

Mammalian chromosomal DNA digestion with restriction endonucleases in silico

Category: [Study of Genome DNA](#)

Hits: 41330

M.A. Abdurashitov, [V.N. Tomilov](#), V.A. Chernukhin, D.A. Gonchar, S.Kh. Degtyarev

Translated from "Ovchinnikov bulletin of biotechnology and physical and chemical biology" V.2, No 3, pp 29-38, 2006

A theoretical method to produce digestion patterns of mammalian chromosomal DNA cleavage by restriction endonucleases was proposed. Based on recently published data on primary structures of genomes, a computer analysis was performed and diagrams of chromosomal DNA fragments distribution were plotted for DNA cleavages at 5'-GGCC-3', 5'-CCGG-3', 5'-GATC-3' and 5'-CC(A/T)GG-3' sequences. Experiments on chromosomal DNA digestion with HaeIII, MspI, Kzo9I and Bst2UI restriction endonucleases, which recognize these sites, were carried out. A correspondence between the computed diagrams and experimentally observed patterns of restriction enzymes cleavage was shown.

DNA cleavage with restriction endonucleases is a widely used method in molecular biological studies and one of the most important instruments of DNA research. Usually, DNA digestion products are analyzed by gel electrophoresis in agarose or acrylamide gels, and the observed pattern of DNA fragments separation, which is specific for each enzyme, is the result of DNA cleavage analysis [1]. A large variety of the known restriction endonucleases (RE) allows to carry out DNA digestion at more than 150 recognition sites [2]. Restriction enzymes analysis is performed for various DNAs, starting from small fragments of tens of nucleotide pairs and up to whole eukaryote genomes of more than 2-3 billion base pairs.

One of the most important elements of restriction enzymes cleavage and analysis is a construction of theoretical pattern of DNA fragments separation in gel. This pattern is determined on the basis of the known primary structure of given DNA. In the case of relatively small DNA (up to 100 000 - 200 000 bp), Vector NTI, Lasergene, Dnasis Max and other software are generally used on a Pentium or Athlon processor based personal computer. However, work with eukaryotic DNAs, primary structures for which were determined in the past decade, requires the availability of computation centers and special software. In this paper we present a simple method for analyzing primary structure of mammalian genomes and for plotting diagrams of chromosomal fragments distribution after DNA cleavage at definite nucleotide sequences. A common Athlon 64 based personal computer may be used to apply this method.

The goals of this work were: 1) to develop software for analysis of large DNA sequences *in silico*; 2) to plot theoretical diagrams of rat, mouse and human chromosomal DNA fragments distribution after DNA cleavage at recognition sites of some restriction enzymes; 3) to obtain experimental data on chromosomal DNAs hydrolysis with respective restriction endonucleases and then compare the observed results with theoretical calculations.

MATERIALS AND METHODS

Software

Borland Delphi (interface, output, treatment) and Microsoft C++ compiler (search routines, multi-threading) were used to develop the software. The final product was optimized to work on systems with two or more processors (the calculations were performed on AMD Athlon64 Core Duo processor).

DNA samples

Male SD rats aged 3-4 months and male A/He mice aged 5-6 months (Breeding Laboratory of Experimental Animals, Institute of Cytology and Genetics, Novosibirsk) were used in the experiments. Genomic DNA from the animal liver was isolated as described previously [3]. Isolation of genomic DNA from human blood leukocytes was carried out according to [4]. Data sampling was performed using not less than three experiments.

Before hydrolysis reaction, all DNA preparations were treated with ribonuclease A (0.1 mg/ml) for 10 minutes at a room temperature and dialyzed in DispoDialyzer MWCO 50,000 tubes ("Sigma", USA) TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA) 100 times of the volume of single DNA preparation at 4°C for 20 hours.

DNA hydrolysis with enzymes

Restriction endonucleases manufactured by SibEnzyme Ltd. were used in the work. DNA cleavage was performed in buffers recommended by the manufacturer at optimal temperatures for 3 h.

Hydrolysis reactions were performed in 40 µl of the incubation mixture containing 6 µg of DNA and 3 µl of enzyme preparation. The enzyme preparations had the following activities: *HaeIII* - 20 u/µl, *Kzo9I* - 5 u/µl, *MspI* - 10 u/µl, *Bst2UI* - 10 u/µl.

Electrophoresis

Electrophoresis in 8% polyacrylamide gel was used to detect DNA fragments from 40 to 500 bp (6 µg of hydrolyzed DNA was applied on gel in each run); 1.5% of low melting point agarose ("Sigma", USA)

was used to separate DNA fragments in the range of 200-2000 bp. Electrophoresis in 1% agarose "Type I-A, Low EEO" ("Sigma", USA) was used to detect DNA fragments of higher molecular weight. 3 µg of hydrolyzed DNA was applied on agarose gel in each run. Tris-acetate buffer was used for electrophoresis in all the cases. After electrophoresis, DNA was visualized with ethidium bromide and photographed in UV light.

RESULTS AND DISCUSSION

The development of the computer analysis and data treatment method

Over the last decade, nucleotide sequences of mouse [5], rat [6] and human [7] chromosomal DNA were determined. Data on the DNA primary structure can be found on the web sites [Entrez Genome](#) , [EMBL Genomes Pages](#) and [Ensemble Genome Browser](#). The primary structures of rat, mouse and human DNAs have been determined by more than 90%, and new data are constantly appearing to increase the portion of known mammalian DNA sequences. In this work we used data on rat, mouse and human genomic DNA sequences from the resource <ftp://ftp.ensembl.org/pub/> (version of June 2, 2006).

In order to develop a program, which would locate recognition sites of restriction endonucleases within extra long DNA sequences and compute DNA fragments lengths formed at these sites cleavage, the following peculiarities were taken into account:

1. Some restriction endonucleases recognize two or more DNA sequences (non-palindrome and/or degenerative recognition sites). In these cases, a search was performed for all possible recognition sites of these restriction endonucleases, including only the main symbols which corresponds to the designations of nucleotide residues ('A', 'G', 'T', 'C').
2. Genomic sequences, available in databases, are a set of lengthy fragments of chromosomal DNA (the so called "contigs"). In most cases, the lengths of these contigs make it impossible to load them wholly into the RAM of an ordinary personal computer. Due to this fact, "contigs" were loaded into the RAM in sections, formatted into string variables. The search for restriction endonucleases was carried out in these sections, which were gradually loaded into the search program.
3. Part of genomic sequences, located on both sides of the contigs was of unknown primary structure. Besides, in order to simplify and accelerate the analysis, all contigs were considered as a single extremely long sequence. Due to these limitations additional and non-existing fragments can appear; however, the significantly small number of such occurrences allows us to propose the feasibility of using this approach.

Computer simulation algorithm included the following stages:

1. The analysis of recognition site and the selection of an optimal method of search (depending on the site degeneration). The following methods of search were used according to the results of recognition site analysis: direct search method, method of substitution or method of substitution with grouping.
2. Data input and search for recognition sites. Due to the large size of the whole genome, data from the file of the genome DNA sequence were loaded into the memory in consecutive order, without service symbols. The main algorithm performed the following functions:
 1. loaded a certain data segment into the buffer,
 2. performed optimized search within it;
 3. added the result and deleted the treated fragment from the buffer in the case of successful search (when the recognition site in the loaded segment was found) , or, in the case of failure (site not found), loaded a new data segment from the genome file, taking into account the boundary effect, i.e. added a section of the old fragment with the length $n-1$ (where n is the length of the analyzed sequence of the recognition site) to the new segment.
3. Computation of the fragments lengths between the nearest recognition sites, calculation of the number of fragments for each length in the range of 1-6000 bp and of the total number of bp in all fragments of the particular length. These data were exported to a file in CSV (comma separated values) format for a further treatment. Calculated data were then imported to Microsoft Excel table and used to plot diagrams.

Thus, the distribution of the number of fragments obtained in the course of chromosomal DNA cleavage at recognition sites of restriction endonucleases depending on the sizes of these fragments, was computed in the simulation process for each recognition sequence. In these distribution diagrams, the value on the ordinate axis was the total number of base pairs in all fragments of the indicated length. This value was calculated using formula:

$$S_i = i \cdot n_i$$

where S_i is the total number of bp, i is the number of bp in the fragment and n_i is the number of fragments containing i nucleotide pairs. In the experiments, S value should be proportional to the intensity of the DNA fragments bands stained by ethidium bromide after separation in gel-electrophoresis.

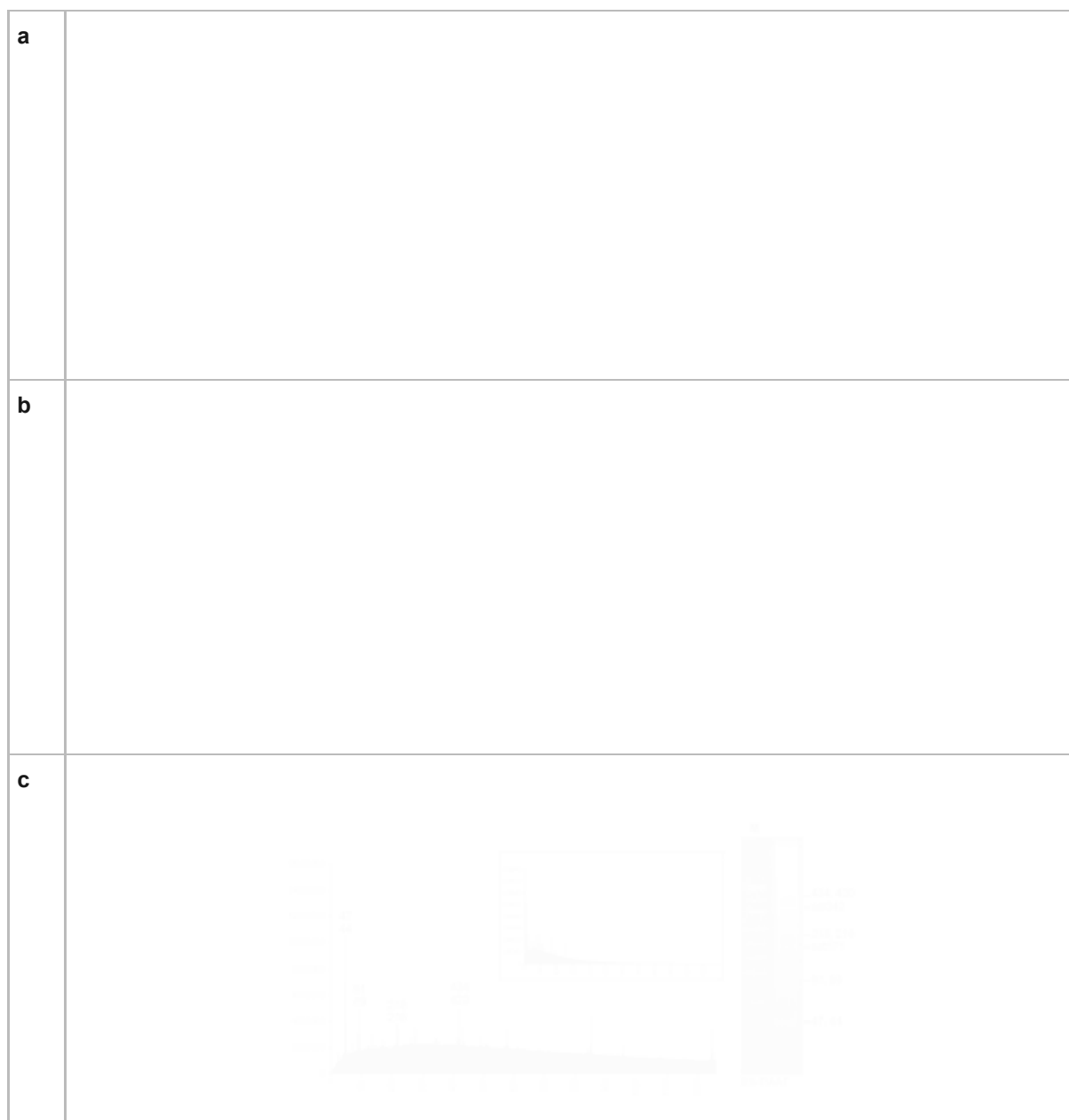


Fig. 1. Distribution diagrams of total DNA fragments lengths (expressed in base pairs) depending on the fragment size for rat (**a**), mouse (**b**) and human DNAs (**c**) cleavage at 5'-GGCC-3' sequence. Shown on the right, there are the experimentally obtained patterns of respective DNA cleavage by restriction endonuclease *HaeIII* (5'-GGCC-3'recognition site). The gel type and concentration are indicated below the picture. **M** - for agarose gels "SE 1 kb" DNA fragment length marker was used with following fragment lengths - 10000, 8000, 6000, 5000, 4000, 2x3000, 2500, 2000, 1500, 2x1000, 750, 500 and 250 bp); pUC19/*MspI* DNA ladder was used for PAAG (fragment lengths - 501, 489, 404, 331, 242, 190, 147, 111, 110, 67, 2x34 bp).

a	
b	
c	

Fig. 2. Distribution diagrams of total DNA fragments lengths (expressed in base pairs) depending on the fragment size for rat (**a**), mouse (**b**) and human DNAs (**c**) cleavage at 5'-CC(A/T)GG-3' sequence. Shown on the right, there are the experimentally obtained patterns of respective DNA cleavage by restriction endonuclease *Bst2UI* (5'-CC(A/T)GG-3'recognition site). The other designations are similar to those shown in Fig. 1.

Figures 1-4 present the obtained diagrams of S_i distribution for rat (1a, 2a, 3a and 4a), mouse (1b, 2b, 3b and 4b) and human (1c, 2c, 3c and 4c) DNA fragments at cleavage at 5'-GGCC-3' (Fig. 1), 5'-CC(A/T)GG-3' (Fig. 2), 5'-GATC-3' (Fig. 3) and 5'-CCGG-3' sequences (Fig. 4). The diagrams show computed data for the range of DNA fragments 1-1200 bp and calculated data for the range of 1-6000 bp in insertions (in a reduced scale). Fragments lengths (bp) are given on the abscissa axis and the total number of bp in all fragments of this length is indicated on the ordinate axis.

a	
b	
c	

Fig. 3. Distribution diagrams of total DNA fragments lengths (expressed in base pairs) depending on the fragment size for rat (**a**), mouse (**b**) and human DNAs (c) cleavage at 5'-GATC-3' sequence. Shown on the right, there are the experimentally obtained patterns of respective DNA cleavage by restriction endonuclease *Kzo9I* (5'-GATC-3' recognition site). The other designations are similar to those shown in Fig. 1.



Fig. 4. Distribution diagrams of total DNA fragments lengths (expressed in base pairs) depending on the fragment size for rat (**a**), mouse (**b**) and human DNAs (**c**) cleavage at 5'-CCGG-3' sequence.

As it can be seen from Figures 1-4, all of the obtained distribution diagrams show sharp extremes for certain lengths of DNA fragments, whose sizes (in base pairs) are indicated by numbers above the peaks. Sizes of these fragments, which show peak values, vary from 40-50 bp for those obtained for human DNA cleavage at 5'-CC(A/T)GG-3', and 5'-GGCC-3' sequences to 5538 bp for that obtained for mouse DNA cleavage at 5'-CCGG -3' sequence. The values of observed peaks also essentially vary from insignificant values to the maximum height of approximately 15000000 bp for human chromosomal DNA cleavage at 5'-GATC-3' (174 and 175-bp fragments) and 5'-CC(A/T)GG-3' (49 and 50-bp fragments) sequences. Presence of peaks in diagrams is probably associated with the existence of the so called repeats in eukaryotic DNAs [8]. The sizes of such repeats vary from several bp

(microsatellites) to 6-7 thousand bp (LINE family repeats), and their number in the genome can reach 500 thousand and more [5-7]. Cleavage of a large number of such repeats at recognition sites of restriction endonucleases will result in the appearance of a large number of fragments of the same length. As a result, in the diagram we can observe peak with a height which is proportional to the number of the fragment repeats in chromosomal DNA.

Determination of nucleotide sequences of eukaryotic genomes, in particular those of rat, mouse and human genomes, made the construction of genomic DNA fragments distribution diagrams to be possible. These diagrams, in essence, are the restriction enzymes analysis of mammalian DNA *in silico*. The presence of gaps and possible errors in already determined DNA sequences are not critical moments in the work. These defects are less than 1% of the determined primary structure [9] and can not significantly change the DNA cleavage pattern due to the large number of DNA repeats in the genomes.

Experimental patterns of chromosomal DNA cleavage

In order to evaluate the proposed method of restriction enzymes analysis of mammalian DNA *in silico*, we carried out experiments on the rat, mouse and human DNA cleavage by restriction endonucleases with respective recognition sites.

Figure 5 illustrates results of the electrophoresis in 1% agarose gel of the products of rat, mouse and human DNA hydrolysis with restriction endonucleases *Kzo9I* (recognition site GATC), *HaeIII* (GGCC), *Bst2UI* (CC(A/T)GG) and *MspI* (CCGG). The other obtained patterns of DNA digestion with restriction enzymes are presented on the right side in Figures 1-3.

As it can be seen from the data presented in Figures 1-3 and 5, in most cases, clearly seen fragments of certain length are formed at cleavage of chromosomal DNAs by restriction endonucleases. Aside from the typical pattern of DNA restriction fragments, individual bands are seen substantially brighter in Fig. 1c and 2b. Sizes of these fragments correspond to the lengths of basic repeats in alpha-satellite DNAs (234 bp - mouse [8], 342 bp and 171 bp - human [10]). In gel photos presented in Fig. 1 and 2, these bands are denoted by the prefix "sat". Evidently, higher intensity of these bands of satellite DNAs is associated with multiply repeated tandem sequences, which present in DNA preparations in quantities considerably exceeding those of LINE1 repeats in lengthy chromosomal DNAs. Cleavage of alpha-satellite DNA with restriction enzyme results in formation of fragments corresponding to one or more basic repeats.

It is known that purification of lengthy chromosomal DNA of large sizes, and in particular from eukaryotes, is problematical due to mechanical destruction of native DNA when using most commonly known techniques. Nevertheless, patterns of restriction fragments, presented in Fig. 1-3 and 5, show that the routine phenol-chloroform method may be used in preparation of DNA samples, which are suitable for restriction enzymes analysis *in vitro*. With that, partial DNA degradation, which usually results in restriction pattern smearing, is compensated by a large number of dominating repeats in eukaryote genomes.

The comparison of theoretically calculated and experimental patterns of restriction analysis

Fig. 1 presents calculation data on cleavage of chromosomal DNAs at GGCC site. As it can be seen from Figures 1 and 5, good correlation can be observed between theoretical and experimental data. Fragments 575-758, 697, 863 and 1173 bp calculated for cleavage of the rat chromosomal DNA, correspond to fragments visible on the gel in Fig. 1a. It should be noted that 370-bp fragment of rat satellite DNA is clearly seen on the gel as well [11].

Fig. 5. Patterns of human, mouse and rat DNAs cleavage of by restriction endonucleases *Kzo9I* (GATC recognition site), *HaeIII* (GGCC), *Bst2UI* (CC(A/T)GG) and *MspI* (CCGG). Electrophoresis in 1% agarose gel. **M** - "SE 1 kB" DNA fragment length marker

The diagram of mouse DNA cleavage depicts 5 fragments of 185, 347, 1176, 1233 and 1906 bp. Among these fragments, only the latter has a high S_i value and is clearly seen on the gel. The 347-bp fragment and the duplicates of 1176 and 1223-bp fragments are also clearly seen in the gel photograph, whereas the 185-bp fragment is hardly observable. A more complex pattern is observed for human DNA cleavage, where aside from calculation fragments with the lengths of 44-47, 89-91, 216-218, even brighter fragments can be observed. These fragments are probably cleavage products of alpha-satellite DNA with the lengths of 171 and 342 base pairs (Fig. 1c).

Data presented in Fig. 2 show that the rat chromosomal DNA cleavage at CC(A/T)GG sequence gives smaller fragments than at GGCC site, however, all of these fragments with size up to 462 bp, are seen on polyacrylamide gel. DNA fragments (713, 1017 and 1690-1691) with higher molecular weight are seen on agarose gel in Fig. 5. The mouse DNA cleavage at CC(A/T)GG site also reveals a set of fragments visible on polyacrylamide gel, as well as two larger 1511 and 1826-bp fragments, which can be seen in Fig. 5. We should also note a presence of a bright band corresponding to a fragment of 234-bp satellite DNA (Fig. 2b). Human DNA cleavage at CC(A/T)GG site gives a unique set of a number of small fragments, which are presented in pairs and are therefore especially clearly seen in the gel photograph in Fig. 2c. The correlation between obtained experimental and theoretical results can be

observed in both the correspondence of the computed and observed fragments number and lengths, as well as in their intensity. As can be seen in the diagram in Fig. 2c, the intensity of the pair of 116, 118-bp fragments is considerably higher than that of 166, 167-bp fragments pairs. This fact corresponds to the electrophoresis data.

The results of rat and mouse DNA cleavage pattern simulation at GATC site, presented in Fig. 3, give a large set of small fragments, which can be seen in the photograph of the corresponding gels. Human chromosomal DNA cleavage gives only one essential peak of DNA fragments at 174-175 bp, also clearly seen in the gel picture.

Fig. 4 depicts diagrams of three DNAs cleavage at CCGG site, and Fig. 5 shows the picture of these DNAs hydrolysis with restriction endonuclease *MspI* (CCGG recognition site). On the rat DNA cleavage diagram, 2 peaks are revealed for 404 and 5537-5538-bp fragments, and bands corresponding to them are also clearly seen in the gel photograph. For mouse DNA, there are 614, 3727 and 4992 bp computed fragments; and in the experimental data, we can observe visible bands, which correspond to these lengths. In human DNA cleavage with restriction enzyme *MspI*, fragments with considerable intensity are neither on the computed diagram nor on the gel photograph.

Thus, as it can be seen from Figures 1-5, with the exception of satellite DNA fragments, most of experimentally obtained fragments lengths, correspond to the peak values on diagrams calculated with the proposed method. On the other hand, theoretically calculated fragments with high value of S_i function can be seen in the gel photographs.

There probably exists a certain threshold value of S_i function, below which the bands are poorly visualized on the gel. Under conditions of our experiments, the threshold value for fragments detection made up approximately 4 million bp in PAAG (approximately 0.13-0.15% of the whole genome length) and 5,5-6 million bp in agarose gel (approximately 0.17-0.21% of the whole genome length). It is necessary to take into account, that the intensity of bands on the gel corresponding to DNA fragments with approximately the same mobility, will increase due to the superposition effect. This effect can be observed in the case of fragments obtained at human DNA hydrolysis of with restriction enzyme *Bst2UI* (CC(A/T)GG recognition site). A similar effect is observed for the rat DNA hydrolysis with restriction endonuclease *MspI*, when high molecular weight fragments, with S function value on the diagram less than 4 million bp, are clearly seen on the gel.

Analysis of the gel photographs, presented in our work, and the theoretical diagrams of DNA fragments distribution clearly demonstrate an essential correlation between the computed data and the experimental results. Additional high intensity bands of alpha-satellite DNA are visible in both mouse DNA (Fig. 2b, 234-bp fragment), and human DNA hydrolosys products (Fig. 1c, 171-bp and 342-bp fragments). A similar band is far less visible in rat DNA hydrolysis (Fig. 1a, 370-bp fragment).

Subsequent works will describe separate studies on restriction enzymes analysis of rat, mouse and human DNA *in silico* at a wide range of recognition sites. They will also provide comparison of obtained data with experimental results on DNA preparations hydrolysis with respective restriction endonucleases.

CONCLUSION

The present work proposes a simple method of locating recognition sites of restriction endonucleases in mammalian genome sequences. This method allows to perform computations of the lengths and quantity of DNA fragments formed at the cleavage, as well as to construct their distribution diagrams.

The software used with this method does not require major server powers and makes it possible to work with lengthy nucleotide sequences of several billion bp on ordinary personal computers.

We believe that this method will be highly useful for the study of the mammalian genomes structure, and the list of site-specific endonucleases used to study eukaryote DNA, will be considerably extended. Though in our work we restricted ourselves to the analysis of only three mammalian genomes, it is obvious that the use of the proposed method of restriction enzymes analysis *in silico* is possible in cases of bacteria, fungi, plants genomes, as well as genomes of other organisms, for which the primary structure DNA has been determined.

The authors thank Dr. V. I. Kaledin Dr. G. V. Vasilyev for assistance in animal experiments and DNA isolation.

REFERENCES

1. Alberts B., Bray D., Lewis J., Raff M., Roberts K., Watson J. // Molecular biology of the Cell. - Moscow: Mir Publishers, 1994. - V. 1. - P. 232-233 (in Russian).
2. Roberts R.J., Vincze T., Posfai J., Macelis D. REBASE - Restriction enzymes and DNA methyltransferases // Nucleic Acids Res. - 2005. - Vol. 33. (Database issue). - D230-D232.
3. Dashkevich V.S., Dymshits G.M., Salganik R.I., Ulanov B.P. About the destabilization of the secondary structure of the DNA of regenerating rat liver in the process of replication // Molecular Biology. - 1972. - Vol.6 -No.5. -P. 689-697 (in Russian).
4. Lindblom B., Holmlund G. Rapid DNA purification for restriction fragment length polymorphism analysis // Gene Anal. Tech. - 1988. - Vol. 5. - P. 97-101.
5. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome // Nature. - 2002. - Vol. 420. - P. 520-562.
6. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution // Nature. - 2004. - Vol. 428. - P. 493-521.
7. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome // Nature. - 2001. - Vol. 409. - P. 860-921.
8. Lewin B. // Genes. - Moscow: Mir Publishers, 1987. - P. 298-308 (in Russian).
9. Finishing the euchromatic sequence of the human genome // Nature. - 2004. - Vol. 431. - P. 931 - 945
10. Willard H.F. Chromosome-specific organization of human alpha satellite DNA // Am. J. Hum. Genet. - 1985. - Vol. 37. - P. 524-532.
11. Pech M., Igo-Kemenes T., Zachau H.G. Nucleotide sequence of a highly repetitive component of rat DNA // Nucleic Acids Res. - 1979. - Vol. 7. - P. 417-432.

ACCEPTED ABBREVIATIONS

a.u. - activity unit

LMA - low melting point agarose

PAAG - polyacrylamide gel

bp - base pairs

RE - restriction endonuclease