Chromosome evolution with naked eye: Palindromic context of the life origin

Sergei Larionov, Alexander Loskutov, and Eugeny Ryadchenko
Physics Faculty, Moscow State University, Moscow 119899, Russia

(Received 4 July 2007; accepted 28 November 2007; published online 1 February 2008)

Based on the representation of the DNA sequence as a two-dimensional (2D) plane walk, we consider the problem of identification and comparison of functional and structural organizations of chromosomes of different organisms. According to the characteristic design of 2D walks we identify telomere sites, palindromes of various sizes and complexity, areas of ribosomal RNA, transposons, as well as diverse satellite sequences. As an interesting result of the application of the 2D walk method, a new duplicated gigantic palindrome in the X human chromosome is detected. A schematic mechanism leading to the formation of such a duplicated palindrome is proposed. Analysis of a large number of the different genomes shows that some chromosomes (or their fragments) of various species appear as imperfect gigantic palindromes, which are disintegrated by many inversions and the mutation drift on different scales. A spread occurrence of these types of sequences in the numerous chromosomes allows us to develop a new insight of some accepted points of the genome evolution in the prebiotic phase. © *2008 American Institute of Physics*.
[DOI: 10.1063/1.2826631]

"When we have shuffled off this mortal coil," William Shakespeare. **As is known, DNA consists of four nucleotides. This fact allows us to sketch out the DNA sequence in a plane and present the nucleotide sequence as a 2D plane walk. By means of this approach we may portray compactly the entire DNA sequence of chromosome, even if it includes tens and hundreds of millions of nucleotides. Owing to this representation, the nature of the sequence organization becomes evident and chromosome images acquire unique "portrait" properties. This method significantly amends and simplifies the process of characterization of quite large functionally essential sites in chromosomes. In addition, we employ the 2D walk method as an interface of genomes' databases. It makes it possible to analyze a wide range of problems: from protein clusters prediction and metabolical network organization to a evolutionary modeling. The proposed analysis allowed us to find a new duplicated giant palindrome in the X human chromosome and advance a schematic mechanism of the appearance of such structures. Of special interest in our study are imperfect gigantic palindromes (up to several tens megabases), which have wide range of ages and functions in different genomes. These palindromes have disintegrated during the evolutionary process by inversion, mutation drift, and other kinds of rearrangement. Abundance of imperfect gigantic palindromes in different species points to the evolutionary significance of such a type of sequences. We suppose that the strategy of the complementary duplication, which has led to the formation of imperfect gigantic palindromes, may have an ancient origin because it is based on the main DNA property; i.e., complementarity. Using the 2D maps of the human chromosomes, we have analyzed their regions that surround the experimentally obtained sites of the replication initiation (so-called replicons). A certain structural resemblance of these sites to imperfect gigantic palindromes is found. It is known that the majority of bacterial genomes that have a unique site of the replication origin, exhibit a composition asymmetry. In the obtained images, the existence of a cooperative composition is shown. Therefore, the mutation process of close bacterial genomes qualitatively is presented like an α-degree rotation of the 2D DNA map. The obtained results display that some of imperfect gigantic palindromes may relate to the replication subsystem also. Certain ideas of a possible role of gigantic palindromes in early genome evolution are discussed in terms of a spin-glass model.**

I. INTRODUCTION

DNA is a macromolecular complex in the form of a double helix consisting of two strands of nucleotides that are connected via hydrogen bonds. Nucleotides are low-molecular compounds that consist of nitrogen bases (purines and pyrimidines), carbohydrates (ribose or deoxyribose), and a phosphate group. The DNA molecule contains two different purines; namely, adenine (A) and guanine (G), as well as two pyrimidines, cytosine (C) and thymine (T). Each pair of nucleotides on opposite complementary strands is associated by hydrogen bonds: a guanine-cytosine pair, by three hydrogen bonds; an adenine-thymine, by two bonds. The phosphate groups run along the outside, while nitrogen bases run inside, so that their planes are perpendicular to the axis of the molecule. Each branch of the helix consists of nucleotide units linked together to form a long polynucleotide strand, which is conventionally represented as a nucleotide alphabet *ATTGCCAA*…, and is considered as the DNA sequence. A double-strand molecule of DNA linked with some proteins and organized in a certain hierarchical manner forms a chromosome.[1]

The term *genome* is used for the complete set of the whole-cell DNA; i.e., the complete sequence of nucleotides. It is conventionally assumed that the main function of the DNA is to process, carry out, and reproduce information, as well as to adapt to a dynamic environment by means of evolution. Moreover, these processes should operate on the basis of the information carried by the same sequence; this imposes specific restrictions on the organization of DNAs.

In this paper, we develop a method that makes it possible to present the whole chromosome in a compact form, even if it includes tens and hundreds of millions of nucleotides (i.e., million bases, Mb), find easily similar fragments, identify functional and structural elements, and detect the self-similarity of some fragments of the DNA sequences. The method is based on the representation of DNAs as a 2D walk of a particle. In this sense, the nucleotide sequence of chromosomes may be easily recognized so that their images have the unique features. This method significantly amends and simplifies the process of characterization of quite large functionally essential sites in chromosomes. In particular, a duplicated gigantic palindrome in the X human chromosome is found and a certain schematic mechanism of its formation is described.

By this method, we have found that some chromosomes or their fragments appear as gigantic palindromes. These palindromes are disintegrated by many inversions and the mutation drift on different scales forming so-called imperfect gigantic palindromes (IGPs). We observe IGPs in genomes of different species and suppose that such structures are a widespread phenomenon.

We also found the similarity of the detected IGPs to certain experimentally determined sites of the replications in different chromosomes. Considering IGPs as a part of the genome evolutionary strategy, a new phenomenology concerning the prebiotic phase is discussed.

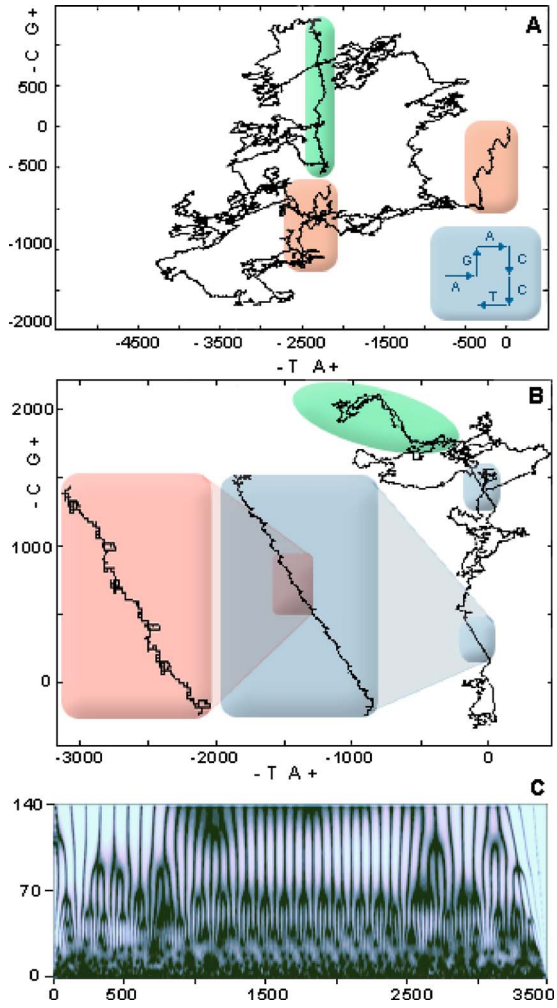

FIG. 1. (Color online) (A) Chromosome 12 of *S. cerevisiae*: Telomeres (small rectangles) and a small part (two units from 150) of the cluster of ribosomal RNAs (prolate rectangle). (B) Chromosome 1 of *S. cerevisiae* (about 230 Kb) containing an IGP (ellipse). (C) The wavelet-transformation of the region 24.5–28 Kb representing the family of flocculation genes FLO1 and FLO9 in the subtelomeric region (see rectangles in B).

II. GENOME AS A 2D DNA WALK

The idea of the correspondence of nucleotide types to coordinate vectors on a complex plane was first suggested in 1962 by Golomb.[2] More recently, in 1985, Gates,[3] and Mizraji and Ninio[4] also mapped some small DNA chains on a plane. Selection of *A-T* and *G-C* coordinate axes, respectively, was dictated by the complementarity of chains and hydrogen bond balance. But as these sequences were very small, they looked pseudorandom, and this type of the sequence representation did not attract specific attention of biologic community. Later, Berthelsen *et al.*,[5] Nandy,[6] and some other authors discussed different forms of graphic representations of DNA. In Refs. 7 and 8 (see also references therein), certain fractal and statistical properties of the 2D DNA walk have been considered.

On the basis of formal treatment of a slightly changed the 2D walk, Vincens *et al.*[9] first developed an algorithm of finding regions of similarity in complete genome sequences. However, characteristic images were beyond the scope of their analysis. As we show below these images are significantly more informative on the whole chromosome level, and our 2D representation provides an insight into the global sequence structure and its properties. Some preliminary ideas of our approach have been published in Ref. 10.

Consider a DNA nucleotide sequence as a walk on a square lattice: Starting from the origin (0, 0) we make a step right for adenine (*A*), a step left for thymine (*T*), a step up for guanine (*G*), a step down for cytosine (*C*). The original sequence of nucleotides is then mapped onto a certain trajectory on the plane with *A-T* and *G-C* coordinate axes [see the inset in Fig. 1(A)]. It is obvious that this representation is composed of two independent components: *A-T* and *G-C*.

Figure 1(A) shows the map of chromosome 12 of *Saccharomyces cerevisiae* (yeast), which contains approximately 1.1×10^6 nucleotides, except a large rRNA cluster (1.5 Mb).[11] The most indicative elements of this representation are the fragments in which the trajectory "condenses" within a certain domain of the *ATGC* plane, skews, and long curved fragments of various shapes. Figure 1(B) demonstrates chromosome 1 of the same yeast. Almost identical periodic fragments with opposite directions, i.e., the complementary parts, can be discerned with naked eye (rectangles

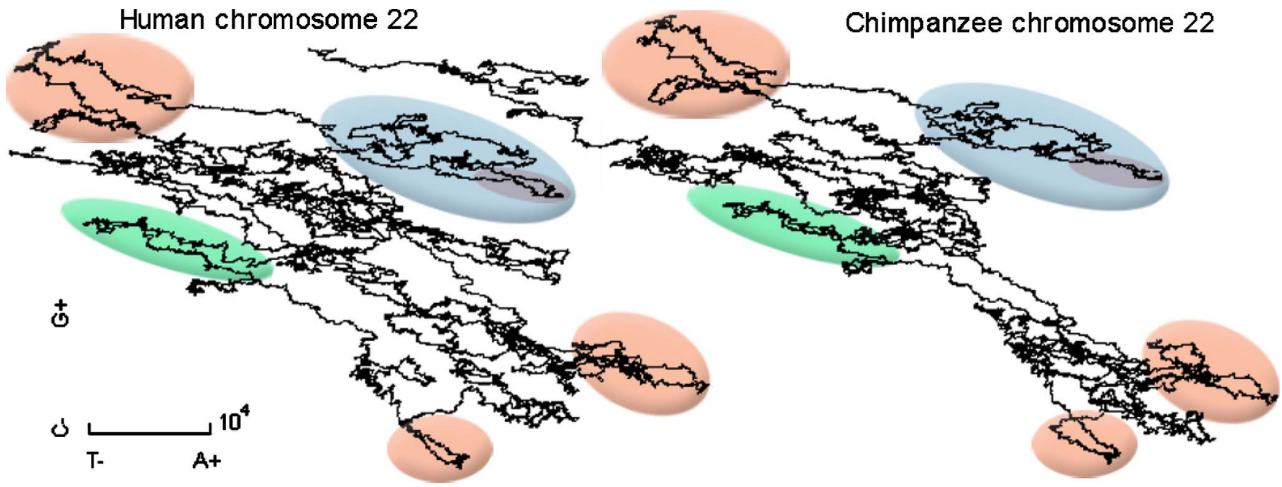

FIG. 2. (Color online) Comparison of human and chimpanzee chromosome 22 (Ref. 14), containing about 34×10^6 and 49×10^6 nucleotides, respectively. Similar selected fragments are marked by the same geometric configuration. The first 15×10^6 nucleotides of the human chromosome 22 are still unknown. In these maps one can easily recognize numerous IGPs.

containing about 4000 bp in length). Let us decompose this part into the *A-T* and *G-C* components and carry out the wavelet transformation of these series. One can see that even such a simple analysis reveals an expected and interesting property of self-similarity [Fig. 1(C)].

In addition, we may zoom in the obtained 2D DNA map and see details in different scales. With the knowledge of the site numbers, it makes possible to get the corresponding information from databases.[10,12,13]

Applying the 2D walk method, one may easily find huge palindromes with a considerable share of pseudorandom inclusions. Keeping in mind characteristic features of such palindromes, they can be called imperfect gigantic palindromes (IGPs). One of these, 35 Kb in length, is selected in Fig. 1(B) by an ellipse. This is the region with the mobile genome element represented by the *Ty*-family of retrotransposons. It should be noted that we use the term "palindrome" (as it is often used in molecular biology and bioinformatics) that does not mean "true palindrome," but complementary palindrome, where a sequence in the chain has an extension with complementary sequence in the opposite direction.

In addition, it is easy to identify telomeres, located at chromosomes ends. In Fig. 1(A) they are selected by the small rectangles. Each fragment has the size of approximately 20 Kb; they are complementary. In Fig. 1 we may clearly see monotonous composition sequences without specific motifs, which can be recognized by the scale magnifying (see a prolate rectangle; here the cluster of ribosomal RNAs is located). We can also see long skews. The slope of these fragments suggest that the complementary strands of the DNA helix are anisotropic and of unequal weight (size).

Using the 2D representation, one can also compare the whole chromosomes (containing even up to tens and hundreds of millions of nucleotides) of different organisms and get a more deep insight into the genome evolution (see below). In Fig. 2 the maps of human and chimpanzee chromosome 22 are shown. The results of this comparison have been reported in Ref. 14, where the examples of the diffusion process in large chromosome parts and formation of numerous IGPs have been presented. In Fig. 2 we easily recognize similar fragments which are marked by ellipses of the same color. We consider these fragments in detail in Sec. IV.

III. DNA COMPLEMENTARITY AND PALINDROMIC CONTEXT OF GENOMES

In this section we explain that palindromes in genome sequences have a wide range of sizes, complexity, age, and functions. It is evident that palindromes originate from a simple property of the complementarity of the DNA structure. This means, naturally, that this type of the interaction should be abundant in genomes. The last investigations of the human genome showed the presence about 12.5×10^6 of the exact small palindromes.[15]

The most part of the DNA and RNA binding sites (replication sites, transcription factor patterns, and so on) which can be recognized by different agents (such as the protein complexes and small RNAs), consists of a complex self-complementary DNA/RNA structures. Such complementary features of DNAs (or RNAs) sequences in the 2D DNA map generate the fractal landscape on small and large scales[16,17] and reflect the second rule of Chargaff: $\Sigma A=\Sigma T$ and $\Sigma G=\Sigma C$ in a single DNA strand on average.

Recently, several papers devoted to the analysis of the human gigantic palindromes have been published. In particular, in papers,[18,19] the human ampliconic gene families have been analyzed. In the Y human chromosome, the authors found the twin gene series located at a distance about 2.9 Mb from each other and formed a palindromic sequence. They used the Dotter method,[20] which works only in strong and long similar cases.

Let us analyze the 2D map of the entire Y human chromosome [Fig. 3(a)] and its part [Fig. 3(b)] 2.9 Mb in length [see gray rectangle in Fig. 3(a)], which includes the longest palindrome (known as P1). We can see that some parts are an almost exact palindrome except for duplicated sites with DAZ genes. This means that during evolution, the diffusion

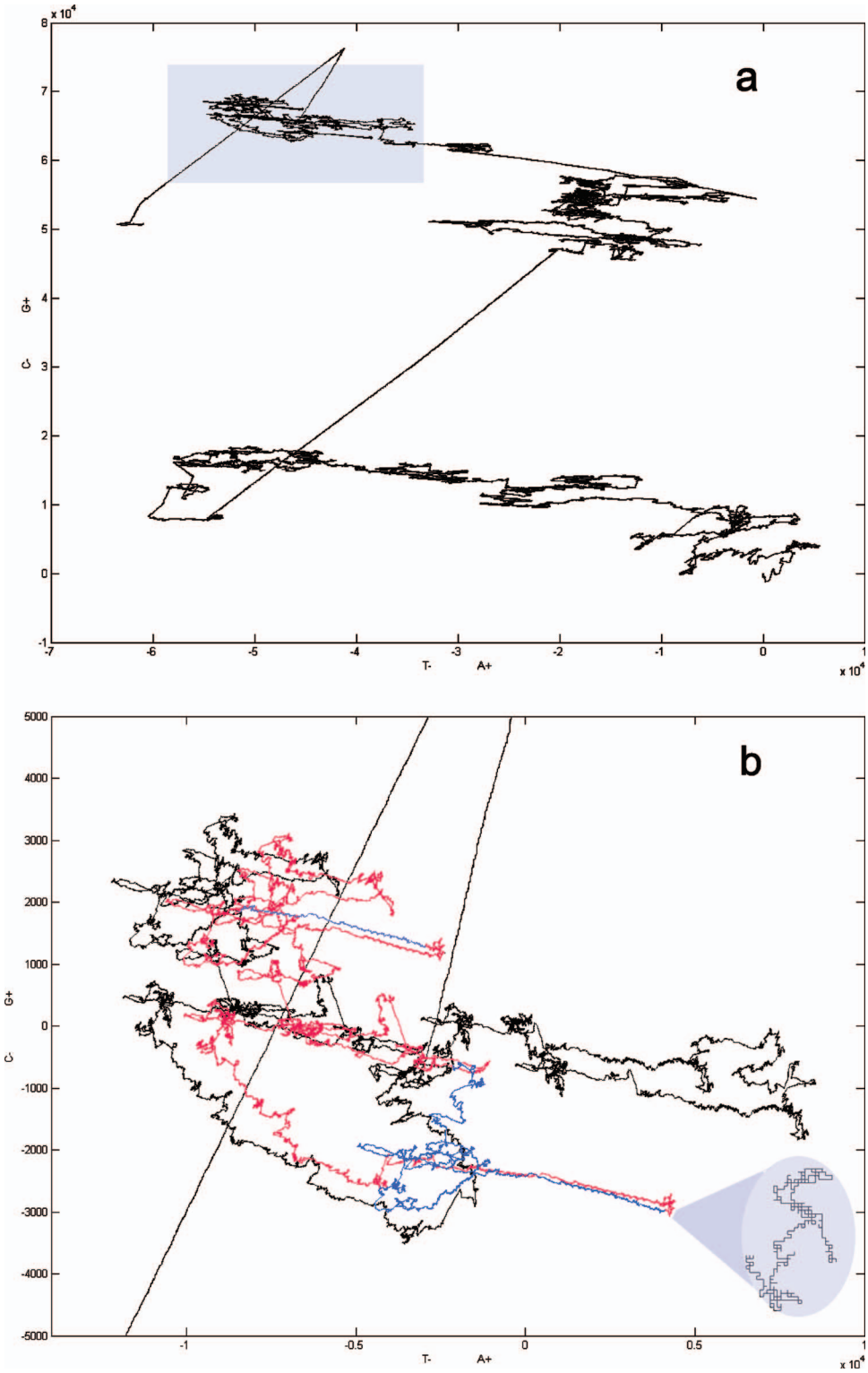

FIG. 3. (Color) Y human chromosome containing about 58 Mb (a) and its greatest Palindrome 1 [(b) and grey rectangle in (a)] with DAZ genes [gray ellipse in (b)]. Surprisingly, in the 2D plane, the shape of the Y human chromosome looks like the letter "Z."

process almost did not take place in this DNA region. Therefore, here we deal with a "quite young" palindrome.

Considering IGPs, owing to the fact that they were formed by means of the complementary duplication, one can naturally wait the presence in these palindromes the inverted repeat phenomena. It is evident that the inverted repeats and the palindromic context of large chromosome sequences are based on simple properties of the DNA helix as the complementarity.

Using the data presented in Ref. 21 by means of the inverted repeat analysis and our 2D walk method, we easily found a new *duplicated gigantic palindrome* in the X human chromosome [Fig. 4(b)]. It is surrounded with a cluster which is similar to the IGP [Fig. 4(a)]. In the human genome, such a duplicated palindrome is detected for the first time. The authors of the paper[21] have analyzed the human genome and found a large number of the inverted repeats up to 1 Mb in length. Most of them, for all the human chromosomes,

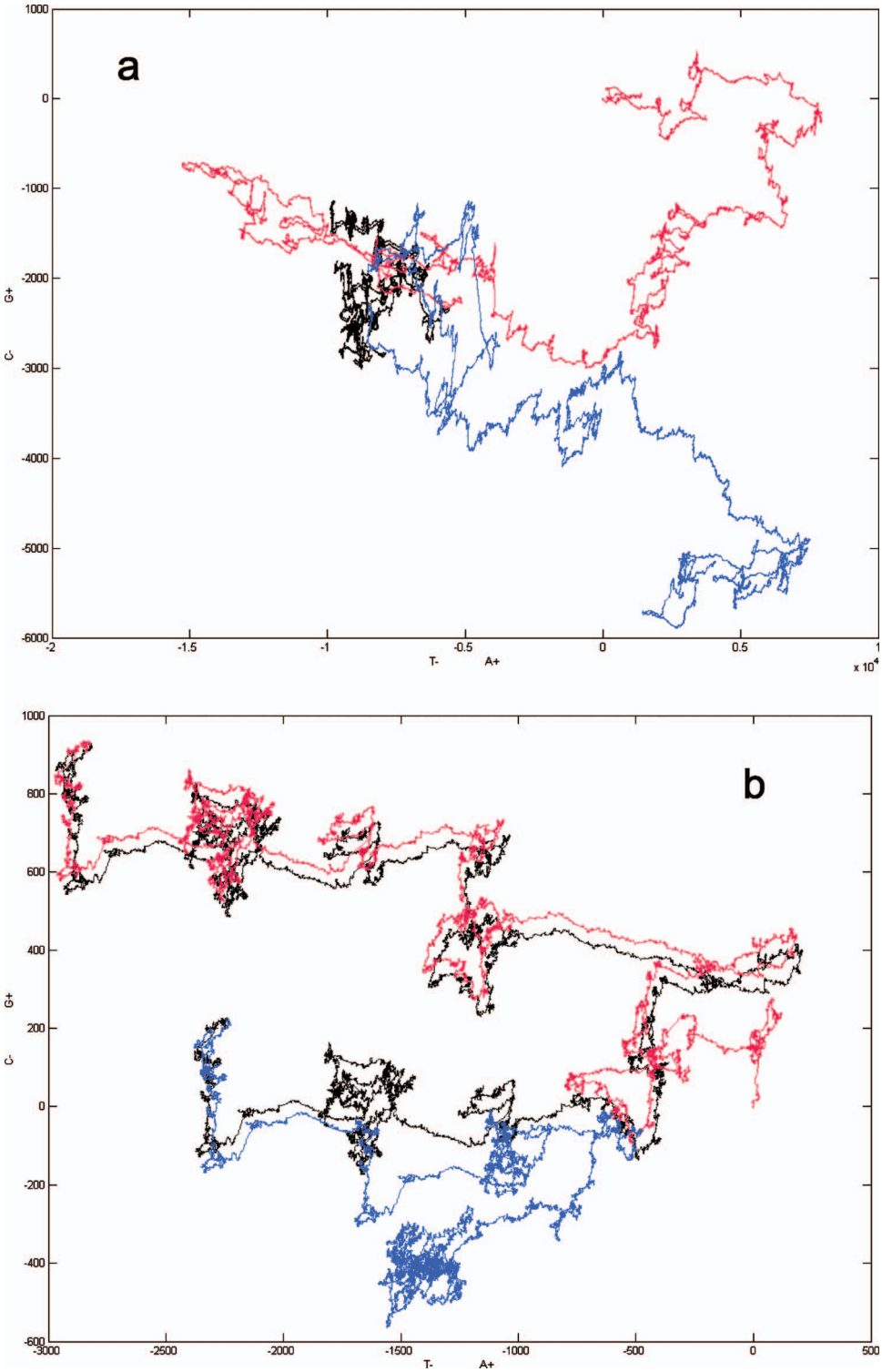

FIG. 4. (Color) The part 51–54 Mb in the X human chromosome (a) and its 300 Kb exact palindrome with a duplication and a small rearrangement (b). This palindrome contains a GAGE gene family.

contain about 100 nucleotides. However, based only on the inverted repeats, it is very difficult to say about the sequence organization and find the complex inversion clusters of IGPs. At the same time, the 2D walk method easily reveals these features in the DNA sequence (Fig. 4).

The described duplicated palindrome [see Fig. 4(b)] could be schematically formed by the mechanism presented in Fig. 5. The initial nucleotide fragment may get a new covalent bond between the complementary chains [Fig. 5(a)]. After double-strand breakage of the DNA, this unified fragment is then joined with a complementary chain [Fig. 5(b)], and so on [see Fig. 5(c)].

It is obvious that if palindromes have a small enough length, then after the diffusion process they can not be recognized. On the other hand, if palindromes are sufficiently large then in a whole they, probably, hold the shape. From this point of view the large almost exact palindromes are "very young." Simultaneously, small palindromes are very conservative and perhaps, they have a certain important function.

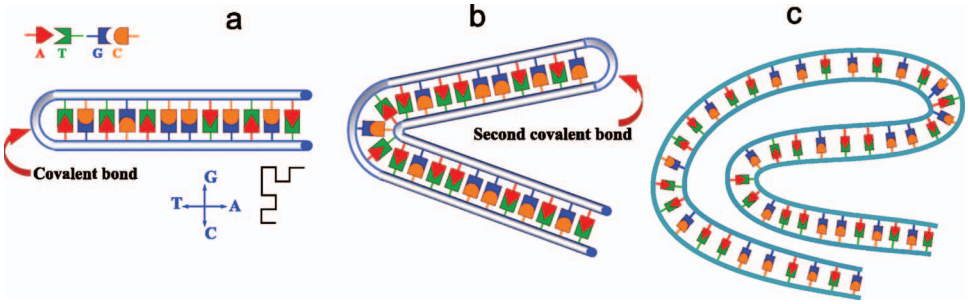

FIG. 5. (Color) A possible schematic mechanism of the creation of the duplicated palindrome.

IV. GIGANTIC PALINDROME DISINTEGRATION

Let us analyze in detail some parts of human chromosome 22 and chimpanzee chromosome 22 in the 2D map (see Fig. 2). In the center, on the left, the IGP of 3 Mb in length are shown. They have a large part of conservative sequences and small sites with a large numbers of indels. The processes of diffusion and pseudorandom sites inclusions may crucially complicate the gigantic palindrome identification. We have observed these types of the sequence organization in many other considered genomes and found that IGPs are not rare event in genomes[22] as it was earlier supposed.[23]

One may assume that a part of the region in the chimpanzee chromosome (a fragment marked by a small gray ellipse, at upper right in the center ellipse (looks like a rabbit) containing about 300 Kb) initially was a palindrome. During evolution, this palindrome has undergone the diffusion process (i.e., mutations, inclusions, deletions and, possibly, some kind of the rearrangement) and disintegrated.[24] With the divergence of human and chimpanzee genomes, the conservation level of sequences of this region was different. In other words, for the human chromosome the fragments which are similar to the palindromes, become blurred (see, e.g., regions marked by the center ellipses in Fig. 2). Therefore, from the standpoint of the divergence time problem, large palindromes may help to analyze the evolutionary process.

The chromosome rearrangements (duplications, inversions, transitions, and translocations) were under scrutiny since the initial works on "chromosome mechanics" by Morgan and Sturtevant (see Ref. 25). The first idea of the palindrome type amplification in the context of the transposition process goes back to McClintock.[26] During the last several decades, huge palindromes, from tens of Kb up to hundreds of Kb in length, were known and analyzed.[27] Some models of the large palindrome formation in different aspects of medical and biological problems were also discussed.[23]

In the evolutionary aspect, the diffusion of gigantic palindromes in the human chromosomes (about several hundred Mb in length) has been described firstly in Ref. 24. The 2D walk images of theses chromosomes are shown in Fig. 6. It is easy to see that the longest human chromosomes (i.e., chromosomes 1–4) consist almost completely of IGPs.

From the technical point of view, owing to the intensive mutation drifts and rearrangements, it is quite difficult to recognize IGPs by other methods because most of them have small enough "hits" in alignments (as an example, see Refs. 28–31) or may require an additional complex analysis.

In Fig. 7 we show the maps of human chromosomes 6 and 9. As it follows from Figs. 6 and 7, such types of palindromes (i.e., IGPs) may be spoken of as "context" or "composite" palindromes because only in a *macroscale* they have

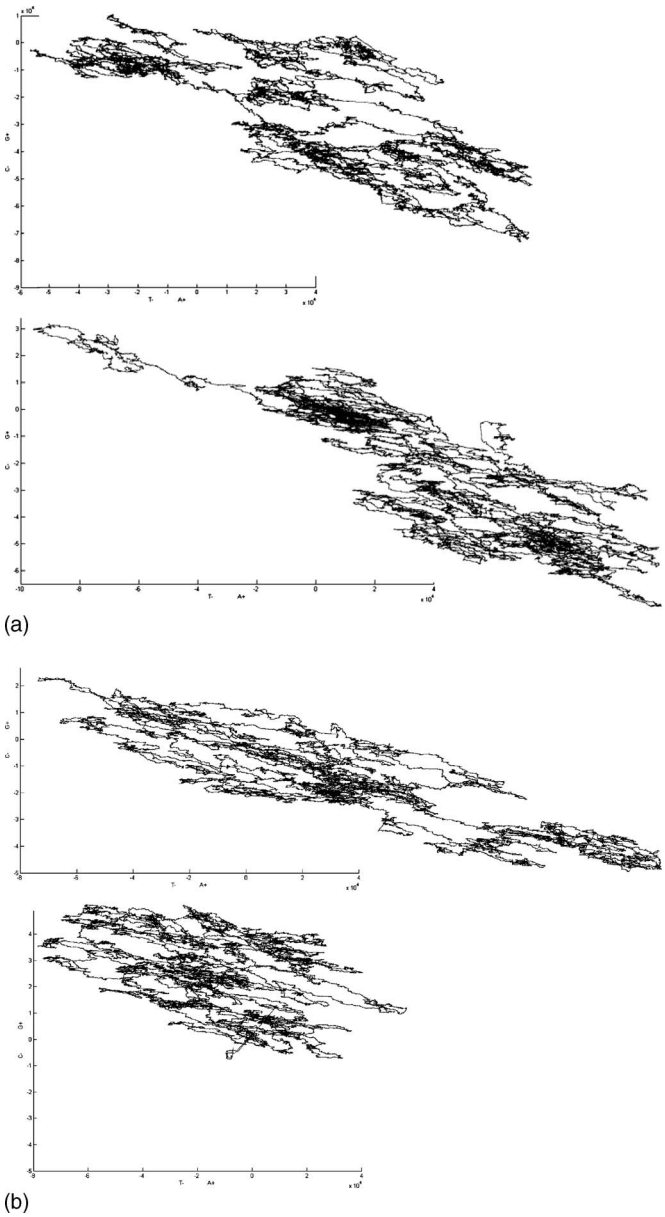

FIG. 6. The first four human chromosomes. The nucleotide lengths are (from top to bottom), 1: 247 249 719 bp; 2: 242 951 149 bp; 3: 199 501 827 bp; 4: 191 273 063 bp.

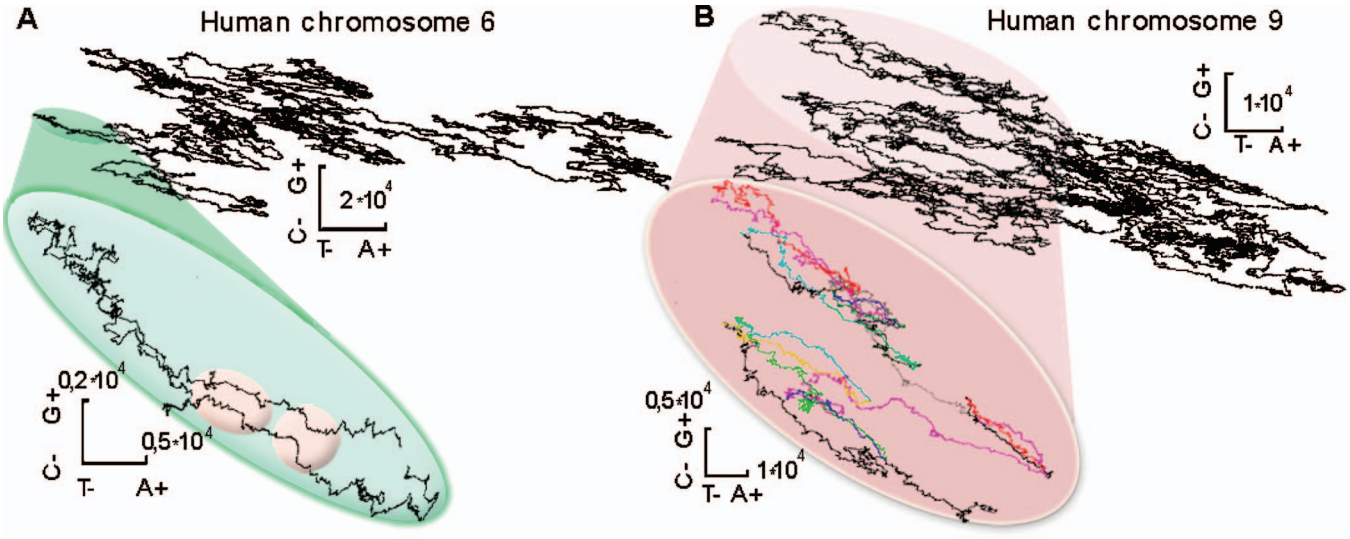

FIG. 7. (Color) (A) Human chromosome 6 (171 Mb): A 3 Mb palindrome (large ellipse) and its most correlated parts (small ellipses within the large ellipse). (B) Human chromosome 9 (140 Mb): A 30 Mb inverse palindrome region (large ellipse). Different parts of its inversion are shown by various grey tints colors. 16 Mb of this chromosome are still unknown.

an organization of the palindromic type. This means that in the 2D map we can see only the direction of the sequence displacement which has the size in terms of the 2D walk, on average, from 1% up to 10% of the real sequence length. That is the reason it is difficult to find this hidden context by others graphical alignment methods (see, e.g., Ref. 20).

At the same time, the nature of such a context does not have the origin of a pseudorandom *AT*-rich sequence, as it was supposed in Ref. 32. In Fig. 7(A) (chromosome 6, small ellipses inside the large ellipse) we see that a large part of sequences has a complex nucleotide composition which correlates with the reverse part of the palindrome. Moreover, after "diffusion processes," the palindromic type of these giant sequences can be identified only on large scales as an inverted quasisymmetrical curve. When we look at these palindromes on a local level, we may infer that they have a low sequence similarity between the arms, and the alignment procedure works here only on small lengths of certain sites.

The described 2D walk analysis allows us to establish (see Refs. 22 and 24) that the palindromic sequences can be much longer [from megabases to several tens of megabases and maybe more; see Fig. 7(B)] and have a wide class of the complexity as a result of the unstable nature on a local level.[33] It is obvious that the chromosome rearrangement (duplications, inversions, transitions, and translocations) and mutations (inclusions and deletions) disintegrate the gigantic palindromes. From the physical point of view this means that the gigantic palindrome disintegration has a wide time-scale hierarchy. Some palindromes may be much older than the divergence time between close species. However, some of them exist just in some species exclusively.

To show this, we present the comparison of human chromosome 12 and chimpanzee chromosome 12 (Figs. 8 and 9). Each of these chromosomes was divided into six parts. Analyzing the obtained structures one can come to the conclusion that all of them consist of IGPs. The most part of human and chimpanzee chromosomes has almost identical form. We may also easily see rearrangement.

Analyzing these images it is necessary to say, however, about gaps in the assembled sequences of these chromosomes (the human chromosome, release 36.2 NCBI and the chimpanzee chromosome, release 2.1 NCBI) [release 46 Ensembl (Aug. 2007, see Figs. 1 and 2 in Supplemental Material[59])]. At the present, in human chromosome 12 there are eight gaps from 16 Kb up to 1.4 Mb in length (in sum about 2046 Kb). Chimpanzee chromosome 12 has 68 gaps of the length from 21 bs up to 1.4 Mb; in sum, 1 936 502 bs. As a result, there are some shifts in the presentations of the corresponding 2D DNA maps. The largest gap (1.4 Mb) in human chromosome 12 is located after 34.660 Mb of the assembled sequence and 89 Kb gaps. In chimpanzee chromosome 12, the greatest gap (also 1.4 Mb) is located after 43.604 Mb in the assembled sequence and 143 Kb gaps.

In addition, if we speak about the rearrangement, we should take into account the existence of the genome variations. As is known, the most variations of the human chromosomes (HapMap project[34]) are presented by single nucleotide polymorphism (SNP); these variations do not appear in the 2D map on the full chromosome level, but they can be detected on the protein level. Recently, different types of variations in genomes on large scale were described. These are copy number polymorphism, large inversion polymorphism, segmental duplications, and structural variation (see Refs. 35–37 and references cited therein).

In particular, large inversion polymorphism in the human genomes has been recently analyzed.[38] The authors found, by the SNP data, a large number of the gigantic inversion regions (about 176 large inversion sites based on the data of 269 human individual genomes). The largest inversions run the length up to several Mb. Specifically, this means that on large scales a certain dynamics is observed. In addition, the

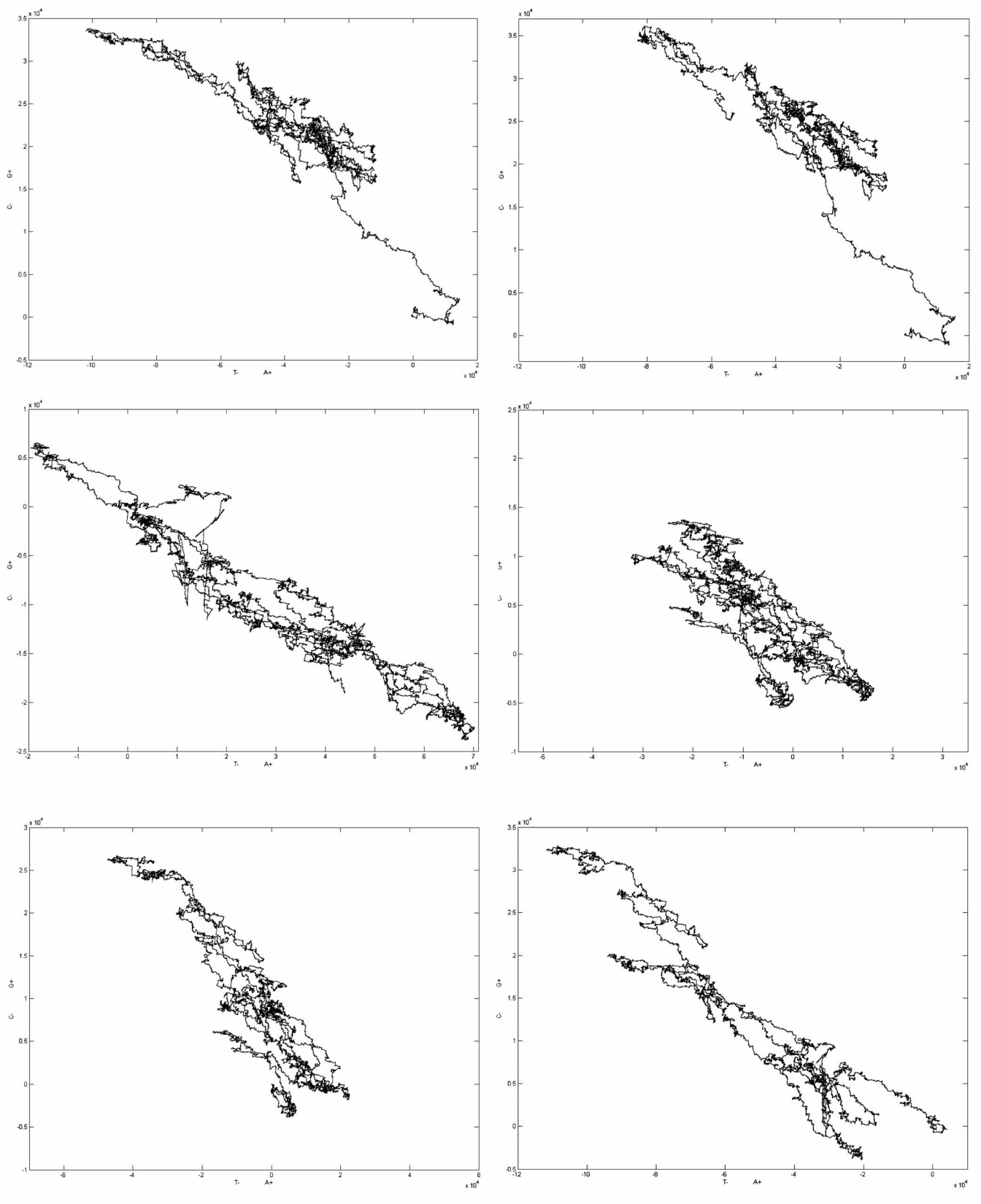

FIG. 8. The first three parts of 12 human (left column) and 12 chimpanzee (right column) chromosomes. The length of each part is about 25 Mb.

inversion polymorphism in the human and chimpanzee genomes has been found in Ref. 39 (see also images in our Supplemental Material[59]). Apparently, this dynamics is related to other time-scale phenomena than structures of IGPs that we observe in the 2D images. Let us turn to Figs. 8 and 9. In these images, the IGPs are presented in both genomes with some differences, but have the same structure in a whole. This means that large variations between the similar

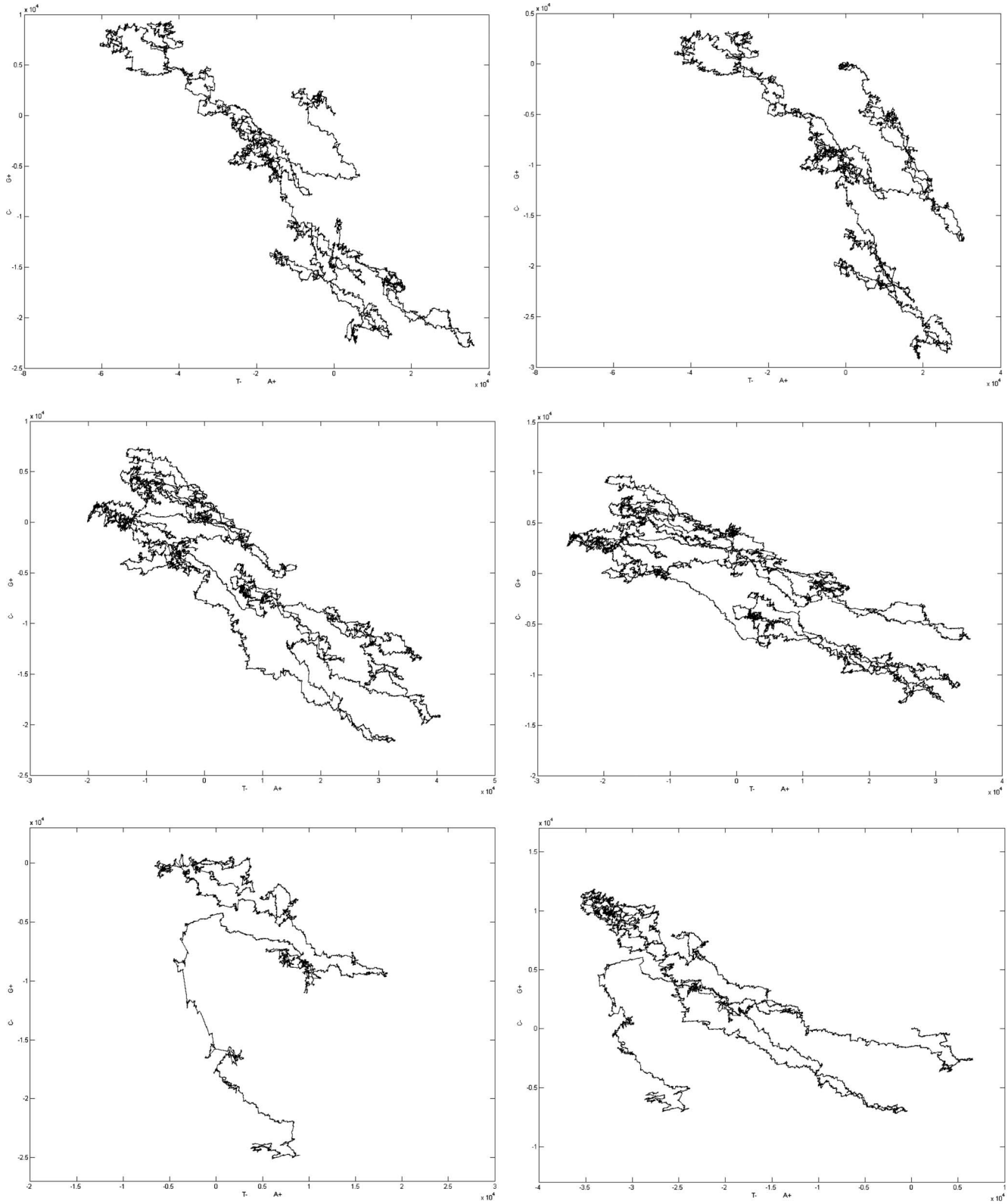

FIG. 9. The last three parts of 12 human (left column) and 12 chimpanzee (right column) chromosomes. The length of each part is about 25 Mb (except for the last part about 7.5 Mb).

chromosomes of human and chimpanzee take place in the other time scales.

We assume that the described 2D DNA presentation can provide a new insight into the evolutionary processes of the chromosome rearrangement theory. During the last years this theory, owing to new assembled genomes, received an additional stimulus.[40,41] Probably, as a part of rearrangement, the complementary duplication of a sequence, i.e., amplification by means of the complementary chain binding (which forms the palindrome), together with the diffusion

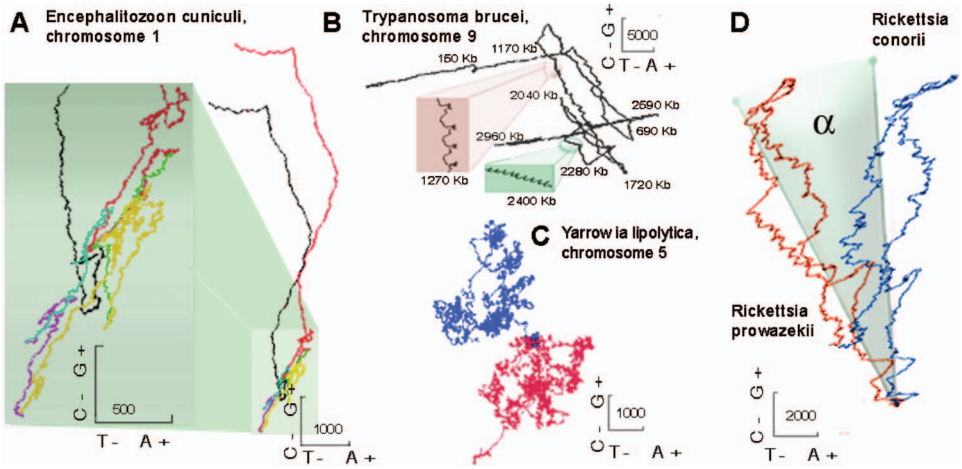

FIG. 10. (Color) (A) Chromosome 1 of the eukaryotic microorganism *E. cuniculi*. (B) Chromosome 9 of *T. brucei* related to parasitic eukaryotic microorganisms. (C) Chromosome 5 (E) of the fungi *Y. lipolytica*. (D) Genome sequences of *R. prowazekii* and *R. conorii*: "the rotation angle α" of the 2D DNA map.

process, expands the strategy of the diversity resource generation.

V. FROM COMPLEMENTARITY TO CO-FUNCTIONALITY

Development of the described 2D walk method has allowed us to find chromosomes in some eukaryotic microorganisms and in prokaryotes that have the shape and/or organization similar to IGPs. The given observation suggests that such a spread occurrence of IGPs may be functionally significant. In this sense, the palindrome disintegration process, as well as a full genome duplication, may be represented as a part of the optimal evolutionary strategy from the first stage of the genome formation.

As an example of the divergence process of IGPs, the length of which is comparable with the whole length of chromosome, let us consider several chromosomes of fungi and reasonable small close bacterial genomes (Fig. 10).

In this figure, chromosome 1 of *Encephalitozoon cuniculi* (related to fungi) consisting of long telomeres (red and black curves) and central part with many palindromic inversions is shown. Parts of palindromes are marked by different colors. In Fig. 10(B), one can see chromosome 9 of *Trypanosoma brucei* (related to parasitic eukaryotic microorganisms) and some types of its fragments. These fragments demonstrate a set of repeats. The structure of this chromosome looks like a duplicated and rearranged palindrome.

In addition, in Fig. 10(C), chromosome 5 of *Yarrowia lipolytica* (related to fungi) is shown. One can see that this map is decomposed into two large clusters, so that either of the two has the sequence about 2 Mb in length. It is known that large parts of genes in this fungi genome were duplicated and triplicated.[42]

Figure 10(D) demonstrates the genome sequences of two bacteria: *Rickettsia prowazekii* and *Rickettsia conorii*. These genomes have been considered in detail by many authors (see, e.g., Ref. 43 and references cited therein). Analyzing this map, one easily comes to the conclusion that it is possible to represent the mutation process qualitatively like an α-degree turn of the curve mass center on the 2D DNA space. This turn represents the difference between corresponding arm skews in these genomes. Lobry and Sueoka[44] noted that the most of bacterial genomes share asymmetry in the nucleotide concentration between arms, which are divided by "Ori" site (from the word "origin"). This is very easy to observe in the 2D map [see Fig. 10(D)].

A question on the composition asymmetry is one of the hot spots in the bacterial genome evolution problem. In Fig. 10(D) we show that the possible origin of this asymmetry has the nature of the ancient duplication of the complementary chain. During the divergence in *R. prowazekii*, about 40% genes (in comparison with *R. conorii*) have been lost. The correlation components and the corresponding interactions between arms had the higher conservation level than other subsystems,[45] which realized the control of the nucleotide compositions within each arm. This means that there is a certain compensation mechanism between the compositions of arms. The nucleotide composition of genes is changed by a concerted way. This coordination implies a global character of their co-functionality that is connected with the property of complementarity of DNA. This follows, properly speaking, from the complementary character of the turn of genome sequences in the 2D DNA map.

We suggest that in this mutation process the correlation component had a complementary palindrome origin at the first stage of the bacterial genome formation. The subsystem of replication may also be connected with this correlation. Thus, the early stages of the evolutionary process could derive, probably, from the complementarity of the DNA chains, through palindrome disintegration, to the co-functionality of the DNA sites.[46]

Turning back to the close examination of the eukaryotic chromosomes, one can note the following. Our qualitative analysis shows that IGPs in large genomes, on the one hand, are prevalent and, in spite of the rearrangement, hold the shape as a whole. On the other hand, they keep some correlated remote sites with the similar complexity and, also, have the common inverted composition. Taking into consideration that IGPs are observed in genomes of different species, we suppose that some of these palindromic types of sequences may be related to the replication subsystem. As we discussed above, in the most of bacteria genomes the replication process start in an "origin site," the location of which divides the genomic sequence into two arms with the asymmetry in

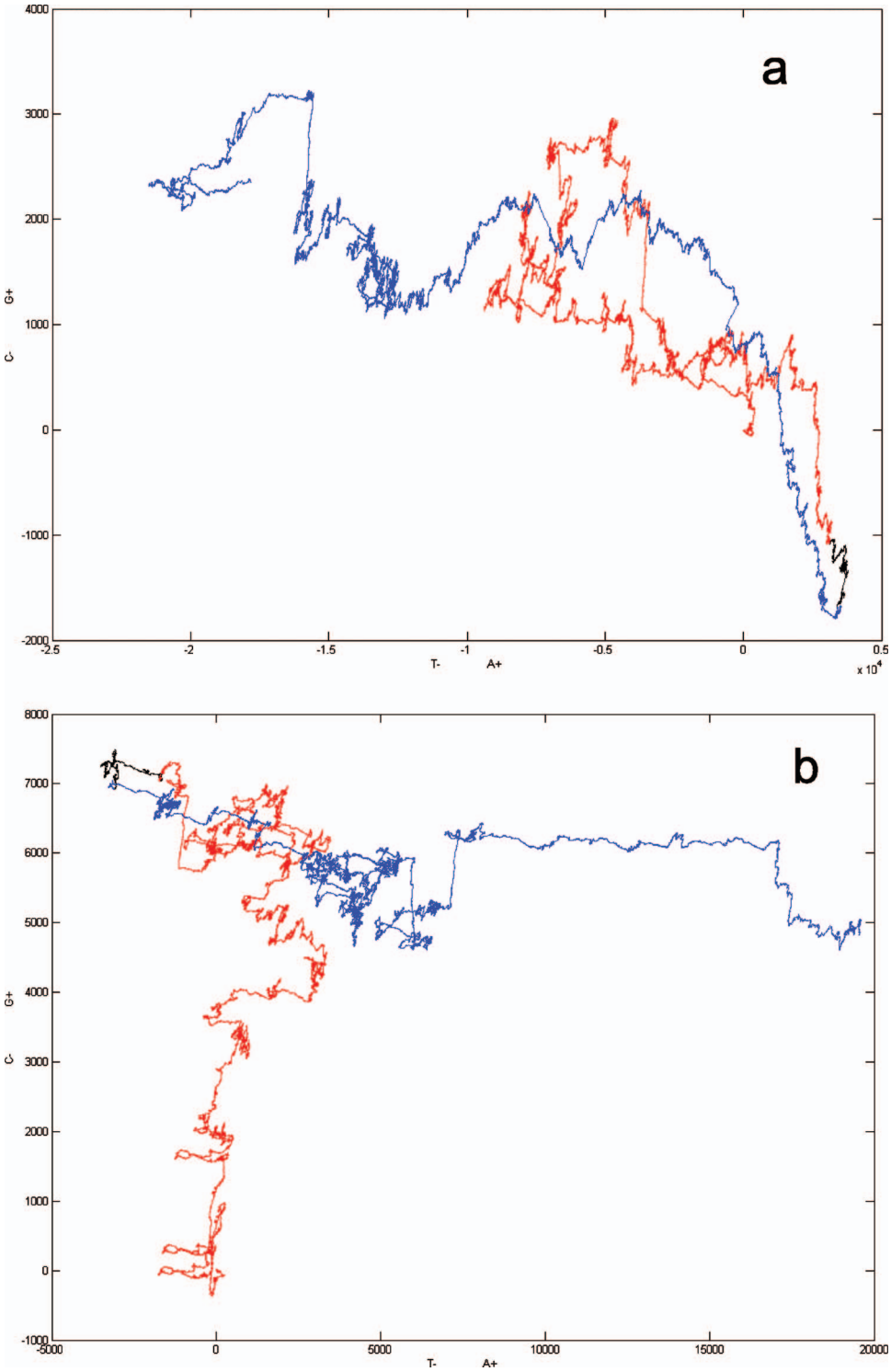

FIG. 11. (Color) A fragment of human chromosome 8, 2 Mb in length, containing the MYC gene in the center, marked by black color (a) and a fragment of human chromosome 11, 2 Mb in length, containing the HBB gene in the center, marked by black color (b).

the nucleotide composition and in the transcription orientation. This subsystem in bacteria is organized in a single replicon with the origin site and with the "termination site" on the opposite position of the circular genome sequence. It is also known[47] that in the chromosomes of high eukaryotes the replication process begins with a number of replication origins, which are distributed along the sequence. Every replicon in the human chromosomes has the length from several hundred Kb to several Mb. We have tested this our assumption on examples of the well known experimentally determined sites in the human chromosomes.[48] Here we present such images in Fig. 11. In the overwhelming majority of cases we observe that the experimentally detected sites of the replication origin are surrounded by the regions similar to the disintegrated palindrome.

Very recently, we considered the 2D map of the chromosome regions described in Ref. 49. The authors analyzed the human genome sequence by the wavelet-transform for the purpose to reveal the sites with the inverted compositions and inversions in the gene transcription direction. The 2D DNA map analysis displays that, at large fragments (more than 2 Mb) the inverted character of sequences can be really

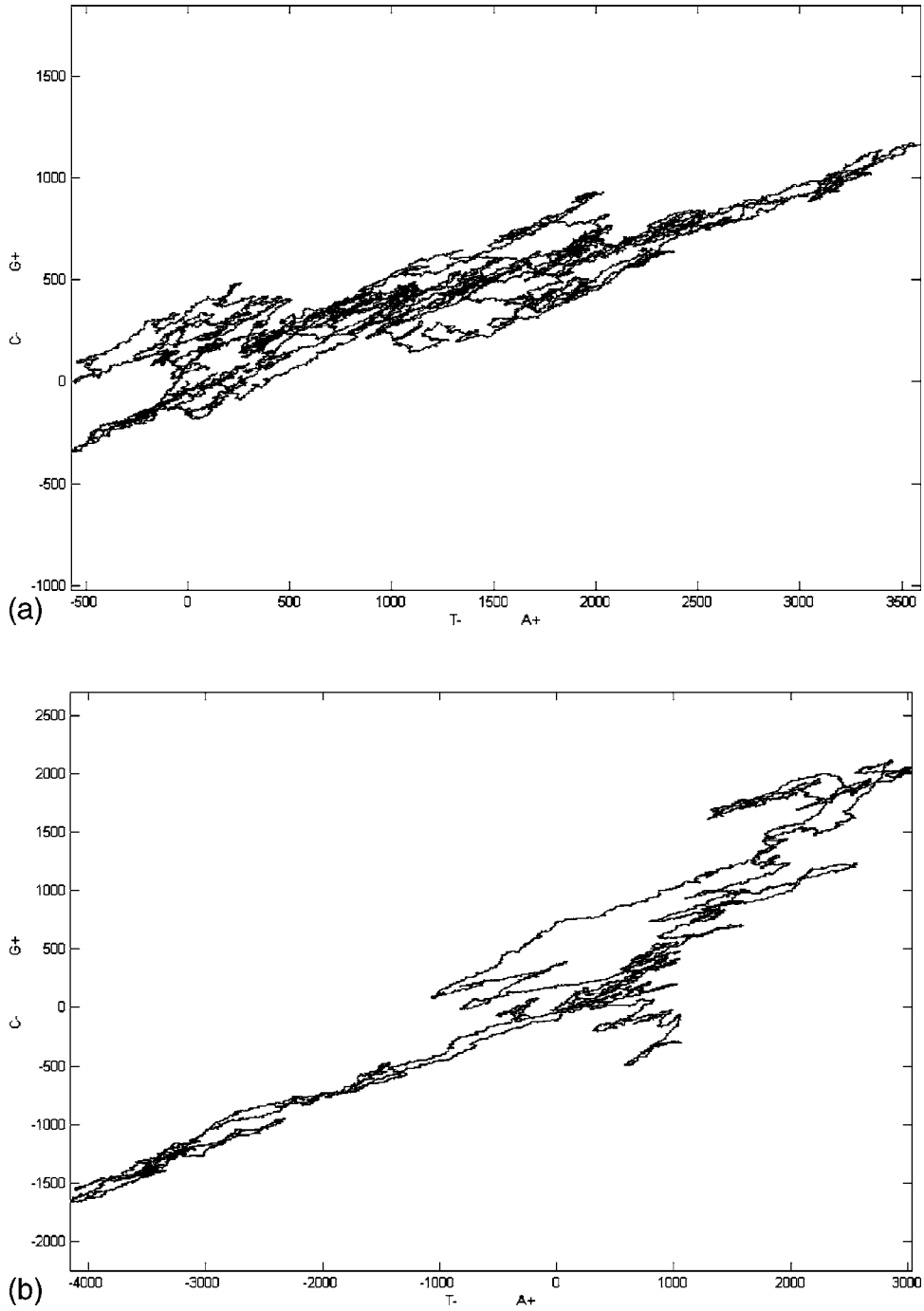

FIG. 12. The chromosome of *N. equitans* (490 Kb, top), and the chromosome of *W. brevipalpis* (697 Kb, bottom).

observed. At the same time, the fragments less than 1 Mb look like a quasirandom walk (see figures in our Supplemental Material[59]). It should be noted that, comparing the results of the paper[49] with the experimentally obtained data,[48] we have found only one coincidence from more than 600 supposed hits. However, the replication origin question is a very complicated problem for the experimental detection. At the present, for the mammalian genomes, there is no sufficient experimental data about the replication subsystems.

In contrast to the these data, now for eukaryotic microorganisms (namely, for *S. cerevisiae*) the database ORI DB[50] is composed. This database contains 732 expected replication sites. 325 of them are experimentally confirmed. In *S. cerevisiae* the sites of the replication initiation have the length about 200 bp and contain autonomously replicating sequence (ARS). Certain sites are included in the imperfect palindrome clusters.[46] It is known that in *S. cerevisiae* the replication origin sites are distributed along the chromosomes with distances between each other about several kilobases. As a result of the distribution of such lengths (which are significantly smaller than the replicons in high eukaryotes), a local rearrangement forms the 2D map as a pseudorandom walk.

Finally, we have additionally made the same 2D analysis of the chromosome of *Nanoarchaeum equitans* and *Wigglesworthia brevipalpis*. The first genome belongs to the archaea domain; the second one is a bacteria (Fig. 12). We have found that all of them consist mainly of the IGPs and rear-

rangement processes. Therefore, the palindromic context is widespread phenomena not only for high eukaryotes, as it was described above, but also inherent in protozoa. Apparently this is very important strategy in the evolutionary process for genomes as a whole.

VI. ANDERSON SPIN-GLASS MODEL AND NETWORK PHENOMENOLOGY. ORIGIN OF METABOLICAL/REGULATORY NETWORKS

A spread occurrence of IGPs, found in different species, suggests that the given type of the sequence organization is the evolutionary significant. We suppose that the strategy of the complementary duplication, which initially leads to the IGP formation, may have a longstanding origin in the sense that it is based on the main DNA property; i.e., complementarity. Probably, the formation of the replication subsystem in genomes with replicones (see Sec. V) is a initial stage in this strategy. Below we describe certain ideas related to this phenomenology.

On the basis of our analysis, we may suppose that long range correlations, which were discussed previously (see, e.g., Ref. 51 and references cited therein), are the consequence of the diffusion and inversion processes.[16] During the evolution, recursion of the palindrome type amplification in the 2D map [see, as examples, Figs. 5, 7(B), and 12] in the presence of diffusion and inversion on macroscales seems to be similar to some elements in the construction of the Smale horseshoe and baker's transformation.[52] However, strictly speaking, this theory has principal differences. With reasonable modifications, spin-glass models[53] are more relevant for the description of the palindrome disintegration process.

For spin-glass models, it would be interesting to analyze the processes with domains that are formed by metabolical or regulator kinds of reactions between the gene products beginning with the first evolutionary steps. In this sense, the palindrome disintegration process can create a genetic network by the biochemical large cycle (hypercycle) doubling/bifurcation, forming interactions with the distant sites. The hypercycle, as a process between the transcript products, is organized in this approach as a unique replicon[54] that is related to the initial palindrome sequence. In our opinion, the mutation drift (diffusion) can play a special role in this process, because in the palindrome sequence each node ("pre-gene") has its own pair in the complementary strand of another arm ("mirror sequence" in the 2D DNA map). All nodes within one palindrome interact locally with their neighbors by the transcripts. If some of them are blocked by the mutation drift, their pairs in the complementary sequence may play a role in the "hypercycle metabolism" via the time synchronization. In the next generation, the replication process can fortify this solution. During billions of years of evolution, the palindrome disintegration process may move off the essentially complementary context of such nodes, but the metabolical/regulatory connection may be conserved.

Genome sites with blocked domains (known as pseudogene sequences) are eliminated with the larger mutation pressure[55] and are reduced or cut with the cluster sequences.[56,57] The local and remote interaction of the transcript process products may form the multidomain proteins and/or small stable cycles. It was also possible that, after elimination, some of local residual sequences may organize a new interaction between the transcript products with the neighbor sites. Our phenomenology is possibly related to the stage that has been mentioned by Eigen[32] in his hypercycle model, when ligase connects different nucleotide sequences to a unified chain. In this stage, a transition from an "RNA world" to a "DNA world" takes place.[32] Perhaps, the described above palindrome disintegration dynamics produces a conflict interaction between local and distant sites in the kinetic subsystem. In other words, this kinetic subsystem may have frustrated properties. Synchronization of such dynamics should be conjugated with compartmentalization of hypercycle.

This, perhaps, in a large ensemble of sequences and after a large number of generations, can create the net for small survived parts in the ensemble with optimal preferential properties (say, the maximal net length or the net distribution as an analog of the Anderson prebiotic model for chains). Such nodes ("pre-gene") may be represented by small sequences (up to $\sim$100 nucleotides[32]), which later may form the contemporary structural domains of proteins or small RNA that play a major role in the regulatory network.

This paradigm also appeals to the last research in the conservatism of the domain order[45] and to the role of the duplication mechanism (but another kind, older) in the network evolution.[58] We suppose that such a duplication context has a very old origin. The alternative splicing, which is the most common mechanism of the large protein diversity, also supposed the primary significance of domains in the protein evolution as a source of the functions combination.

VII. CONCLUDING REMARKS

A special interest of our paper is gigantic palindromes, which we found in large quantities in various genomes. The size of such palindromes is in a wide interval: from 5–10 Kb to several tens Mb and more. Most of these sequences have a large number of mutations and look like noisy palindromes that we call imperfect gigantic palindromes (IGPs). Alignment of these regions between sequences that compose the palindromes, gives us a low similarity, so that identifications of the IGP is possible only by the described the 2D method. This type of the similar sequences is said to be "context-homology" sequences. We found that huge palindromes, megabases in length, are common events in many species rather than a case of illness mutation (e.g., a cancerous amplification).

The 2D map analysis of a large number of chromosomes allowed us to detect new phenomena. In particular, we identify a duplicated gigantic palindrome in the X human chromosome and advance schematic phases of such a palindrome formation.

The performed comparisons of the corresponding human and chimpanzee chromosomes clearly point to the fact that some their fragments are almost identical. On the other hand, during evolution, these two genomes was different that led to a qualitatively distinction in the respective palindromes.

Our investigation makes it possible to reveal that, perhaps, some chromosomes of fungi and bacteria have in their

origin the disintegrated gigantic palindromes. We assume that the strategy of the complementarity duplication may be a reason of the IGP creation. Formation of the replication subsystem in genomes is probably regarded as one of the phases of this strategy. A prevalence of IGPs and their resemblance to the experimentally revealed regions of replications in the numerous chromosomes allows us to discuss certain details of the genome evolution in the prebiotic phase. Such a phenomenology of the chromosome organization, perhaps, may also be explained in terms of a spin-glass model.

In our opinion, the performed qualitative analysis provides an opportunity for a more careful development of identification algorithms and permits to discover unknown sequence properties.

ACKNOWLEDGMENTS

We would like thank S. Rybalko for his assistance, A. Dzhanoev for discussions of mathematical aspects, M. Poptsova for assistance and discussion on bacterial genomes analysis, N. Brilliantov for discussion of selected problems, I. A. Zakharov-Gezehus for helpful discussions of Rickettsia genomes comparisons, B. Dujon who proposed to consider the *Yarrowia lipolytica* genome, and H. Renauld, who directed our attention to *Trypanosoma brucei* genome and comments. We also thank A. Khokhlov, P. Schuster, A. Bairoch, J. Thornton, A. Valencia, M. Linial, Sh. Pietrokovski, L. Hurst, J. Skolnik, and I. Friedberg for their interest and useful comments.

[1]V. V. Lobzin, and V. R. Chechetkin, "Order and correlations in genomic DNA sequences. The spectral approach," Usp. Fiz. Nauk **170**, 57–83 (2000); [Phys. Usp. **43**, 55–81 (2000)].

[2]S. W. Golomb, "The genetic code," in *Mathematical Problems in the Biological Sciences*, edited by R. Bellman (Amer. Math. Soc., Providence, RI, 1962).

[3]M. A. Gates, "Simpler DNA sequence representations," Nature **316**, 219–219 (1985).

[4]E. Mizraji and J. Ninio, "Graphical coding of nucleic acid sequences," Biochimie **67**, 445–448 (1985).

[5]Ch. Berthelsen, J. A. Glazier, and M. H. Skolnik, "Global fractal dimension of human DNA sequences treated as pseudorandom walks," Phys. Rev. A **45**, 8902–8913 (1992).

[6]A. Nandy, "New graphical representation and analysis of DNA sequence structure. I: Methodology and application to globin genes," Curr. Sci. **66**, 309–314 (1994).

[7]G. Abramson, P. A. Alemany, and H. A. Cerdeira, "Noisy Levy walk analog of two-dimensional DNA walks for chromosomes of *S. cerevisiae*," Phys. Rev. E **58**, 914–918 (1998).

[8]A. Rosas, E. Nogueira, and J. F. Fontanari, "Multifractal analysis of DNA walks and trails," Phys. Rev. E **66**, 061906 (2002).

[9]P. Vincens, L. Buffat, C. Andre, J.-P. Chevrolat, J.-F. Boisvieux, and S. Hazout, "A strategy for finding regions of similarity in complete genome sequences," Bioinformatics **14**, 715–725 (1998).

[10]S. A. Larionov, A. Loskutov, and E. V. Ryadchenko, "Genome as a two-dimensional walk," Dokl. Phys. **50**, 634–638 (2005).

[11]Y.-H. Kim, D. Ishikawa, H. Ph. Ha, M. Sugiyama, Y. Kaneko, and S. Harashima, "Chromosome XII context is important for rDNA function in yeast," Nucleic Acids Res. **34**, 2914–2924 (2006).

[12]http://www.ncbi.nlm.nih.gov

[13]E. Birney, D. Andrews, M. Caccamo, Y. Chen, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, S. Graf, M. Hammond, J. Herrero, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, F. Kokocinski, E. Kulesha, D. London, I. Longden, C. Melsopp, P. Meidl, B. Overduin, A. Parker, G. Proctor, A. Prlic, M. Rae, D. Rios, S. Redmond, M. Schuster, I. Sealy, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, A. Stabenau, J. Stalker, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and T. J. Hubbard, "Ensembl 2006," Nucleic Acids Res. **34**, D556–D561 (2006).

[14]S. A. Larionov, A. Loskutov, and E. V. Ryadchenko, "What can we learn from 2D DNA walk?" Proc. of Int. Conf.: ISMB-2005, SIG "Automatic function prediction," Detroit, June, 2005.

[15]L. Lu, H. JiaP. Droge, and J. Li, "The human genome-wide distribution of DNA palindromes," Funct. Integr. Genomics **7**, 221–227 (2007).

[16]S. A. Larionov, A. Loskutov, A. D. Rybalko, and E. V. Ryadchenko, "Genome as a fractal set generated by inversion-diffusion dynamics," in *Nonlinear Waves-2006*, edited by A. V. Gaponov-Grekhov and V. I. Nekorkin (Institute of Applied Physics, Nizhnii Novgorod, 2007), pp. 491–508 (Russian).

[17]S. A. Larionov, A. Loskutov, E. V. Ryadchenko, and S. Rybalko (unpublished).

[18]H. Skaletsky, T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S. F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S. P. Yang, R. H. Waterston, R. K. Wilson, S. Rozen, and D. C. Page, "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes," Nature **423**, 825–837 (2003).

[19]B. K. Bhowmick, Y. Satta, and N. Takahata, "The origin and evolution of human ampliconic gene families and ampliconic structure," Genome Res. **17**, 441–450 (2007).

[20]E. L. L. Sonnhammer, and R. Durbin, "A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis," Gene **167**, GC1–GC10 (1995).

[21]P. E. Warburton, J. Giordano, F. Cheung, Y. Gelfand, and G. Benson, "Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes," Genome Res. **14**, 1861–1869 (2004).

[22]S. A. Larionov, A. Loskutov, and E. V. Ryadchenko, "Lessons of large scale comparisons from 2D DNA walk," Proc. of ESF Research Conf.: Comparative Genomics of Eukaryotic Microorganisms, Sant Feliu de Guixols, Spain, 12–17 November 2005.

[23]H. Tanaka, D. A. Bergstrom, M. C. Yao, and S. J. Tapscott, "Widespread and nonrandom distribution of DNA palindromes in cancer cells provides a structural platform for subsequent gene amplification," Nat. Genet. **37**, 320–327 (2005).

[24]S. A. Larionov, A. Loskutov, E. V. Ryadchenko, and S. Rybalko, "Gigantic palindrome diffusion and certain features of genomes evolution," Proc. of the Int. Symposium on Evolution of Biomolecular Structure, University of Vienna, Austria, 25–27 May 2006.

[25]A. H. Sturtevant, *A History of Genetics* (Harper and Row, New York, 1965).

[26]B. McClintock, "The stability of broken ends of chromosomes in Zea mays," Genetics **41**, 234–282 (1941).

[27]M. Ford, and M. Fried, "Large inverted duplications are associated with gene amplification," Cell **45**, 425–430 (1986).

[28]S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," J. Mol. Biol. **215**, 403–410 (1990).

[29]A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg, "Alignment of whole genomes," Nucleic Acids Res. **27**, 2369–2376 (1999).

[30]M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, NISC Comparative Sequencing Program, E. D. Green, A. Sidow, and S. Batzoglou, "LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA," Genome Res. **13**, 721–731 (2003).

[31]A. Y. Ogurtsov, M. A. Roytberg, S. A. Shabalina, and A. S. Kondrashov, "OWEN: aligning long collinear regions of genomes," Bioinformatics **18**, 1703–1704 (2002).

[32]M. Eigen, *Self-organization of Matter and the Evolution of Biological Macromolecules* (Springer-Verlag, Berlin, 1971).

[33]S. M. Lewis, E. Akgun, and M. Jasin, "Palindromic DNA and genomic stability: further studies," Ann. N.Y. Acad. Sci. **870**, 45–57 (1999).

[34]The International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs," Nature **449**, 851–861 (2007).

[35]J. A. Bailey, Zh. Gu, and R. A. Clark, "Recent segmental duplications in the human genome," Science **297**, 1003–1007 (2002).

[36]A. J. Sharp, D. P. Locke, S. D. McGrath, Z. Cheng, J. A. Bailey, R. U. Vallente, L. M. Pertz, R. A. Clark, S. Schwartz, R. Segraves, V. V. Oseroff, D. G. Albertson, D. Pinkel, and E. E. Eichler, "Segmental duplications and

copy-number variation in the human genome," Am. J. Hum. Genet. **77**, 78–88 (2005).
[37]A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and Ch. Lee, "Detection of large-scale variation in the human genome," Nat. Genet. **36**, 949–951 (2004).
[38]V. Bansal, A. Bashir, and V. Bafna, "Evidence for large inversion polymorphisms in the human genome from HapMap data," Genome Res. **17**, 219–230 (2007).
[39]L. Feuk, J. R. MacDonald, T. Tang, A. R. Carson, M. Li, G. Rao, R. Khaja, and S. W. Scherer, Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies," PLoS Genet. **1**, e56 (2005).
[40]G. Bourque, and P. A. Pevzner, "Genome-scale evolution: Reconstructing gene orders in the ancestral species," Genome Res. **12**, 26–36 (2002).
[41]J. A. Bailey, R. Baertsch, W. J. Kent, D. Haussler, and E. E. Eichler, "Hotspots of mammalian chromosomal evolution," Genome Biol. **5**, R23.1–R23.7 (2004).
[42]B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. de Montigny, Ch. Marck, C. Neuveglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J.-M. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. de Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, Ch. Hennequin, N. Jauniaux, Ph. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J.-M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G.-F. Richard, M.-L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, Ch. Bouchier, B. Coudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, and J.-L. Souciet, "Genome evolution in yeasts," Nature **430**, 35–44 (2004).
[43]H. Ogata, S. Audic, P. Renesto-Audiffren, P. E. Fournier, V. Barbe, D. Samson, V. Roux, P. Cossart, J. Weissenbach, J. M. Claverie, and D. Raoult, "Mechanisms of evolution in Rickettsia conorii, and R. prowazekii," Science **293**, 2093–2098 (2001).
[44]J. R. Lobry and N.Sueoka, "Asymmetric directional mutation pressures in bacteria," Genome Biol. **3**, R1–R14 (2002).
[45]T. Dandekar, B. Snel, M. Huynen, and P. Bork, "Conservation of gene order: A fingerprint of proteins that physically interact," Trends Biochem. Sci. **23**, 324–328 (1998).
[46]S. A. Larionov, A. Loskutov, E. V. Ryadchenko, and S. Rybalko, "Visual genomics methods: gigantic palindromes and protein clusters prediction," Proc. of the Int Symp.: In-silico Analysis of Proteins: Celebrating the 20th Anniversary of Swiss-Prot, Fortaleza, Brazil, 30 July–4 August 2006.
[47]R. Berezney, D. D. Dubey, and J. A. Huberman, "Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci," Chromosoma **108**, 471–484 (2000).
[48]I. Lucas, A. Palakodeti, Y. Jiang, D. J. Young, N. Jiang, A. A. Fernald, and M. M. Le Beau, "High-throughput mapping of origins of replication in human cells," EMBO Rep. **8**, 770–777 (2007).
[49]M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, and C. Thermes, "Human gene organization driven by the coordination of replication and transcription," Genome Res. **17**, 1278–1285 (2007).
[50]C. A. Nieduszynski, S. Hiraga, P. Ak, C. J. Benham, and A. D. Donaldson, "OriDB: a DNA replication origin database," Nucleic Acids Res. **35**, D40–D46 (2007).
[51]C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, "Long-range correlations in nucleotide sequences," Nature **356**, 168–170 (1992).
[52]A. S. Mikhailov and A. Loskutov, *From Chaos to Noise* (Springer, Berlin, 1996).
[53]P. W. Anderson, "Suggest model for prebiotic evolution. The use of chaos," Proc. Natl. Acad. Sci. U.S.A. **80**, 3386–3390 (1983).
[54]F. Jacob, S. Brenner, and F. Cuzin, "On the regulation of DNA replication in bacteria," Cold Spring Harb Symp. Quant Biol. **28**, 329–438 (1963).
[55]T. Miyata and H. Hayashida, "Extraordinarily high evolutionary rate of pseudogenes: Evidence for the presence of selective pressure against changes between synonymous codons," Proc. Natl. Acad. Sci. U.S.A. **78**, 5739–5743 (1981).
[56]M. Touchon and E. P. C. Rocha, "Causes of insertion sequences abundance in prokaryotic genomes," Mol. Biol. Evol. **24**, 969–981 (2007).
[57]Y. Ejima and L. Yang, "Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling," Hum. Mol. Genet. **12**, 1321–1328 (2003).
[58]J. B. Pereira-Leal, E. D. Levy, and S. A. Teichmann, "The origins and evolution of functional modules: lessons from protein complexes," Philos. Trans. R. Soc. London, Ser. B **361**, 507–517 (2006).
[59]See EPAPS Document No. E-CHAOEH-18-003801 for figures illustrating the human chromosome Y with different releases, the inversion polymorphism in the human and chimpanzee genomes, as well as two sites obtained from Ref. 49. This document can be reached through a direct link in the online article's HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).