

No Isochores in the Human Chromosomes 21 and 22?

David Häring and Jaroslav Kypr¹*Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, CZ-61265 Brno, Czech Republic*

Received December 11, 2000

The human genome is described in the literature as being composed of the isochores, i.e., long (hundreds of kilobases) segments with a homogeneous (G + C) content. We calculated the (G + C) content variations along the DNA molecules of the human chromosomes 21 and 22 and found the variations to be higher everywhere compared to the randomized sequences. Hence the (G + C) content is certainly not homogeneous on the isochore scale in the two human chromosomes. In addition, we found no significant difference between the two human molecules and the genome of *E. coli* regarding the (G + C) content variations. Hence no isochores are either present in the DNA molecules of the human chromosomes 21 and 22, or the isochores are also present in the genome of *Escherichia coli*. In any case, the present communication demonstrates that the isochores should be defined in unambiguous molecular terms if they are to be used for an up-to-date genome structure characterization. © 2001 Academic Press

Key Words: (G + C) content variations; isochores; human chromosomes 21 and 22; *E. coli*.

In the presence of silver or other DNA binding ligands, density gradient centrifugation separates nuclear DNA of warm-blooded vertebrates into components called isochores (1). The centrifugation analysis led to a widely accepted view that more than a half of the human genome is composed of two families of (G + C)-poor isochores while three families of (G + C)-rich isochores constituted the genome remainder (2). The isochores were postulated to be large (300 kb or longer) genomic DNA segments having a homogeneous (G + C) composition (3). Now the recently published (almost) complete nucleotide sequences of the human chromosomes 22 (4) and 21 (5) provide an opportunity to look at the isochores in detail and define their properties on the molecular level. This is what we wanted to do in this work. However, we were surprised to be unable to

detect any distinct (G + C) content homogeneity along the DNA molecules of the human chromosomes 21 and 22 on the isochore scale level. In addition, the (G + C) content variations or homogeneity was found to be essentially the same along the two human chromosomes as it was along the chromosome of *Escherichia coli*. These findings can be explained in various ways that are presented under Discussion of this communication.

MATERIALS AND METHODS

The complete genome sequences of *Escherichia coli* (6, the EMBL Accession Number U00096) and the sequences of human chromosomes 21 (5, GenBank ID: NT002836) and 22 (4, GenBank ID: NT_001039) were obtained from the EMBL (7) and GenBank (8) databases. Randomized sequences, serving as statistical controls, were generated using the RANDOM utility (9, 10).

The (G + C) content distributions were obtained with SEQSTAT (9, 10). The (G + C) contents were calculated in segments having 10 kb, 100 kb and 1 Mb in length. The segments overlapped by 3/4 of their length. The (G + C) content variation was defined as standard deviation divided by the average value of (G + C) content in segment of certain length (11). The distributions of (G + C) variation were calculated from (G + C) content distribution (based on 10 kbp segments) in 100 kbp segments overlapping by 3/4 of their length. Length of 100 kbp was the average length of the DNA fragments used in the experimental isochore identification (1–3).

Next we searched for contiguous chromosome regions where the (G + C) content variation was lower than the average value of (G + C) content variation in the corresponding randomized sequence. Further we searched for contiguous chromosome regions inside of which the (G + C) content fluctuated in the range of 1–3% of the total nucleotide content. The contiguous regions of homogenous (G + C) content were determined as follows (Fig. 1).

The first chromosome segment was chosen as the beginning of the homogenous region. The homogenous region was then extended with the neighboring (overlapping) segment if the range of (G + C) content fluctuation within the region, including the added segment, was smaller than or equal to the given limit. If so, the next neighboring (overlapping) segment was added to the region and so on until the fluctuations exceeded the limit. Then the same procedure was started from the second chromosome segment in order to find another homogenous region. Then the search started from the third chromosome segment etc. until the whole chromosome was processed. Only the homogenous regions of at least three times the segment length (i.e., consisting of at least 9 overlapping segments, Fig. 1) were retained for further analysis. In case there were several overlapping regions, the longest one was left intact and the shorter regions were truncated. Finally we constructed maps of the homog-

¹ To whom correspondence should be addressed. Fax: ++420541240497. E-mail: kypr@ibp.cz.

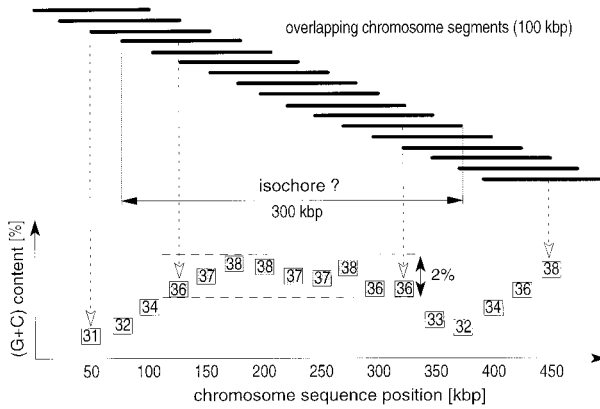


FIG. 1. Scheme of the search algorithm used to identify contiguous regions of homogenous (G + C) content, i.e., isochores. In this example the (G + C) content was calculated in 100 kbp chromosome segments overlapping by 3/4 of their length. The isochores had to have a minimum length of 300 kbp inside which the (G + C) content fluctuated within 2% of the (A + C + G + T) content.

enous (G + C) regions and calculated distributions of the total lengths of the homogenous regions according to their (G + C) content.

RESULTS

Isochores were postulated to be long (300 kb or longer) genome regions showing a homogeneous (G + C) content (3). Homogeneity implies low variations. That is why we calculated variations (the standard deviation divided by the average value) of the (G + C) content along the whole DNA molecules of the human

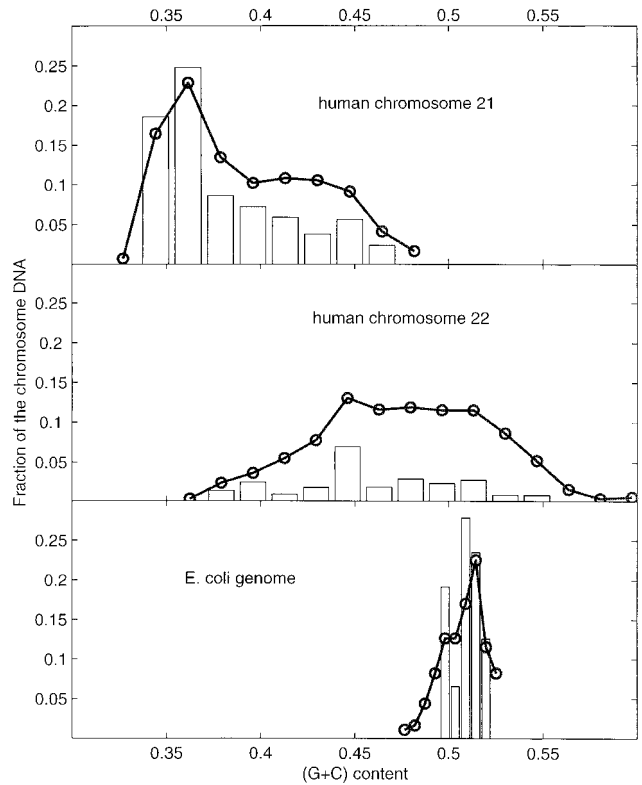


FIG. 3. Histograms reflect the amounts of (G + C) homogenous regions longer than 300 kbp depending on their (G + C) content in the human chromosomes 21, 22 and *E. coli* genome. The amounts of 100 kbp chromosome segments are represented by the thick curves.

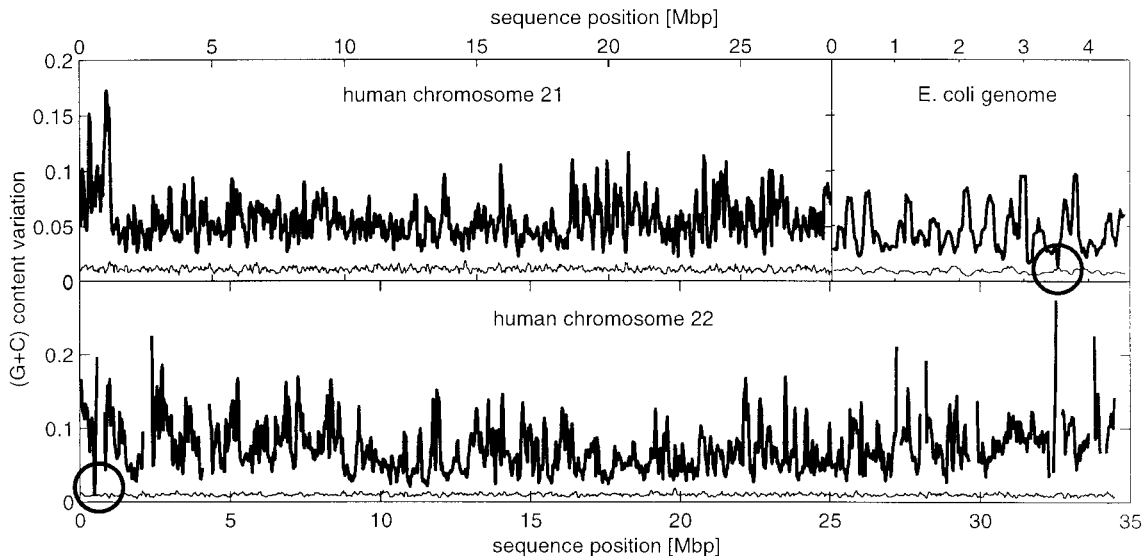


FIG. 2. Distribution of (G + C) content variation in the human chromosomes 21 and 22 and in the genome of *E. coli*. Bold and light curves represent (G + C) content variation along the native and the corresponding randomized chromosome sequences, respectively. The two positions where the (G + C) content variation is similar in the native and randomized sequences, are marked with circles.

TABLE 1

Contiguous Regions of the (G + C) Content Fluctuating within the Indicated Ranges in Human Chromosomes 21, 22 and *E. coli* Genome (the Lengths are in kbp)

Chromosome*	Regions \geq 30 kbp**			Regions \geq 300 kb***			Regions \geq 3 Mb****		
	Number	Longest	Total length	Number	Longest	Total length	Number	Longest	Total length
1% (G + C) fluctuation; (A + C + G + T) = 100%									
CHR21	11	40	345	19	675	7,300	1	7,000	7,000
CHR21-R	548	137.5	24,542.5	1	28,500	28,500	1	28,500	28,500
CHR22	9	80	357.5	3	375	1,050	0	0	0
CHR22-R	634	165	28,392.5	1	34,550	34,550	1	34,500	34,500
ECOLI	5	37.5	162.5	3	425	1,075	0	0	0
ECOLI-R	85	125	3,932.5	2	3,975	4,700	1	4,500	4,500
1.5% (G + C) fluctuation									
CHR21	65	52.5	2,187.5	37	875	16,175	3	7,250	14,750
CHR21-R	355	297.5	28,942.5	1	28,500	28,500	1	28,500	28,500
CHR22	55	82.5	1,945	12	575	4,325	0	0	0
CHR22-R	458	265	34,510	1	34,550	34,550	1	34,500	34,500
ECOLI	19	52.5	645	7	625	3,300	1	3,750	3,750
ECOLI-R	57	285	4,690	1	4,625	4,625	1	4,500	4,500
2% (G + C) fluctuation									
CHR21	181	62.5	6,385	44	1,400	21,850	3	7,500	16,250
CHR21-R	159	667.5	29,192.5	1	28,500	28,500	1	28,500	28,500
CHR22	150	122.5	5,327.5	24	625	8,700	1	3,000	3,000
CHR22-R	226	742.5	35,282.5	1	34,550	34,550	1	34,500	34,500
ECOLI	38	75	1,432.5	8	725	4,100	1	4,500	4,500
ECOLI-R	29	675	4,757.5	1	4,625	4,625	1	4,500	4,500
2.5% (G + C) fluctuation									
CHR21	338	77.5	12,440	42	1,775	24,600	5	7,750	23,750
CHR21-R	48	1,867.5	28,852	1	28,500	28,500	1	28,500	28,500
CHR22	243	122.5	9,082.5	34	900	13,100	1	3,250	3,250
CHR22-R	85	1,480	35,822.5	1	34,550	34,550	1	34,500	34,500
ECOLI	59	87.5	2,395	8	750	4,475	1	4,500	4,500
ECOLI-R	12	1000	4,917.5	1	4,625	4,625	1	4,500	4,500
3% (G + C) fluctuation									
CHR21	433	87.5	17,262.5	38	1,925	25,650	6	8,000	28,000
CHR21-R	8	7,827.5	28,555	1	28,500	28,500	1	28,500	28,500
CHR22	337	127.5	13,167.5	44	950	18,100	3	3,500	10,000
CHR22-R	25	4,765	35,362.5	1	34,550	34,550	1	34,500	34,500
ECOLI	79	107.5	3,340	6	1,200	4,675	1	4,500	4,500
ECOLI-R	6	1,810	5,227	1	4,625	4,625	1	4,500	4,500

* CHR21 denotes the native DNA sequence of human chromosome 21; CHR21-R denotes its randomized counterpart, etc.

** Regions consisting of 10 kb chromosome segments overlapping by 7.5 kbp.

*** Regions consisting of 100 kb chromosome segments overlapping by 75 kbp.

**** Regions consisting of 1 Mb chromosome segments overlapping by 750 kbp.

chromosomes 21 and 22 in order to identify the isochores. To our surprise, however, we saw no 300 kb or longer region in these DNA molecules whose (G + C) content variation was lower than in the corresponding randomized sequences. There only was a single 120 kb segment (at position 0.5 Mbp, Fig. 2) in the DNA of

human chromosome 22 whose (G + C) content variation was below the maximum (G + C) content variation found in the corresponding randomized sequence when the segments were 100 kb in length. Hence this result makes us conclude that, with respect to randomized sequences, the (G + C) distribution is not homogeneous

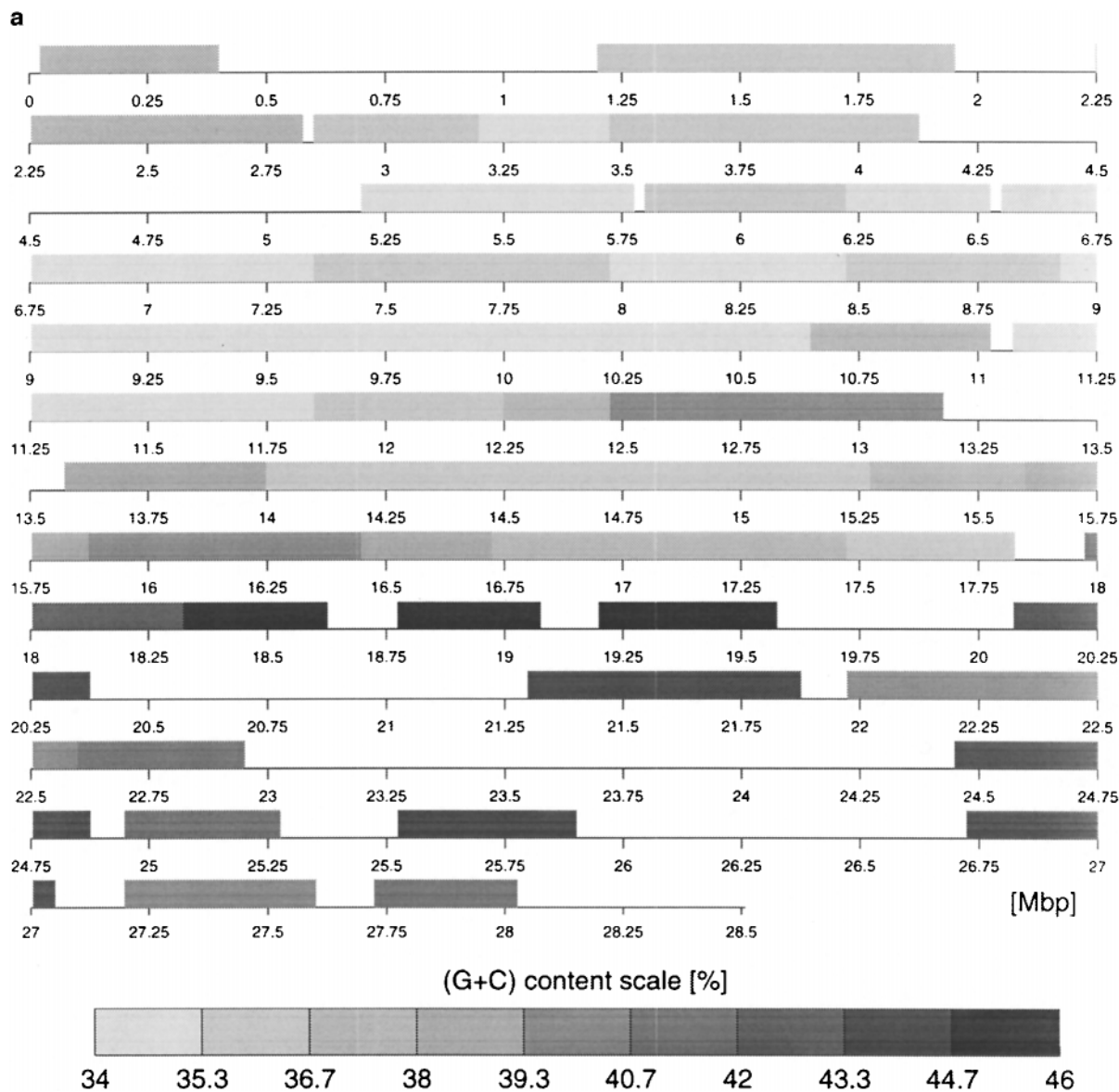


FIG. 4. Regions of homogenous (G + C) content longer than 300 kbp along the human chromosome 21 (a), 22 (b), and *E. coli* genome (c). The (G + C) content fluctuation within the homogenous regions is 2% of the (A + C + G + T) content. The regions are composed of 100 kbp chromosome segments overlapping by 3/4 of their length.

at all in any part of the human chromosomes 21 and 22 on the isochore length scale.

Next we calculated the (G + C) variations in the *Escherichia coli* genome as well, using the same approach, and found that the variations were almost identical as observed above with the human chromosomes 21 and 22 (there was only a single chromosome segment of (G + C) content variation lower than maximum (G + C) variation in the corresponding randomized sequence, at position 3.5 Mbp, Fig. 2). Hence we have to conclude that the isochores either occur in the *Escherichia coli* genome as well, or they do not occur in the human chromosomes 21 and 22, or the (G + C)

content variation as defined above is not the proper quantity to identify the isochores.

To elucidate the latter possibility, we divided the DNA molecules of the human chromosomes 21 and 22 into overlapping 10, 100 and 1,000 kbp segments, calculated the segment (G + C) contents and compared them along the chromosome DNA sequence as described under Materials and Methods. This approach yielded contiguous regions of chromosomal DNA inside which the (G + C) level oscillated within a narrow range. The same analysis was repeated with *E. coli* genome and with corresponding randomized sequences. The results demonstrate (Table 1) first of all

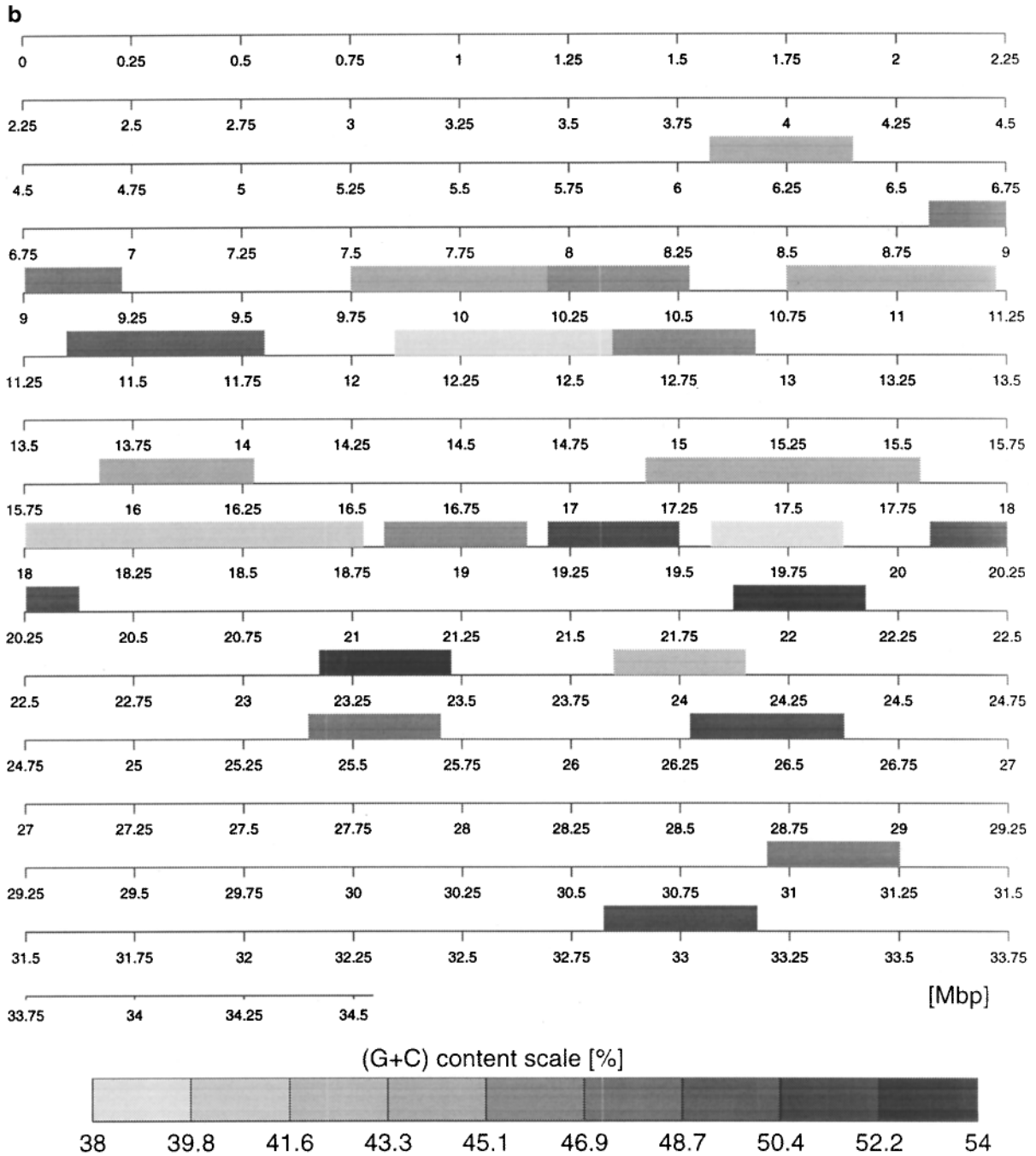


FIG. 4—Continued

that (G + C) content homogeneity is much more frequent in the randomized sequences than in the native sequences irrespective of whether the native sequence originates from human chromosome 21, 22 or the *E. coli* genome. For example, the total size of the homogenous (G + C) regions in the randomized sequences is at least four times the length of the homogenous (G + C) regions in the native sequences (oscillations within 1%, 100 kbp segments), regardless of their origin.

Histograms showing the distribution of (G + C) homogenous regions in the human chromosomes 21 and 22 and the genome of *E. coli* demonstrate (Fig. 3, oscillations within 2%, 100 kbp segments) that the most populated homogenous regions have (G + C) contents of 36, 45, and 51% in the human chromosomes 21, 22, and the *E. coli* genome, respectively. Distribution of these homogenous regions along the human chromosomes 21, 22 and the *E. coli* genome is shown in Fig. 4.

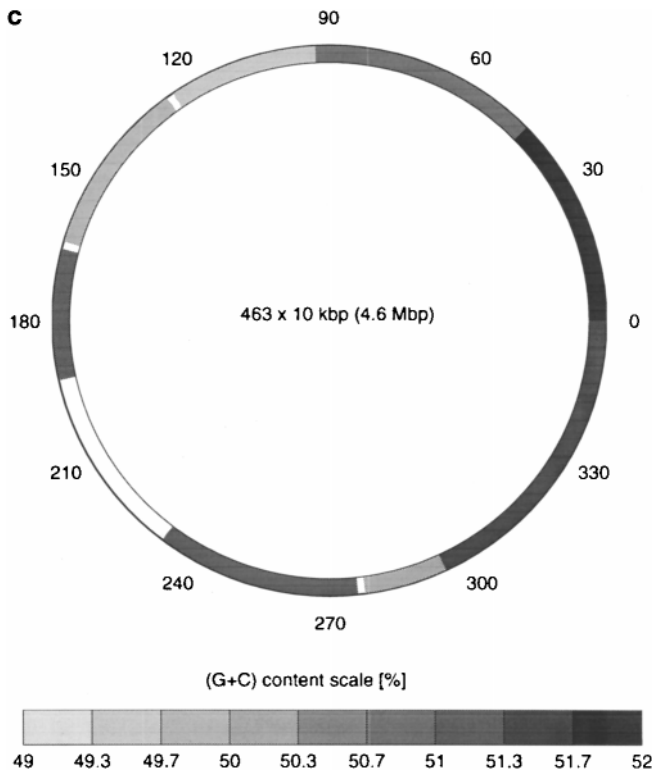


FIG. 4—Continued

The (G + C) homogenous regions occupy 77, 25, and 90% of the human chromosomes 21, 22, and *E. coli* genome, respectively.

DISCUSSION

DNA molecules constituting organism chromosomes are the material substances inherited in the process of evolution. Availability of their complete nucleotide sequences makes possible to analyze their structure and compare the structures of the DNA molecules of various chromosomes. In our opinion, this will be a more fruitful approach to trace their evolutionary relationships than the tedious and ambiguous identification and comparison of proteins the chromosomal DNA molecules code for.

One of the fundamental properties of the chromosomal molecules of DNA is their (G + C) content. It has a taxonomy value (12) and it determines amino acid composition of the encoded proteins (13). The (G + C) content also relates to codon usage in genes (14) and the succession of replication of various genomic regions (15). The (G + C) content distribution also stands behind isochores (1), i.e. long chromosomal DNA regions in higher vertebrates, where the (G + C) content level was postulated to be "homogenous" on the basis of centrifugation studies of the genomic DNA fragments

in the presence of silver or other DNA binding ligands (2, 3).

The recently published (almost) complete nucleotide sequences of the human chromosomes 21 (5) and 22 (4) provided the first possibility to see the isochores on the molecular level. Previous attempts to see the isochores only used relatively short sequenced parts of the chromosomes (16–19) or were focused on silent codon positions (20), which led to inconclusive results regarding the isochores.

The present results are conclusive but rather surprising because they show that while the analyzed chromosomes are different with respect to their ranges of (G + C) content, there is no significant difference between the DNA molecules of *E. coli* and the human chromosomes 21 and 22 regarding the homogeneity of their (G + C) distribution on the isochore length scale. Hence we have to conclude that no isochore either is present in any of the two analyzed human chromosomes, which is rather surprising, or that the isochores also occur in the genome of *E. coli* which is in conflict with the concept that the isochores are a specific property of higher vertebrates.

The chromosomes 21 and 22 only constitute about 2% of the human genome so that it is evident that the two smallest chromosomes will only marginally influence properties of the genomes as a whole, including the occurrence, absence or properties of isochores. Yet it is startling how much more inhomogeneous the (G + C) content distribution is along almost the entire lengths of the two human chromosomes compared to the randomized sequences (Fig. 2). In addition, if we admit that the isochores do exist in the human chromosomes 21 and 22, then their (G + C) contents are entirely different (Fig. 3) from the (G + C) contents of the human isochores inferred from the centrifugation studies (1–3). For a more conclusive opinion, we must wait for a complete nucleotide sequence of the DNA of a bigger human chromosome. In any case, the present analysis demonstrates that the isochores should be defined in unambiguous molecular terms to be useful in up-to-date genome analysis.

ACKNOWLEDGMENT

This work was supported by Grant A5004802/1998 from the Grant Agency of the Academy of Sciences of the Czech Republic.

REFERENCES

1. Bernardi, G., Olofsson, B., Filipski, J., *et al.* (1985) *Science* **228**, 953–958.
2. Bernardi, G. (1995) *Annu. Rev. Genet.* **29**, 445–476.
3. Bernardi, G. (2000) *Gene* **241**, 3–17.
4. Dunham, I., Shimizu, N., Roe, B. A., *et al.* (1999) *Nature* **402**, 489–495.

5. Hattori, M., Fujiyama, A., Taylor, T. D., *et al.* (2000) *Nature* **405**, 311–319.
6. Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., *et al.* (1997) *Science* **277**, 1453–1474.
7. Stoesser, G., Tuli, M. A., Lopez, R., and Sterk, P. (1999) *Nucleic Acids Res.* **27**, 18–24.
8. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., *et al.* (2000) *Nucleic Acids Res.* **28**, 15–18.
9. Häring, D., and Kypr, J. (1999) *J. Biomol. Struct. Dyn.* **17**, 275–280.
10. Häring, D., and Kypr, J. (2000) *Biochem. Biophys. Res. Commun.* **272**, 571–575.
11. Haring, D., and Kypr, J. (1999) *J. Theor. Biol.* **201**, 141–156.
12. Hori, H., and Osawa, S. (1986) *Biosystems* **19**, 163–172.
13. Collins, D. W., and Jukes, T. H. (1993) *J. Mol. Evol.* **36**, 201–213.
14. Porter, T. D. (1995) *Biochim. Biophys. Acta* **1261**, 394–400.
15. Deschavanne, P., and Filipinski, J. (1995) *Nucleic Acids Res.* **23**, 1350–1353.
16. Ikemura, T., and Aota, S. (1988) *J. Mol. Biol.* **203**, 1–13.
17. Ikemura, T., Wada, K., and Aota, S. (1990) *Genomics* **8**, 207–216.
18. Fickett, J. W., Torney, D. C., and Wolf, D. R. (1992) *Genomics* **13**, 1056–1064.
19. Fukagawa, T., Sugaya, K., Matsumoto, K., *et al.* (1995) *Genomics* **25**, 184–191.
20. Bradnam, K. R., Seoighe, C., Sharp, P. M., and Wolfe, K. H. (1999) *Mol. Biol. Evol.* **16**, 666–675.