4, the two quantities show a good correlation with 0.902 of Pearson's correlation coefficient. From the slope of the fitting line, a 20-kb genomic sphere requires 320–400 bp for 1-nm spatial movement in the fission yeast genome, which is three times larger than that for the budding yeast genome (110–150 bp/nm) (42).

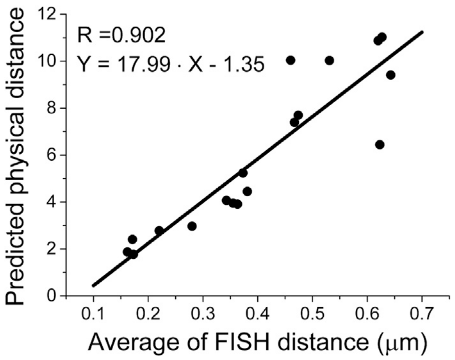

FIGURE 4 Physical distance and contact frequency. Experimental data on the physical distances between pairs of genomic loci were obtained from Tanizawa et al. (14). The physical distances corresponding to the experimental data are derived from contact frequencies and plotted (*solid circles*) according to the experimental data. The unit of the vertical axis is the average length of a 20-kb genomic segment. A linear fitting line is shown (*solid line*).

Segment size effect

For various segment sizes, we evaluate the mean contact frequencies from the experimental data and then derive the same quantities directly from the mean segment activity curves using Eq. 12. For each genome, the parameters obtained by fitting both the mean segment activity curve and the mean contact frequency curve with the minimum segment size were used to generate other mean contact frequency curves with larger segment sizes. In Fig. 5, the mean contact frequencies from the experimental data are plotted as open circles and those derived from the mean segment activity curves are depicted as solid lines for the human and yeast genomes. Pearson's correlation coefficients range from 0.9981 to 0.9994 for the human genome and range from 0.9933 to 0.9991 for the yeast genome. The conversion accurately predicted the experimental data, indicating the validity of the LECC model. The mean contact frequencies between the nearest neighbors are consistently lower than

TABLE 1 Parameters used in fitting the mean segment activity curves

Genome	Cut[a] (kb)	v	P[c] (= $1/\xi$, kb)	c_2[d] (1×10^{-6}/kb)	c_3[d] (kb)	$\overline{R}^2$ [e]
Human	30	0.4428 ± 0.0006[b]	55.58 ± 2.36	1.35 ± 0.06	−3.29 ± 0.16	0.99991 ± 0.00001
	50	0.4429 ± 0.0007	55.16 ± 3.22	1.35 ± 0.03	−3.26 ± 0.21	0.99991 ± 0.00001
Yeast	20	0.4836 ± 0.0016	93.03 ± 4.20	85.7 ± 1.4	6.17 ± 0.94	0.99977 ± 0.00003
	50	0.4857 ± 0.0011	98.87 ± 4.70	87.0 ± 0.7	7.61 ± 0.84	0.99977 ± 0.00003

[a] Denotes the sequence separation criterion for removing paired reads.

[b] Denotes the number of monomers per persistence length, which is one-half of the Kuhn length.

[c] Denotes the standard deviation, which was evaluated from the parameters obtained from five equally divided experimental data sets.

[d] Denotes fitting parameters for Eq. 8 normalized by c_1.

[e] Denotes adjusted R-square values.

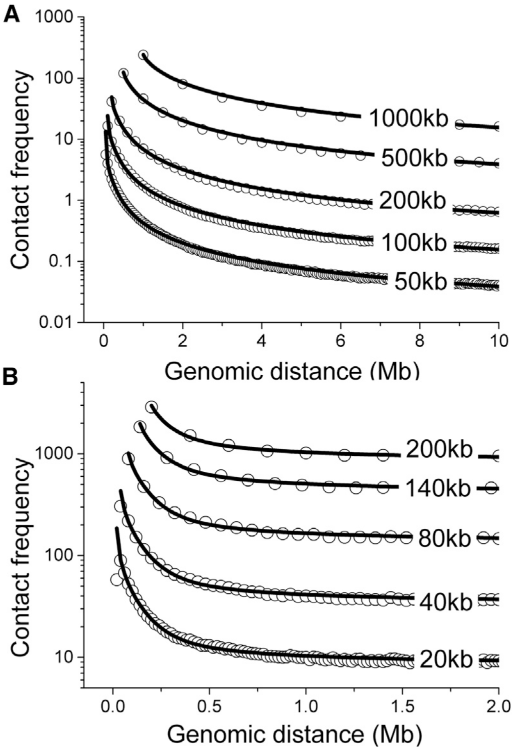

FIGURE 5 Mean contact frequency curves according to segment size. Mean contact frequencies are calculated from experimental data according to the segment size and are plotted (*open circles*) for the human (*A*) and fission yeast (*B*) genomes. The mean contact frequency curves for various segment sizes are extracted from a mean segment activity curve for each genome and plotted (*solid lines*). Segment sizes are denoted on each line.

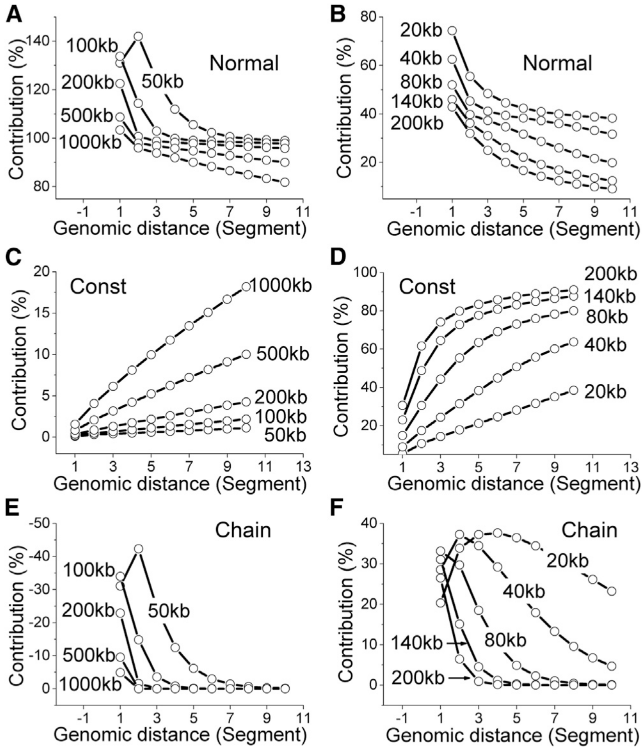

FIGURE 6 Contribution of the three components to the mean contact frequency curves according to segment size. For various segment sizes, the contributions of the components of the mean contact frequency curves are plotted (*open circles*). (*A*, *C*, and *E*) Component plots of the human genome; (*B*, *D*, and *F*) component plots of the fission yeast genome.

the expected value due to the removal of paired reads with a genomic distance cutoff during the course of the preparation of the contact frequencies from the experimental data.

For various segment sizes, the relative contributions of the three components to the mean contact frequency curves are shown in Fig. 6. For the segment sizes >500 kb in the human genome, the overall shape of the mean contact frequency curve is primarily determined by the normal component, and the overall contribution of the other components is <25%. For the sizes <50 kb, the contribution of the chain-persistence component is comparable to the contribution of the normal component because of the large curvature of the chain-persistence curve. For the size of 200 kb, the contribution of the constriction component in the human genome is <5% whereas the contribution in the fission yeast genome is ~90%. The rapid growth of the constriction component in the fission yeast genome is related to both the small nucleus size and the large surface contact between the chromosome and the nuclear periphery.

DISCUSSION

In this study, we suggested the use of the LECC model to help us understand the physical origin of the contact frequency. The primary assumption of the model is that the total number of contacts between a target segment and all of the surrounding segments is linearly proportional to the exposed surface area of the segment. In the model, we regarded a chromosome as a polymer composed of consecutive segments, and altered our representation of consecutive segments into a trajectory of a segment over time. We successfully derived this model from the Langevin equation, which we then applied to derive the physical distance between segments. An exact equation for converting the mean segment activity into a mean contact frequency was also presented. The model accurately fit both the mean contact frequencies and the mean segment activities obtained from chromosome conformation capture data. From the model, we established the relationship between the contact frequency and the physical distance between genomic loci.

The scaling exponent used in Eq. 8 is determined by the folding properties in the polymer models. Using the random loop model, Mateos-Langerak et al. (23) demonstrated that the scaling exponent is related to the looping probability such that more frequent loops between genomic loci reduce the scaling exponent. We obtain a smaller scaling exponent for the human genome than for the fission yeast genome, indicating that the human genome forms more loops than the fission yeast genome. We also obtain Kuhn lengths for both genomes from the chain-persistence component. The number of monomers per Kuhn length in the human genome is 110 kb, which is fivefold larger than that of the chromatin fibers reported in a previous study (which ranged from 14 to 46 kb) (43). Because the Kuhn length obtained in our model is related to the spatial movement of the genomic segments, the length is different from the ordinary Kuhn length of the chromatin fibers. The chromatin fibers of mice at meiotic prophase were found to form small loops of <200 kb (44). The measure 110 kb is close to the size of chromatin loops in interphase chromosomes and is also a low-frequency rhythm of GC content along the genome (45–47). The loop size was used to simulate human chromosomal characteristics in the MLS model (27). The 110-kb genomic length from the chain-persistence component lowers the contact frequency of the human genome compared to the overall contact frequency pattern, supporting existence of small loops in the MLS model in the interphase chromosome because contact probability between genomic loci in the same loop will be low. The Kuhn length of the yeast genome is estimated to be twofold longer than that of the human genome.

In the reported biphasic relationship of physical distance between genomic loci, the mean-square end-to-end distance between segments of G_0/G_1-phase human nuclei followed the random coil model in the range from zero to 2.0 Mb with the scaling exponent close to 0.5; the distance followed the fractal globule model in the region >10 Mb with an exponent of ~0.33 (25,26). In our model, both the low contribution of the constriction component and the large contribution of the chain-persistence component increase the scaling exponent of the human genome to approach the random-walk model at a small genomic distance region, whereas the large contribution of the constriction component at the large genomic distance region decreases the scaling exponent to approach the fractal globule model. Another characteristic of chromosome conformation is the relationship between the contact frequency and the genomic distance. The scaling exponent of the relationship was experimentally measured to be -1.08 (11,25). We find that the relationship between the contact frequency and genomic distance changes according to the used segment size. For the 1-Mb segment size, the dominant component for the mean contact frequency curve in the human genome is the normal component. Based on the equation stating that the mean contact frequency can be approximated to the second derivative of the mean segment activity function, the scaling exponent is approximately equal to $2\nu - 2$ (-1.1 for the human genome), which coincides with the experimental value.

The contribution of the chain-persistence component to the contact frequencies increases for smaller segment sizes. This contribution is comparable to that of the normal component for small segment sizes. In the wormlike chain model, the contribution always decreases the mean-square end-to-end distance, which reduces the contact frequencies, because the direction of the contribution is fixed by the additional boundary condition requiring that the first derivative of the square distance be equal to zero at an initial position (36). However, in our model, the direction of the contribution varies with species. The component in the yeast genome increases the contact frequency whereas this component decreases the contact frequency in the human genome.

In our model, each segment in a chromosome experiences a force that draws both ends of the segment closer together. This force shortens the physical distance between segments with a long genomic distance and is assumed to be caused by chromosome compartmentalization. The proportional constant representing the force in the human genome is orders-of-magnitude lower than in the fission yeast genome, which coincides with the fact that the segments in the fission yeast genome are confined to a smaller space than those in the human genome (a nucleus radius of 5 μm for the human genome (15) and 0.71 μm for fission yeast (14)).

We compared experimentally measured physical distances between genomic loci with the distances derived from contact frequencies using the LECC model. Two quantities show a good correlation, indicating that the LECC model properly captures the characteristics of chromosome conformation. The line density of the fission yeast genome is three times larger than that of the budding yeast genome.

There are two possible reasons for the difference:

1. Chromosomes of the fission yeast are confined in a small nucleus compared to the budding yeast, and
2. We used 20-kb genomic distance for measuring the density, which increases the line density of the fission yeast chromatids because average density increases as the genomic length increases.

The agreement of the LECC model with the experimental results supports the hypothesis that the physical concept used in this model is valid. In the large-scale chromosomal organization, we found that the mean contact frequency curve consists of three components: normal, constriction, and chain-persistence. We suggested a physical origin for each component. The normal component is related to the normal folding characteristics of chromatin and is similar to a random-walk motion without an external force. The constriction component arises from compartmentalization and reduces the physical distance between the segments. The chain-persistence component is related to the properties of the contact between closely located segments and formation of small chromatin loops.

Furthermore, we suggested a method for converting the segment activities into contact frequencies. The conversion was accomplished using an exact equation, which was applied to evaluate contact frequencies between segments of various sizes. The model successfully explained the experimentally verified relations: the biphasic relation of the mean-square end-to-end distance, the segment activity, and the contact frequency as a function of genomic distance (11,26).

Additionally, the physical distances derived from contact frequencies have a good correlation with those from experiment. Interestingly, the LECC model was developed based on the assumption that the contact frequency is proportional to the contact surface area, not the contact volume, which was suggested in the fractal globule model (25). Understanding the physical origins of the contact frequency will provide a rational way to normalize the contact frequency map, which determines both the quality of long-range interactions revealed by chromosome conformation capture data and the quality of further analysis for assembling a three-dimensional structure of the genome inside a nucleus.

This work was supported by the Stem Cell Research Program (grant No. 2012M3A9B4027957), the Korea Healthcare Technology R&D Project, Ministry for Health, Welfare & Family Affairs, Republic of Korea (grant No. A092006), and the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (grant No. NRF-2010-0012649).

REFERENCES

1. Cremer, M., J. von Hase, …, T. Cremer. 2001. Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome Res.* 9:541–567.
2. Cremer, T., and C. Cremer. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* 2:292–301.
3. Tanabe, H., F. A. Habermann, …, T. Cremer. 2002. Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications. *Mutat. Res.* 504:37–45.
4. Fussner, E., R. W. Ching, and D. P. Bazett-Jones. 2011. Living without 30 nm chromatin fibers. *Trends Biochem. Sci.* 36:1–6.
5. Luger, K., M. L. Dechassa, and D. J. Tremethick. 2012. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell Biol.* 13:436–447.
6. Woodcock, C. L., and R. P. Ghosh. 2010. Chromatin higher-order structure and dynamics. *Cold Spring Harb. Perspect. Biol.* 2:a000596.
7. Joti, Y., T. Hikima, …, K. Maeshima. 2012. Chromosomes without a 30-nm chromatin fiber. *Nucleus.* 3:404–410.
8. Maeshima, K., S. Hihara, and M. Eltsov. 2010. Chromatin structure: does the 30-nm fiber exist in vivo? *Curr. Opin. Cell Biol.* 22:291–297.
9. Schalch, T., S. Duda, …, T. J. Richmond. 2005. X-ray structure of a tetranucleosome and its implications for the chromatin fiber. *Nature.* 436:138–141.
10. Eltsov, M., K. M. MacLellan, …, J. Dubochet. 2008. Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. *Proc. Natl. Acad. Sci. USA.* 105:19732–19737.
11. Lieberman-Aiden, E., N. L. van Berkum, …, J. Dekker. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 326:289–293.
12. Duan, Z., M. Andronescu, …, W. S. Noble. 2010. A three-dimensional model of the yeast genome. *Nature.* 465:363–367.
13. Sexton, T., E. Yaffe, …, G. Cavalli. 2012. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell.* 148:458–472.
14. Tanizawa, H., O. Iwasaki, …, K. Noma. 2010. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.* 38:8164–8177.
15. Kalhor, R., H. Tjong, …, L. Chen. 2012. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* 30:90–98.
16. Dekker, J., K. Rippe, …, N. Kleckner. 2002. Capturing chromosome conformation. *Science.* 295:1306–1311.
17. Zhao, Z., G. Tavoosidana, …, R. Ohlsson. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* 38:1341–1347.
18. Tolhuis, B., R. J. Palstra, …, W. de Laat. 2002. Looping and interaction between hypersensitive sites in the active β-globin locus. *Mol. Cell.* 10:1453–1465.
19. Dostie, J., T. A. Richmond, …, J. Dekker. 2006. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 16:1299–1309.
20. Sajan, S. A., and R. D. Hawkins. 2012. Methods for identifying higher-order chromatin structure. *Annu. Rev. Genomics Hum. Genet.* 13: 59–82.
21. van Steensel, B., and J. Dekker. 2010. Genomics tools for unraveling chromosome architecture. *Nat. Biotechnol.* 28:1089–1095.
22. Hu, M., K. Deng, …, J. S. Liu. 2012. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics.* 28:3131–3133.
23. Mateos-Langerak, J., M. Bohn, …, S. Goetze. 2009. Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl. Acad. Sci. USA.* 106:3812–3817.
24. Hahnfeldt, P., J. E. Hearst, …, L. R. Hlatky. 1993. Polymer models for interphase chromosomes. *Proc. Natl. Acad. Sci. USA.* 90:7854–7858.
25. Mirny, L. A. 2011. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* 19:37–51.
26. Yokota, H., G. van den Engh, …, B. J. Trask. 1995. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G_0/G_1 interphase nucleus. *J. Cell Biol.* 130:1239–1249.
27. Munkel, C., and J. Langowski. 1998. Chromosome structure predicted by a polymer model. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics.* 57:5888–5896.
28. Gehlen, L. R., G. Gruenert, …, J. M. O'Sullivan. 2012. Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions. *Nucleus.* 3:370–383.
29. Yaffe, E., and A. Tanay. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* 43:1059–1065.
30. Imakaev, M., G. Fudenberg, …, L. A. Mirny. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods.* 9:999–1003.
31. Bolzer, A., G. Kreth, …, T. Cremer. 2005. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* 3:e157.
32. Misteli, T., and E. Soutoglou. 2009. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat. Rev. Mol. Cell Biol.* 10:243–254.
33. Teller, K., I. Solovei, …, T. Cremer. 2007. Maintenance of imprinting and nuclear architecture in cycling cells. *Proc. Natl. Acad. Sci. USA.* 104:14970–14975.
34. Strickfaden, H., A. Zunhammer, …, T. Cremer. 2010. 4D chromatin dynamics in cycling cells: Theodor Boveri's hypotheses revisited. *Nucleus.* 1:284–297.
35. Hansen, J. P., and I. R. McDonald. 1990. Theory of Simple Liquids. Academic Press, San Diego, CA.
36. Bloomfield, V. A., D. M. Crothers, and I. Tinoco, Jr. 1974. Physical Chemistry of Nucleic Acids. Harper & Row, New York.

37. Albiez, H., M. Cremer, …, T. Cremer. 2006. Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. *Chromosome Res.* 14:707–733.

38. Tark-Dame, M., R. van Driel, and D. W. Heermann. 2011. Chromatin folding—from biology to polymer models and back. *J. Cell Sci.* 124:839–845.

39. Langmead, B., C. Trapnell, …, S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

40. Reference deleted in proof.

41. Meldi, L., and J. H. Brickner. 2011. Compartmentalization of the nucleus. *Trends Cell Biol.* 21:701–708.

42. Bystricky, K., P. Heun, …, S. M. Gasser. 2004. Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. *Proc. Natl. Acad. Sci. USA.* 101: 16495–16500.

43. Rippe, K. 2001. Making contacts on a nucleic acid polymer. *Trends Biochem. Sci.* 26:733–740.

44. Heng, H. H., J. W. Chamberlain, …, P. B. Moens. 1996. Regulation of meiotic chromatin loop size by chromosomal position. *Proc. Natl. Acad. Sci. USA.* 93:2795–2800.

45. Laemmli, U. K., E. Käs, …, Y. Adachi. 1992. Scaffold-associated regions: *cis*-acting determinants of chromatin structural loops and functional domains. *Curr. Opin. Genet. Dev.* 2:275–285.

46. Arneodo, A., C. Vaillant, …, C. Thermes. 2011. Multi-scale coding of genomic information: from DNA sequence to genome structure and function. *Phys. Rep. Rev. Phys. Lett.* 498:45–188.

47. Nicolay, S., F. Argoul, …, A. Arneodo. 2004. Low frequency rhythms in human DNA sequences: a key to the organization of gene location and orientation? *Phys. Rev. Lett.* 93:108101.