Molecular Biology, Vol. 35, No. 3, 2001, pp. 318–323. Translated from Molekulyarnaya Biologiya, Vol. 35, No. 3, 2001, pp. 376–382.

GENOMICS. PROTEOMICS. BIOINFORMATICS

UDC 577.1

Occurrence of MIR Elements in the Complete Nucleotide Sequence of Human Chromosome 22

E. V. Korotkov[1] and M. A. Korotkova[2]

[1] *Center of Bioengineering, Russian Academy of Sciences, Moscow, 117312 Russia; E-mail: katrin2@mail.ru*

[2] *Moscow Engineering and Physical Institute, Moscow, 115409 Russia*

Received July 31, 2000

Abstract—The location of mammalian interspersed repeats (MIRs) and their density have been determined in the complete nucleotide sequence of human chromosome 22. The approach developed by us has allowed detection of 9675 MIRs at a statistically significant level, which by 15% exceeds the MIR number revealed by all previous approaches. It has been demonstrated that a considerable amount of MIRs missed by the algorithms applied earlier occurs in known DNA sequences of the human genome. The study of the MIR density revealed substantial irregularity of their distribution along the chromosome. The data on the MIRs thus found and the computer program searching for diverged sequences are available by E-mail: katrin2@mail.ru or katrin22@mtu-net.ru.

Key words: MIR, mathematical algorithms, evolutionary divergence, chromosome

INTRODUCTION

Steady accumulation of data on complete genomes of many organisms makes especially topical the development of mathematical methods to study symbol sequences. Obviously, mathematical modeling can substantially speed up the evaluation of biologically meaningful information contained in complete genomic sequences of many species. A final goal of such studies appears to be elucidating the mechanisms of genetic regulation of cell activity and creating a corresponding computer model. Such a model could allow intelligent control of cell activity according to human will, which opens new possibilities for improving human existence. However, to reach the goal, it is necessary to solve the following tasks. First, this requires the study of the structural arrangement of nucleotide sequences in full genomes, i.e., revealing different kinds of repeated sequences, various symmetrical reflections and periodic structures, and many others. Second, one needs to learn how to reliably determine the location and biological function of various genes and exons in full genomes, as well as the regulatory elements, by using mainly computer methods. Third, we should be able to construct huge genetic networks *in silico* using both computer methods and all the biological information accumulated. Solving this task requires also the development of extensive databases on the structure and function of numerous proteins and on their role in metabolism.

This paper is concerned with solving the first task. In the complete nucleotide sequence of human chromosome 22, we have determined the location of mammalian interspersed repeats (MIRs) by the original method developed by us earlier [1–3] and have also estimated the MIR density. The method [1] allowed us to reveal at a statistically significant level twice as many MIRs as by the CENSOR program, and 15% more MIRs than by the RepeatMask program [4–7, 21]. We have also demonstrated that some repeat families detected earlier have similarity with MIRs. The data obtained indicate that a substantial amount of MIRs missed by the algorithms applied earlier occurs in known DNA sequences of the human genome. In addition, the more ancient sequences are analyzed, the larger number of MIR copies is missed by the standard methods of similarity search. This means that the share of repeated sequences and transposon-like elements in mammalian genomes may be substantially higher than 42% [8]. A text file containing data on the location of MIRs revealed in human chromosome 22, the MIR sequences themselves, and a computer program to search for any drastically diverged sequences are available by E-mail: katrin2@mail.ru or katrin22@mtu-net.ru

RESULTS AND DISCUSSION

To search for MIRs in chromosome 22, we used the mathematical approach and the position-specific weight matrix that were applied to find MIRs in vertebrate genomes [1]. We used the Monte Carlo method, as before, to determine the threshold value Z that, when exceeded, testifies to the occurrence of MIR in a nucleotide sequence. For this purpose, we have generated a random text of more than 600 Mb in length

0026-8933/01/3503-0318$25.00

(20 full sequences of chromosome 22) with the same triplet correlation as in chromosome 22, since MIRs also occur in DNA coding regions [19]. Then we applied the program package developed earlier [1] to reveal the MIR-like sequences in this random text. As a result, no similarity has been revealed when Z exceeded 7.0, while two similarities have been found when Z was between 6.5 and 7.0. Thus, the probability of detecting a random similarity calculated on the full length of chromosome 22 is about 0.1 when Z is from 6.5 to 7.0. Therefore, the threshold for detecting MIRs is as follows: a nucleotide sequence is MIR if $Z > 6.5$.

A search for MIRs in chromosome 22 has been done on the basis of the weight matrix introduced earlier [1]. We split the full nucleotide sequence of chromosome 22 into 600-kb fragments designated AA to CD as in [9, 10]. As a result, 9675 MIRs have been found with $Z > 6.5$, including 8201 sequences with $Z > 7.0$. At the same time, only 4860 MIRs have been revealed before by the CENSOR package [9, 10]. Thus, we have detected 4815 MIRs more than by the previous algorithms [4–7].

If the probability of random detection of a sequence with Z between 6.5 and 7.0 is 0.1, then the probability of detecting even 5 sequences with Z falling into this interval will be negligible. Thus, we can take that almost all 9675 sequences revealed by us belong to the MIR family.

All MIRs thus revealed have been organized as a text file containing the coordinates of a MIR sequence, the Z values reflecting the statistical significance of observed similarity, and the optimal alignment of sequences (see Tables 1–4).

We have compared the sets of MIRs revealed by us in this paper with those found earlier [9, 10]. It turned out that the approach developed by us missed 82 MIRs detected in [9, 10]. This could be explained by the occurrence of large insertions and deletions (more than 20 bp in length) in these sequences, since we introduced limits on the insertion/deletion size [1]. The possibility of large insertions/deletions can be taken into account in the present algorithm, but this would increase severalfold the computation time.

Conversely, the earlier algorithms [9, 10] have missed 4815 MIRs with considerably high Z (>6.5). Five such sequences are shown in Tables 1 and 2. This argues against the algorithms used to search for various repeated sequences (http://www.girinst.org). Missing such a large number of MIRs is connected with the great extent of their divergence, since a search of simple homology does not allow detection of considerably diverged sequences. Thus, the use of the position-specific matrix introduced by us appears highly expedient. In addition, the applied approaches to similarity search, such as Blast or PSI-Blast, in order to speed up computations, require the occurrence of certain regions, so-called hits, containing no insertions/deletions, where the similarity should be of a certain statistically significant level above a threshold. In this situation, very ancient repeat families [1, 17, 18] having accumulated a large number of insertions, deletions, and nucleotide substitutions may have no such hits, which can make it impossible to detect most members of such a family. The fact that all the characteristics of ancient repeats were taken into account in our algorithm [1] together with the use of the position-specific matrix allowed us to reveal a substantially greater amount of MIRs than it has been done before [9, 10].

Table 1. Coordinates and statistical significance of MIRs missed by the CENSOR program

No.	Chromosome 22 region	MIR coordinates	Coordinates of similarity in the MIR consensus sequence	Z
1	AA	52299–52423	56–183	8.5
2	AA	365414–365594	60–245	9.0
3	AA	378094–378313	19–239	9.6
4	AA	389032–389194	92–260	9.8
5	AA	442629–442750	137–260	8.3

We also assessed the superposition of the unknown MIRs revealed by us and of the other repeat families described in [9, 10] and deposited at http://www.girinst.org. The superposition was observed for LINE2 (including the fragment of LINE2 called LINE2B), MER2, and the Alu elements. As follows from Tables 3 and 4, substantial correlation exists between MIRs and the LINE2 fragments and Alu elements. The similarity of MIR and LINE2 has been also demonstrated earlier [8, 11, 12]. Our results point out that MIRs may be almost entirely included in some LINE2 sequences. The similarity between LINE2 and MIRs was observed in the most conservative part of MIR, nucleotide positions 80 to 180 of the consensus sequence. This testifies to their close evolutionary relations. At the same time, in most cases, some LINE2B fragments coincide with MIRs. This indicates that our approach cannot distinguish such LINE2B fragments as a separate family and they are all included in the MIR family.

The similarity between the Alu elements and MIRs is mainly observed in a MIR region from position 10 to 130. This MIR region has substantial homology with tRNA genes [13, 14]. The Alu elements at their beginning have the same homology with tRNA genes [15]. We observed no similarity between MIRs and the Alu elements in the MIR consensus region beginning from the 120th base. Therefore, we explain the similarity revealed between MIRs and the Alu ele-

Table 2. Optimal alignment of MIRs revealed in chromosome 22 (upper sequence) and the MIR consensus sequence. Numeration as in Table 1

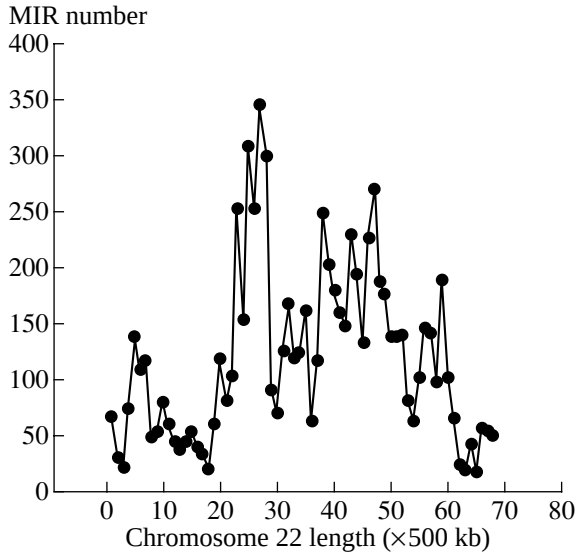
No.	Nucleotide sequences
1	aaat cct gact ct gccac- t a- at ct t t ct cagct t agg- aagt cat t t agc- t g- at gt ggct ccagt t t t ct cact ccat gt t at aat agggt t aaat caa- at gt cccaaact ca- a- g- t t ct t a- cagga
	gaat cccagct ct gccact t aat agct gt gt gacct t gggcaagt t act t aacct ct ct gagcct - cagt t t cct ca- t c- t gt aa- aat ggggat aa- t - aat a- gt acct acct cat agggt t gt t gt gagga
2	cct - gc- at t ccaat t - t t - gct gt at aac- t - gggcacaaat t acat aaccct ggat agagga- cagct t t cat gaact aat aact cat ct gt gat - at aggaat aggcacact - acagagat - t t ct gaggat t
	cccagct ct gccact t act agct gt gt gacct t gggca- ag- t t act t aa- cct ct ct - gagcct cag- t t t cct cat ct g- t a- aaat g- gggat aat aat agt acct ac- ct cat agggt t gt t gt gaggat t
	- - a- gag- t aat gaa- gt caaat gt t t gt ct t agt gcct accacact gt aggt ccct aat
	t aat gagt t aat acat gt aaagt gct t agaacagt gcct ggcacat agt aagt act caat
3	t aagt agcagt aaat ccgg- t cct gaat a- ct gact t t gaca- ct cagct t t ct ccacat cct t cct gt cact gcct t t gag- a- ct - act t cagaat ct t ccct t agct t ct at t t ct c- cat t t gt aaaat ggg
	t aag- agcacaggct ct ggagccagact gcct gggt t cga- at cccagct ct g- ccact t act agct gt - g- t gacct t gggcaagt t act t - a- a- - cct ct ct gagcct cagt t t c- ct cat ct gt aaaat ggg
	t - t gat gagggt a- - t - ct t cat cag- t agct gt gacaat aaaaat gggat cat - cat gcat cct cct t agccccat gagt aagct cccagt aagt g
	gat aat aat agt aact acct cat agggt t gt t gt gaggat t - aaat gagt t aat acat gt aaagt gct t agaacagt gcct g- gcacat agt aagt a
4	t aggcaagt t ac- - - - ct t t t ct - - gcct - - gt t t cact ct t c- a- aaaat t agagat t caat aat accgacct gat t t ct t t gggt t - t cat aagt aggaaat aaaat aat acat agaaaagact t ggaat at t gc
	t gggcaagt t acct t acct ct ct t agcct t agt t t c- ct cat ct gt aaaat - ggggat - - aat aat agt - acct - acct cat agggt t gt t gt gaggat t aaat gagt t aat acat gt aaagt gct t agaacagt gc
	ct gacacaaaat agt t gact t aa- aac- gt t aat t at t a
	ct ggcacat agt aagt - act caat aaat gt t agct at t a
5	aaaa- gagt - t aat at t t gt acct act t cct aat agggct - t cacaat gat t aaa- gagat aat acat g- aaagt gcacagcaaaaaggct ggt ccat agt aaagact t cat a- t t at t gat t t at t aa
	aaat ggggat aat aat agt acct acct c- a- t agggt t gt t gt gaggat t aaat gagt t aat acat gt aaagt gct t agaacagt gcct ggcacat agt aagt act caat aaat gt t - agct at t a

ments by a common origin of their certain regions from tRNA genes. This result testifies to the sensitivity of our algorithm.

We have also studied the distribution of the revealed MIRs along chromosome 22 (figure). We split the nucleotide sequence of chromosome 22 into 500-kb segments and estimated the number of MIRs in each segment. As follows from the figure, the distribution of MIRs is extremely uneven, and the difference in the MIR copy number per 500 kb may exceed ten times.

Such difference in MIR density may be explained in two ways. First, this might be connected with the different availability of the chromosome 22 sites for MIR insertions owing to the different extent of chromatin compactness. In this case, the less compact DNA regions might be more populated with MIRs than the more compact ones. Second, such irregularity could be explained by the evolutionary origin of chromosome 22. Since MIRs were suggested to originate 500 million years ago as a part of the CORE-SINE family [1, 17, 18], it is unlikely that the human chromosome 22 existed at that time in the present condition. Reasonably, it arose by various chromosome rearrangements and included DNA fragments with different MIR density.

We have also studied the occurrence of the Alu elements in chromosome 22. For this purpose, we created a set of 3200 Alu elements by the method of extended nucleic acid similarity [16]. The consensus sequence of the Alu elements [9, 10] served as the original sequence for creating the set. Earlier, 24,643 Alu elements have been revealed in chromosome 22 [9, 10]. Our program allowed detection of another 89 Alu elements and their fragments at a statistically significant level ($Z > 6.5$), while we missed 123 Alu elements probably because of large insertions and deletions. These results point out that the program package developed by us is efficient (an error about 0.5%) in revealing highly homologous repeat families, despite the possibility of missing the sequences with large insertions/deletions. The data on MIR detection demonstrate that our approach is especially useful for finding considerably diverged sequences, when homology between family members is less than 60%.



MIR density in human chromosome 22. Each dot indicates the number of MIRs revealed in the 500-kb DNA segment.

It is interesting to compare the results obtained in this study with those obtained in [9] and in the original work on chromosome 22 [20]. The RepeatMasker program [21] has revealed 8426 MIRs and 20,188 Alu elements [20], while our approach, as well as the CENSOR program [22], has detected 4000 Alu repeats more than RepeatMasker. This testifies that the limits on insertion/deletion size introduced by us are insignificant for revealing repeats. We have found 1249 MIRs more than RepeatMasker, and about twice as much MIRs as CENSOR. This means that using the

Table 3. Examples of similarities between the MIR consensus sequence and members of other repeat families in chromosome 22

No.	Chromosome 22 region	Repeat name	Repeat coordinates in chromosome 22	Coordinates of the MIR-like sequence in chromosome 22	Coordinates of similarity in the MIR consensus sequence	Z
1	BM	LINE2 (d)	475589–475824	475631–475814	73–256	7.6
2	AD	LINE2B (d)	541545–541646	541523–541694	71–251	9.2
3	AH	Alu-Jb (c)	286332–286613	286445–286571	9–121	7.3
4	AU	Alu-Jb (d)	394402–394606	394423–394532	145–33	7.5
5	BR	MER2 (c)	558954–559139	558849–559073	258–33	8.5
6	BO	Alu-J (c)	509561–509863	509617–509757	8–147	7.4
7	BU	Alu-Sx (d)	552981–553262	552931–553081	201–44	8.4
8	AB	Alu-Sp (d)	204635–204784	204635–204784	190–33	11.7

Table 4. Optimal alignment of MIRs revealed in chromosome 22 (upper sequence) and the MIR consensus sequence. Numeration as in Table 3

No.	Nucleotide sequences
1	cttgac-agctct-taccactctctg-aagtt-cttagc-tgtctgttgtct--gtttcc-cacttactggaacgtgaggttcatgagaatagagatcttgcttgtcttgttctctattag-attctaaa--agcta-gccta ctt-actagctgtgtgac-ct-tgggcaagttacttaacctctctg-agccttagtttcctca-t--ctgtaaaatg-ggg--ataataatag-tacctacctcatagggttgt-tgtgaggat-taaaagagttaatacatg gaata-gcctagaatagttcctggcacagagcaggtgtgcagtaaatatttgtt taaagtgcttagaacagtgcctggcacatagtaagtactcaataaatgttagct
2	cacatacctagctatttctacctcattgggtaagtcattta-ccactctgtgcctcagttt-a-c-tctgta-gtttaccattagact-gtgagct-ccttg-aggg-actt-tg-tca-taa-tca-ctgttacatccca-g cactta-ctagctgtgtg-a-c-c-ttgggcaagttacttaacctctctgagcctcagtttcctcatctgtaaaatggggataataatagt-acctacctcatagggttgttgtgaggattaaatgagttaatacatgtaaag tgcctcacactatgcctggcccttaagaagtgctcaataaatgt tgcttagaacagtgcctggcacatagtaagtactcaataaatgt
3	gtgcagtggc-acaatca-aggcttactgcagcctcgacctcctgggctcagaagatcccccaacctcagcctcctgagtagctggggcttgaacacctggggtcaagtgatctacccaccttg-gcctc gcatagtggttaagagcacaggct--ctggagcc-agactgcctgggttc-ga--atccc---agctctgccactt-actagctgtg--t-ga-c-cttggg--caagttacttaacctctctgagcctc
4	ccccatctctatcaaaataaaaaaattagccacacat-ggtg--gc-acatgcctgaggtc-c-cagctactcaggaggctgaggtaggagttgagcctgggaggtcaagcctgcag ggggtaaa-at-gtctactccttgactccgagtctctccaatccattgaacgggttccagtgtgtcgatca-ttcaccgtctcgaccctaagcttgggtccgtcag-accgaggtc
5	acagctaacatttcctaag-ggttaggacctgccaggcactgtgacaagcactctacatttgttctgtggcttagtcctca--g-aactctgtgaagttagtattattcaaggagtccc----t-cagcatgggggaattg-g tatcgattgtaaataactcatgaatgatacacggtccgtgacaagattcgtgaaatgtacataattgagtaaattaggagtgttgttgggatactccatccatgataa--taataggggtaaaatgtc-tactcctttgactc ttcacgacctggtggataacaaaatctg-cgatgatcaagtcccttacgtaaaggggcgtagtatct-gcattt-aacccacgt-gt-agcctccag cgagt-ct----ctccaattcattga-acgggttccagtgtgtcg-at-catt--cac-cgtc-tcgaccctaagcttgggtccgtcagaccgaggtc
6	agtgcagtgg-tacgatcatggtgcactgcagcctcaaccgcctgtgctcaaagcattctcctgcctcagcctccttagtagctg-g-gaccacaggcatgt-gct-a-ccacaactg-gc-t-aatttttttatttctagag agcatagtggttaagagcacag-gctctggagcc-agactgcctgggttcga---a-tc-ccagc-tctgcc-acttactagctgtgtgaccttgggcaagttacttaacctc-tctgagcctcagtttcctcatctgta-aa at-gagat □ atggggat
7	atggatgagct--ttcattccccacagtaa-cctttaaggt-ga-gct-caat-actcccattttgcagatgagcaaac--aggctcggaatggtacaggga-g-gccaaacgcggt-agctcatgcctgtaa-tcccagcac tacataattgagaaaattaggagtgttgttgggatactccatccatgataataataggggtaaaatgtctactcctttgattccgagtctctccaattcattgaacgggtt-c--cagt-gtgt-cgatcattcaccgtctcg ttggggaggccaatccaggtgg a--ccctaagcttgggtccgtc
8	ca-ttgattctcacaataattctatgtgctaggtactta-aagcatccccatttt-cagaatgtaggaaac--aggcataaagaagttaaat-acttggcc-agattactc--ct-gtaa-tcccagcacttggggaggccaa gtaaattaggagtgttgttgggatactccatccatg-ataataataggggtaaaatgtc-tac-tcctttgattccgagtctctccaattcattgaacgggttccagtgtgtggatcattcaccgtctcga--ccctaagctt ggcaggcag-atggcttgag gggtccgtcagaccgaggtc

position-specific matrix and avoiding hits is the most appropriate way to reveal both relatively homologous and drastically diverged repeat families.

To summarize, the results obtained in this study demonstrate that in the human genome repeated sequences are far from having been fully revealed. This mainly concerns the most ancient repeat families, whose members have quite considerably diverged from each other.

REFERENCES

1. Korotkov, E.V., Korotkova, M.A., and Rudenko, V.M., *Mol. Biol.*, 2000, vol. 34, pp. 553–559.
2. Korotkov, E.V., *Mol. Biol.*, 1991, vol. 25, pp. 250–263.
3. Korotkov, E.V., *Izv. Ross. Akad. Nauk, Ser. Biol.*, 1992, no. 4, pp. 660–672.
4. ftp://ncbi.nlm.nih.gov/repository/repbase/SOFTWARE/
5. Milosavljevic, A. and Jurka, J., *Computer Applications in Biosciences*, 1993, vol. 9, pp. 407–411.
6. Shpaer, E.G., *Microbial and Comparative Genomics*, 1998, vol. 2, pp. 75–86.
7. Shpaer, E.G., *Math. Modelling and Sci. Computing*, 1998, vol. 9, pp. 45–52.
8. Smit, A.F.A., *Curr. Opin. Genet. Devel.*, 1999, vol. 9, pp. 657–663.
9. *Repbase Update*, Jurka, J., Ed., Genetic Information Research Institute, 1997, (http://www.girinst.org/Repbase_Update.html).
10. *Database of Repetitive Elements (repbase)*, Jurka, J., Ed., NCBI Database Repository, 1995, (ftp://ncbi.nlm.nih.gov/repository/repbase/).
11. Jurka, J., *Curr. Opin. Struct. Biol.*, 1998, vol. 8, pp. 333–337.
12. Jurka, J. and Klonowski, P., *J. Mol. Evol.*, 1996, vol. 43, pp. 685–689.
13. Smit, A.F.A. and Riggs, A.D., *Nucleic Acids Res.*, 1995, vol. 23, pp. 98–102.
14. Jurka, J., Zietkiewicz, E., and Labuda, D., *Nucleic Acids Res.*, 1995, vol. 23, pp. 170–175.
15. Okada, N., *Curr. Opin. Genet. Dev.*, 1991, vol. 4, pp. 498–504.
16. Korotkov, E.V. and Korotkova, M.A., *DNA Research*, 1996, vol. 3, pp. 157–164.
17. Gilbert, N. and Labuda, D., *Proc. Natl. Acad. Sci. USA*, 1999, vol. 96, pp. 2869–2874.
18. Gilbert, N. and Labuda, D., *J. Mol. Biol.*, 2000, vol. 298, pp. 365–377.
19. Tulko, J.S., Korotkov, E.V., and Phoenix, D.A., *DNA Sequence*, 1997, vol. 8, pp. 31–38.
20. Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., *et al.*, *Nature*, 1999, vol. 402, pp. 489–496.
21. www.genome.washington.edu/UWGC/analysistools/repeatmask.htm
22. Jurka, J., Klonowski, P., Dayman, V., and Delton, P., *Computers and Chemistry*, 1996, vol. 20, pp. 119–122.