Doklady Physics, Vol. 50, No. 12, 2005, pp. 634–638. Translated from Doklady Akademii Nauk, Vol. 405, No. 6, 2005, pp. 755–759.

PHYSICS

Genome as a Two-Dimensional Walk

S. A. Larionov, A. Yu. Loskutov, and E. V. Ryadchenko

Presented by Academician A.R. Khokhlov April 22, 2005

Received April 26, 2005

The problem of the identification and biological significance of chromosome fragments and complete genomes is approached on the basis of the representation of a sequence of DNA nucleotides as a two-dimensional walk. Self-similarity properties have been analyzed; similar fragments of chromosomes, as well as some known functional and structural elements, have been distinguished. Completely and partially decoded chromosomes have been considered; in particular, fragments of the 22nd chromosome of a human and a chimpanzee have been compared.

It is well known that 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. Molecules of DNA contain two different purines, namely, adenine (*A*) and guanine (*G*), as well as two pyrimidines, namely, 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 string of characters drawn from the so-called nucleotide alphabet *ATTGCCAA*… and 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.

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 DNA is to carry, process, 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 very same sequence; this imposes specific restrictions on the organization of DNA.

The organization of sequences of various DNA fragments and their functional meaning is currently an important and urgent problem. The point is that, by now, a considerable number of sequenced chains of genomes have been obtained; however, the functional organization of these sequences has yet to be explained.

In this paper, we suggest a method that makes it possible to present the whole chromosome (even if it contains more than one million nucleotides) in a compact form, to easily find similar fragments, to identify functional and structural elements, and to detect the self-similarity of some fragments of the DNA sequence. The method is based on the representation of DNA as a plane walk of a particle.

Represent the sequence of nucleotides as a plane walk on a square lattice starting from the origin (0, 0) in the following way. Read the nucleotide chain in the order of appearance of the bases *A*, *T*, *G*, and *C*. In encountering adenine (*A*), make a step right, when thymine (*T*), a step left, when guanine (*G*), a step up, and when cytosine (*C*), a step down. Denote these coordinates by *AGTC* moving counterclockwise from the *x*-axis. Then, the original sequence of nucleotides corresponds to a certain *walking trajectory* on the plane *AGTC*. This representation of DNA is composed of two sequences *A–T* and *G–C*, which cannot be reduced to each other. The sequences may be considered separately and, moreover, may be represented as time series. The series, in turn, may be studied by well-known methods of calculus, such as wavelet transformation.

This method seems to be mentioned for the first time in 1962 by S.W. Golomb, one of the pioneer investigators of the genome, in [1], where he represented the DNA sequence on the complex plane by associating the nucleotide types with the coordinate vectors. However, at the time, the DNA code had not been discovered in full. Twenty years later, small sequences of the decoded DNA were considered as plane walks [2, 3]. There, the choice of the coordinates *G–C* and *A–T* was determined by considerations of the complementarity of the strands by the balance of hydrogen bonds along the strand. This

Moscow State University,
Vorob'evy gory, Moscow, 119992 Russia
e-mail: Loskutov@chaos.phys.msu.ru

1028-3358/05/5012-0634$26.00

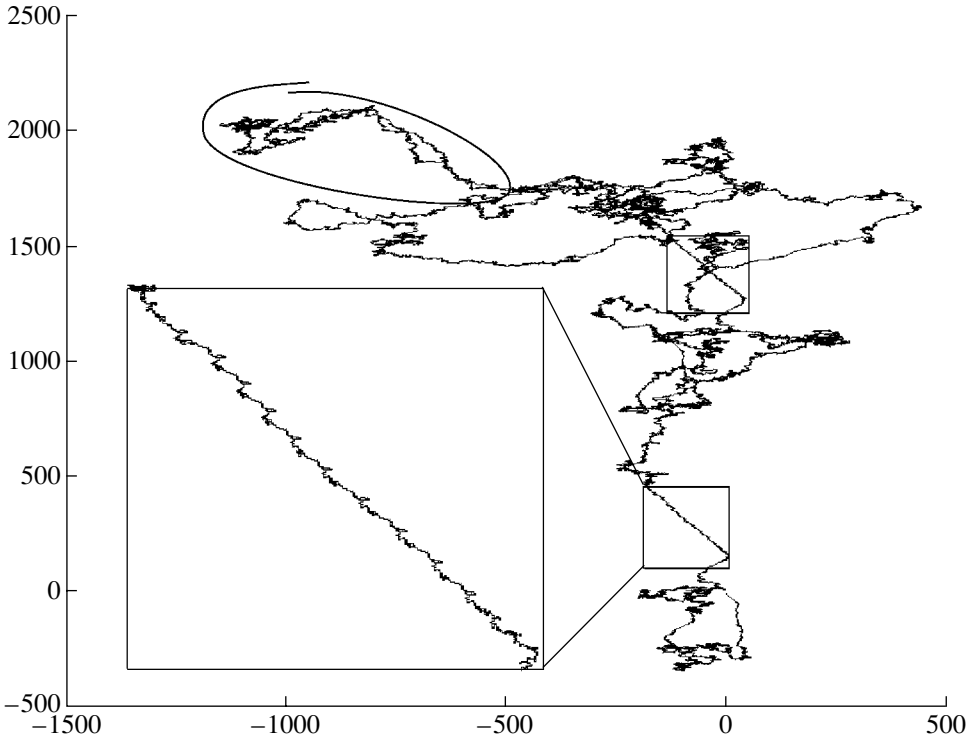

Fig. 1. *AGTC*-representation of the first chromosome of *S. cerevisiae.*

approach was subsequently provided with a strict justification (see, for instance, [5]).

Obviously, there are many sequences that may be considered as plane walks; then, there arise fractal structures that may be studied, etc. (see [6, 7] and references therein). However, as applied to DNA, functional and structural fragments of a chromosome may be distinguished by a typical walk "pattern" only for a sufficiently large number of units. Moreover, the number of units in the sequence of nucleotides may be of the same order of magnitude as the length of the sequence of the whole chromosome. Earlier, it seemed impossible to perform identification by this method with the use of only small fragments of chromosomes. The other methods of identification based on the alignment algorithms are rather labor-intensive and are not so demonstrative. It is from this viewpoint that *ATGC* sequences are considered in this paper.

Moreover, modern computer techniques allow automated processing by this method. One tentative attempt has already been made by a group of researchers (see [4]), who considered a somewhat different version of the plane walk. The authors formally used the method of a two-dimensional walk to construct an algorithm for comparing sequences. However, typical walk patterns, which might be crucial for understanding the organization of the structure of the sequence of chromosomes and its properties, had not been analyzed.

Today there are only a few species of living organisms for which the sequences of DNA of all their chromosomes have been completely decoded. Among them is the yeast cell *Saccharomyces cerevisiae*, by whose example we illustrate the analysis of the *AGTC* map.

Figure 1 shows the first of its 16 chromosomes, which contains approximately 230000 nucleotides. Almost identical large fragments (the square selections) are seen by the naked eye. Moreover, it follows from the construction that these fragments are passed in the opposite directions. This suggests that the fragments are complementary. Note that the length of each fragment is approximately 4000 units, which means that the use of another method (such as the alignment) for distinguishing these fragments would require incomparably greater investigative resources. These fragments are representatives of the family of flocculation genes FLO1 and FLO9 in the subtelomeric region.

Consider one of these fragments in more detail (see the insert to Fig. 1). It is seen to have an almost periodic spatial structure. Decompose this fragment into components (*A*–*T*) and (*G*–*C*). The wavelet transformation applied to these components explicitly shows that the selected fragment also possesses the property of self-similarity.

Employing the method of two-dimensional walk, one can easily find huge palindromes with a considerable share of pseudorandom inclusions. One of these is selected at the top left of Fig. 1. Its total length

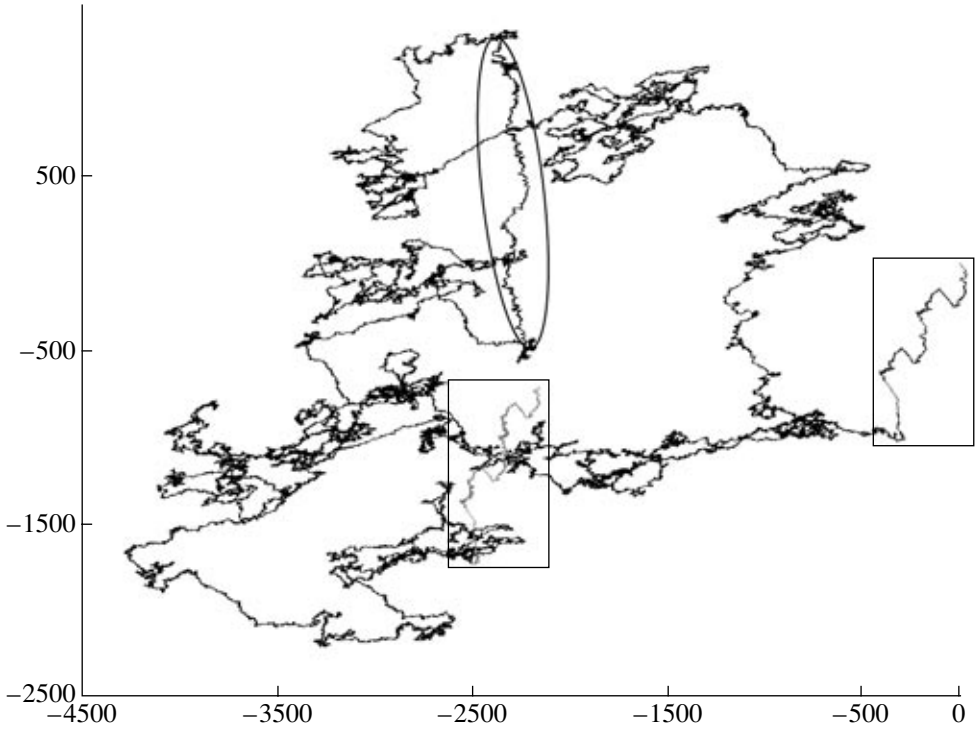

Fig. 2. Telomeres (the rectangles) of the 12th chromosome of *S. cerevisiae* and the cluster of ribosomal RNAs (the ellipse).

is 35000 nucleotides. Such a fragment cannot be identified by other methods; and even if it could, it would then require, at the least, an additional careful examination. This is the region with mobile genome transpositions represented by the *Ty*-family of retrotransposons.

Moreover, it is easy to identify telomeres. Figure 2 shows the 12th chromosome of the same yeast cell *Saccharomyces cerevisiae*. The telomeres are located at its ends; they are selected by rectangles. Each fragment has a size of approximately 20000 nucleotides. It is not difficult to visually detect similar fragments in other chromosomes without performing any statistical analysis. Note that these fragments are complementary.

The most indicative elements of this representation are the fragments in which the trajectory concentrates within a certain domain of the *ATGC* plane, skews, and long curved fragments of various shapes (see Figs. 1 and 2).

It is obvious that, throughout long skews, certain nucleotides dominate in the sequence. This is clearly seen in Fig. 2, where such a fragment is selected by an ellipse (here, the cluster of ribosomal RNAs is located). What is identified here is either satellite sequences or the averaged selected concentration of nucleotides without specific motifs, which may be seen by scaling up the observed fragment. In view of the different concentrations of purines and pyrimidines, the slope of this fragment suggests that the complementary strands of the DNA helix are of unequal weight (size) and anisotropic. Therefore, it appears interesting to investigate the length and directional distribution of the skews.

Fix a frame of the size N nucleotides with due regard for the appropriate scale and consider the motion of this frame along the walking trajectory on the *AGTC*-plane with the step of one nucleotide. Construct the diagram of such a motion in the coordinate system *AGTC* as follows. At each step of the motion, fix the radius vector that joins the beginning and the end of the frame. This radius vector is characterized by its length and direction. Plot the end of this radius vector on the diagram. Then, moving the frame by one step, we obtain a new radius vector. Plot its end on the diagram and proceed further in the same way. We thus obtain a plane diagram (Fig. 3). Each point i of this diagram ($0 < i < M - N$; M is the number of nucleotides in the sequence; and N is the size of the frame) shows how far and in which direction the representative point has moved from the beginning of the ith segment in N steps. This construction makes it possible to easily distinguish fragments of the domination of certain nucleotides. In particular, the outliers (see Fig. 3) characterize the direction and the length of the observed skews. Obviously, for each sequence, there exists a specific indicative size of the frame.

Finally, with the use of the above representation, one can easily compare sufficiently long fragments of chromosomes of different organisms. For a human and a chimpanzee, such a comparison has recently been made by a team of authors [8] by means of alignment.

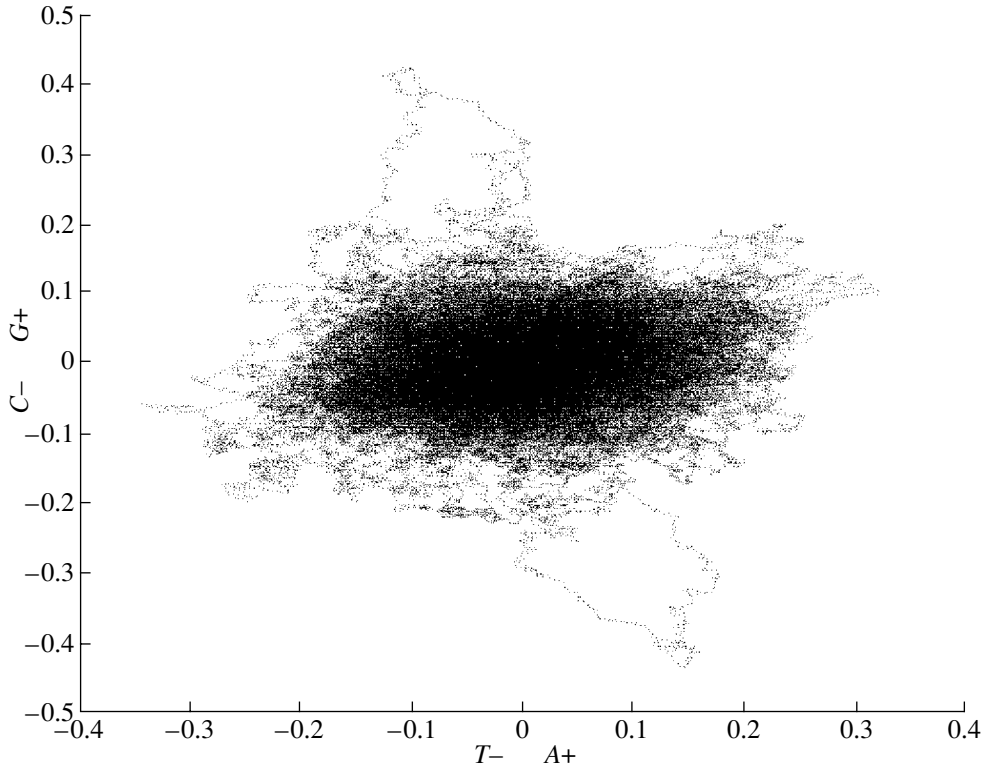

Fig. 3. Normalized length distribution of the 12th chromosome of *S. cerevisiae.*

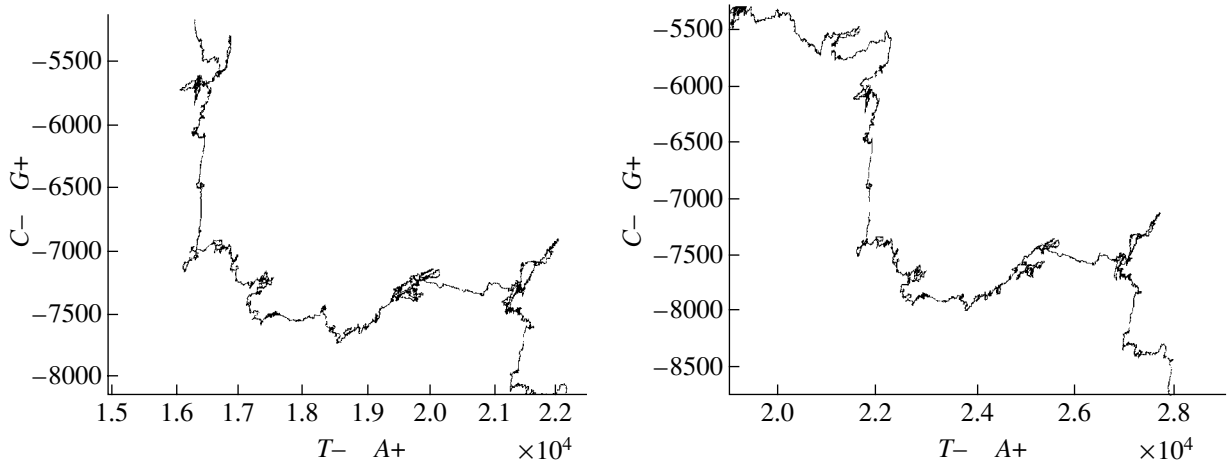

Fig. 4. Fragment (1934000–2134000) of the human 22nd chromosome (on the left) and the fragment (2176500–2376500) of the chimpanzee 22nd chromosome (on the right). They are seen to be almost identical.

Although the method of alignment constants is rather efficient, detection of similar fragments turns out to be quite labor-intensive, especially with account for the large number of missing fragments. At the same time, the above-described method enables us to easily render the general sequence fragmented for further alignment. Figure 4 presents the fragment (1934000–2134000) of the human 22nd chromosome compared with the fragment (2176500–2376500) of the chimpanzee 22nd chromosome. These fragments are seen to be almost identical.

The application of the above-described treatment of DNA as a random walk is much broader than simply the identification and comparison of fragments. Combined with other modern methods (such as fractal and Fourier analysis [9], wavelet transformation [10], the sliding

window method, etc.), it enables one to carry out a detailed fragmenting and, being rather demonstrative, becomes an instrument for advancing and testing the hypotheses on the organization of complete genomes and their properties.

REFERENCES

1. S. W. Golomb, in *Mathematical Problems in the Biological Sciences*, Ed. by R. Bellman (Amer. Math. Soc., Providence, 1962; Mir, Moscow, 1966).
2. M. A. Gates, Nature **316**, 219 (1985).
3. M. A. Gates, J. Theor. Biol. **119**, 319 (1986).
4. P. Vincens, L. Buff, C. Andre, *et al.*, Bioinformatics **14**, 715 (1998).
5. Ch. Berthelsen, J. A. Glazier, and M. H. Skolnik, Phys. Rev. A **45**, 8902 (1992).
6. G. Abramson, P. A. Alemany, and H. A. Cerdeira, Phys. Rev. E **58**, 914 (1998).
7. A. Rosas, E. Nogueira, and J. F. Fontanari, cond-mat/0209396.
8. H. Watenabe, A. Fujiyama, M. Hattori, *et al.*, Nature **429**, 382 (2004).
9. V. V. Lobzin and V. R. Chechetkin, Usp. Fiz. Nauk **170**, 57 (2000) [Phys. Usp. **43** (1), 55 (2000)].
10. A. Arneodo, Y. D. Aubenton-Carafa, B. Audit, *et al.*, Physica A (Amsterdam) **249**, 439 (1998).

Translated by A. Pankrat'ev