Molecular Biology, Vol. 34, No. 4, 2000, pp. 468–473. Translated from Molekulyarnaya Biologiya, Vol. 34, No. 4, 2000, pp. 553–559.

UDC 577.1

MIR: Family of Repeats Common to Vertebrate Genomes

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

[1] *Bioengineering Center, Russian Academy of Sciences, Moscow, 117312 Russia;*
E-mail: katrin22@mtu-net.ru
[2] *Moscow Institute of Engineering and Physics, Moscow, 115409 Russia*

Received December 10, 1999

Abstract—We studied the occurrence of mammalian interspersed repeats (MIRs) in DNA and RNA of vertebrates, invertebrates, and bacteria using the data from GenBank. A special algorithm based on a weight position matrix with optimal alignment using dynamic programming was developed to search for the traces of MIR dissemination. This allowed us to search for highly divergent MIRs carrying deletions and insertions. MIRs were detected in genomes of various fishes, including *Latimeria*. This suggests that the origin of MIRs dates back more than 400 million years. The method to search for similarity between highly divergent sequences may be used to find the genome fragments from various ancient repeat families and from various gene families.

Key words: mammalian interspersed repeats, dynamic programming, genome evolution

INTRODUCTION

Mammalian interspersed repeats (MIRs) are a very ancient repeat family of the SINE type. This family at present remains rather poorly known, though in the human genome it is represented by several thousand copies [1–7] found in coding regions of many human genes [8]. These repeats were detected in genomes of many mammals [5–7] and in genomes of birds [8], therefore, they originated more than 180 million years ago. Considering that the MIRs are highly divergent and may show no obvious homology, a large part of them would not be revealed by the methods of experimental biology based on DNA hybridization. Standard algorithms basing on the search for homologies between symbolic sequences of the FASTA type [9] are also unsuitable for MIR search. Special algorithms should be developed to detect the divergent MIR sequences in the data banks. The aim of this work was to study MIR dissemination in vertebrate genomes and to develop new programming algorithms to search the data banks for MIRs and other highly divergent sequences. Earlier we detected the MIRs using the method of enlarged similarity of the DNA nucleotide sequences with introduction of certain weight functions based on the available MIR set [4, 8, 10]. However, these approaches failed to detect the highly divergent MIR copies carrying deletions or insertions. Introduction of a weight position matrix for the MIRs using the method of enlarged similarity with further optimal alignment by dynamic programming allowed us to detect the MIRs in some fish genomes. We showed that some of the known fish repeat families had originated from the MIR.

The study of MIRs is important for better understanding the mammalian genome evolution. It should allow one to reproduce the evolution history for various genomes. The developed algorithms may be used to detect and analyze any other highly divergent gene families or any other sequences.

METHODS

Earlier we used the method of enlarged DNA nucleotide sequence similarity [10, 11] to find 49 MIR fragments (MB1 sequences) of about 150 bases [4]. At this step the method of enlarged similarity is essential, because it allows one to detect highly divergent MIR copies and generate a representative training MIR set. The presence of highly divergent MIR copies in the human genome is related to the long time of MIR existence [8]. Since it was shown that MIRs are 260 bases long [6], we completed each of the found MIRs to 260 bases adding 55 bases from each 3' and 5' end. This MIR set was used as a training set to calculate the weight of each nucleotide for all MIR positions, for further application of the weight function method [8] to create a more representative MIR set. For the training set of 49 MIR sequences the weight $v(i,j)$ was calculated as

$$v(i, j) = f(i, j)\ln\{f(i, j)/p(i)\}, \qquad (1)$$

where i is a base (a, t, c, g) and $j = 1, 2, \ldots, 260$, $f(i,j)$ shows the frequency of the base i in position j, $p(i)$ is a frequency of the base i in all set of the 49 sequences. Searching for the MIRs in the human genome clones

0026-8933/00/3404-0468$25.00

we calculated the total weight of a subsequence with coordinates $k+1, \ldots, k+260$ as:

$$V(k) - \sum_j v(S(j+k), j+k), \qquad (2)$$

where j varies from 1 to 260; k from 1 to $L-259$; L is the length of the sequence from GenBank; $S(j+k)$ shows the appearance of the base in position $j+k$ of the studied sequence. In order to estimate statistical significance $V(k)$ we used $Z = (V_m - V)/\sigma(V)$, where V_m and $\sigma(V)$ are the mean and the standard deviation of $V(k)$ for the set of random nucleotide sequences with base composition corresponding to that in human DNA. Analyzing DNA sequences of the primates from GenBank, we selected the cases when Z was higher than 6.0. The number of such MIR sequences was over $2 \cdot 10^4$.

Since MIR represent a very ancient repeat family aging more than 180 million years, it is reasonable to expect, beside nucleotide substitutions, the presence of various deletions and insertions. To search for MIR carrying deletions and insertions we developed an algorithm combining the weight position matrix [12] and dynamic programming methods [13]. The obtained sample of $2 \cdot 10^4$ MIRs was used to generate a new weight function $wes(i, j)$ as the weight of nucleotide i (i = a, t, c, g) in position j ($1 \le j \le 260$). This weight is introduced according to (1), but for the MIR set of more than $20 \cdot 10^4$. Using (1) to introduce weight functions appears reasonable, since more frequent nucleotides have higher weights. For example, if $p(i) = 0.2, f(i, j)) = 0.4$. In this case $wes(i, j) = 0.4 \log 2$. In case where $p(i) = 0.1, f(i, j)) = 0.2$ $wes(i, j) = 0.2 \log 2$, i.e. two times less. On the other hand, if one introduces the weights as suggested earlier [33], these weights would be equal. It is more logical to consider higher weight for more frequent nucleotides.

At the same time, using (1) has a drawback: rare nucleotides would have almost zero weight. In order to avoid this, we found the mean weight $w_{mid}(j)$ for each position j corresponding to one nucleotide, and then used the modified weight as $w(i, j) = wes(i, j) - 1.1 w_{mid}(j)$. In this case even very rare nucleotides would have negative weight according to (1), though the function $w(i, j)$ at small frequencies $f(i, j)$ is somewhat nonmonotonous.

The resulting weights for each base in the set of $2 \cdot 10^4$ sequences are shown in the figure. Then we applied the dynamic programming algorithm [13] to fill the matrix $F(k, j)$, where k shows the number of the nucleotide in subsequence S of the analyzed 280-symbol DNA, j shows the position of the relative matrix W, elements of which are the weights $w(i, j)$ calculated as above. We consider $f(k, j)$ as an element of matrix F:

$$f(k, j) = \max\{f(k-1, j) - v_d; \\ f(k, j-1) - v_d; f(k-1, j-1) + w(s(k), j); 0.0\} \qquad (3)$$

Zero column and zero row may be defined as $f(k, 0) = f(0, 0) - kv_d; f(0, j) = f(0, 0) - jv_d$. Element $f(0, 0) = 0.0$. The weight of the deletion v_d was taken to be –0.3, equal to the mean weight $w(i, j)$. In the process of calculations we fill completely the matrix $F(k, j)$, and then find its maximal element $f_{max}(k_m, j_m)$. Depending on the f_{max} position we determine, as earlier [13], the optimal alignment as the way from the maximal element to the first zero with coordinates (k_0, j_0). This allows us to find the "maximal subsequence" [13]. As a result, an alignment will show this maximal subsequence S_m, and the location of the related W matrix will be marked by the nucleotides of higher weight. In order to speed up the calculation, we considered the maximal nucleotide subsequence with $|(k-j)| < 10$. This limits the possible number of insertions and deletions, but allows one to run the calculation more than 10 times faster as compared with the F matrix.

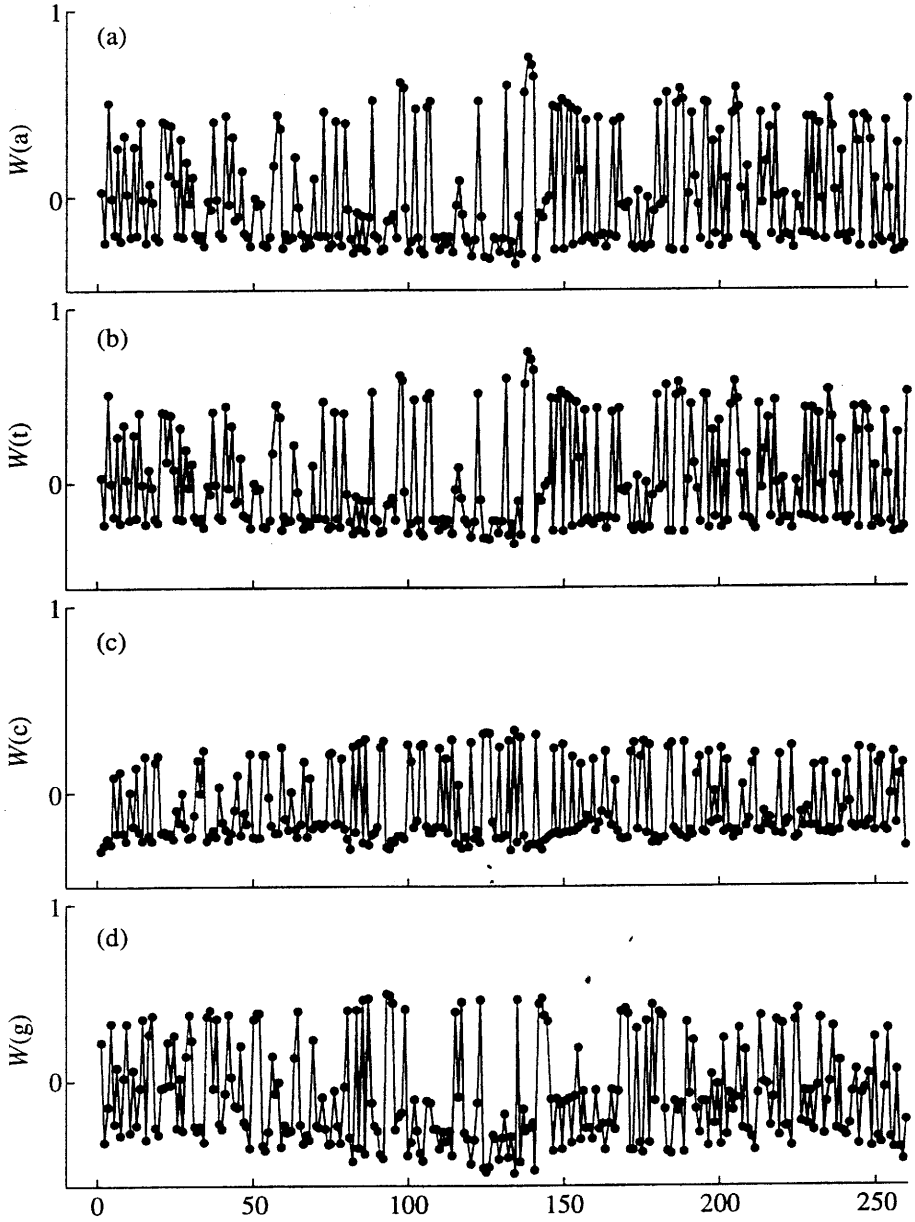
The sequence S may be considered as a window. Analyzing a DNA sequence from data bank we move this window with a step of ten bases to obtain a profile of the analyzed sequence as function $F_{max}(k)$. The length of the sequence S (window) was considered to be 280 bases to allow search for maximal subsequence S_m considering the possibility of insertions and deletions in the query sequence S.

To estimate the statistical significance of the optimal alignment, the alignment matrix F' is filled using the formula:

$$f'(k, j) = \max\{f'(k-1, j) - v_d; \\ f'(k, j-1) - v_d; f'(k-1, j-1) + w(s(k), j)\}. \qquad (4)$$

Then we find $d = f'(k_m, j_m) - f'(k_0, j_0)$. If we use (4), the distribution of d is close to normal for the given coordinates of the start and the end of the maximal subsequence [14]. We confirmed this using the Monte Carlo method on $5 \cdot 10^7$ randomly generated sequences. The result is true for any coordinates of the start and the end of the maximal subsequence if $|(k-j)| < 10$ and $\{(k_m - k_0)^2 + (j_m - j_0)^2\} > 3600$. This allows one to estimate the statistical significance of the maximal subsequence using $Z = (d - d_m)/\sigma(d)$. The values of d_m and $\sigma(d)$ may be found applying the Monte Carlo method to randomly generated DNA sequences with the same frequencies and triplet correlation of the bases as in a human DNA sequence from GenBank.

In order to reveal nonrandom maximal subsequences, it is essential to find Z_0, providing <5% probability to randomly find a sequence with $Z > Z_0$ in the



Weight functions of nucleotides a, t, c, and g for the MIR, bases 1–260.

whole text (a sequence from vertebrate genome). To do this, one should analyze the distribution of $d = f'(k_m, j_m) - f'(k_0, j_0)$ for a sample of more than 10^{12}, a very laborious procedure. Alternatively, we found Z_0 applying the Monte Carlo method to a random text of 600 mln symbols, about 20 times more than the total size of the vertebrate DNA sequences presented in GenBank. The random text was generated using base frequencies and triplet correlation as in the vertebrate sequences GenBank. For this sample we found no cases of similarity at $Z > 7.0$. We also analyzed bacterial sequences of more than 150 mln symbols total length, without MIR. No similarity was found at $Z_0 >$ 7.0. Therefore, we assume that this Z_0 value is critical and consider that we find MIRs in a vertebrate sequence from GenBank, if Z obtained for alignment is >7.0. The software to search for the MIR written in Fortran-77 as available at E-mail request to: katrin2@mail.ru.

RESULTS AND DISCUSSION

In this work we showed the presence of MIRs in various vertebrate genomes. Most of the results was obtained at a significance level $Z > 8.0$. Importantly, we confirmed the presence of MIRs in the chicken

Table 1. Coordinates and significance level *Z* for vertebrate and invertebrate sequences from GenBank

N	GenBank accession number	Position in MIR	Position in GenBank	*Z*	Species and sequence type	Reference
1	AB001865	135–6	621–756	7.4	*Coregonus lavaretus*, noncoding region	15
2	AB009705	51–120	213–280	7.4	*Tropheus moorii*, AFC repeat, 144–472	16
3	AB016549	41–120	54–129	8.0	*Julidochromis transcriptus*, AFC repeat, 1–333	17
4	AB016567	37–150	51–173	8.5	*Julidochromis transcriptus* AFC repeat, 1–376	17
5	AF061275	17–124	5305–5416	8.7	*Ictalurus punctatus*, noncoding region	18
6	AOIRHODOPS	123–4	9095–9210	10.6	*Anolis carolinensis*, intron	19
7	AOIRHODOPS	223–41	6304–6488	7.3	*Anolis carolinensis*, intron	19
8	CCJAK1S07	3–137	2945–3081	7.8	*Cyprinus carpio*, intron	20
9	CHKFASA	57–181	4230–4342	10.0	*Gallus gallus*, noncoding region	21
10	FHU59833	13–211	211–401	8.4	*Fundulus heteroclitus*, noncoding region	22
11	GGU46503	152–66	1663–1744	8.3	*Gallus gallus*, noncoding region	23
12	LCU08034	71–144	8330–8410	7.2	*Latimeria chalumnae*, intron	24
13	OGU34716	1–147	707–853	8.2	*Oncorhynchus gorbuscha*, intron	25
14	ONHHPAF	7–148	79–220	7.3	*Oncorhynchus keta*, repeat Hpa I, 79–264	26
15	ONHSINES5	7–147	169–314	7.1	*Oncorhynchus kisutch*, repeat HpaI 163–372	27
16	ONHSINESD	1–148	56–203	8.4	*Oncorhynchus gorbuscha*, repeat HpaI, 55–240	28
17	AF036751	7–64	32–88	7.5	*Schistosoma mansoni*, RNA polymerase III-binding sites, boxes A and B, 31–49, 74–85	29
18	BFY18367	142–4	13895–14026	9.7	*Branchiostoma floridae*, noncoding region	30
19	PV14GENE	43–178	2296–2423	12.9	*Plasmodium vivax*, noncoding region	31
20	PVPVA1	170–20	336–477	12.1	*Plasmodium vivax*, noncoding region	32

genome [8] (Table 1, nos. 9, 11). We also detected MIRs in the genome of *Latimeria* (Table 1, no. 12). This suggests that the MIRs may originate as early as in Devon, more than 400 million years ago. The long existence of MIRs is confirmed by their presence in the genomes of various fish species (Table 1, nos. 1–5, 8, 10, 13–15) and lizard *Anolis carolinensis* (Table 1, nos. 6, 7). Interestingly, repeat families AFC and *Hpa*I from fish genomes are similar to MIR. This suggests the origin of these families and the MIR family from the SINE repeat family, which existed in the common ancestor of the studied species.

We failed to detect MIRs in other vertebrate species. This may be explained by an insufficient number of the available sequences, on the one hand, and by the high nucleotide divergence in the present traces of the MIR dissemination, on the other. One may also suggest that the MIR originated only in those species of the ancient vertebrates which gave rise to modern mammals and primates. These possibilities may be tested when a sufficient number of sequences is available and when better algorithms allowing detection of more divergent copies of the DNA sequences originating from a common ancestor are developed.

We have also analyzed the occurrence of the MIRs in invertebrate genomes. Similarity with MIRs was revealed for the satellite DNA containing binding sites for RNA polymerase III (Table 1, no. 17) and for the genomic sequence of *Branchiostoma floridae* (Table 1, no. 18). Similarity of the MIRs and the region of RNA polymerase III binding is well explained by homology of the first 70 nucleotides of the MIR with the genes encoding tRNA and transcribed with RNA polymerase III. The presence of the MIR-like sequence in the genome of *Branchiostoma floridae* may indicate that the MIR family is even more ancient. However, this should be proved by finding at least some more MIR copies in invertebrate genomes. This would be possible when a greater number of sequences is available from the invertebrates. We also detected MIR in the genome of *Plasmodium vivax*. Earlier, Alu repeats were found in *P. vivax* [32]. Probably in this case there was a transfer of genomic material from human to *P. vivax*, because *P. vivax* exists for a long time within the human organism. The transferred DNA fragment probably contained both the Alu repeat and MIR.

In all cases, similarities were found within the most conserved MIR fragment (bases 85–155) and

within the tRNA-like MIR fragment (bases 1–85) [4, 6]. One may suggest that the cases of similarity are related to the RNA polymerase III-binding sites. In this case, similarity would be detected mainly in the bases 1–85 of the MIR. To test this, we calculated the weight function $w(i, j)$ only for nucleotides 1–85 from the MIR sample of $2 \cdot 10^4$ sequences. Then we applied our algorithm to all sequences where traces of MIR dissemination had been detected (Table 1). Similarity was observed at $Z < 6.0$. This indicates that vertebrate MIRs have retained both the tRNA-like part and the most conserved MIR part starting from base 85 (Table 2).

In conclusion, we have shown that the MIRs form a repeat family aging about 400 million years. To our present knowledge, MIR is the only repeat family so widespread in the vertebrate genomes. The algorithms described in this work may be used to analyze any other family of the repeats as well as any family of genes to find similarity between distinct sequences that originated hundreds of millions years ago.

REFERENCES

1. Degen, S.J. and Davie, E.W., *Biochemistry*, 1987, vol. 26, pp. 6165–6167.
2. Donehower, L.A., Slagle, B.L., Wilde, M., Daglington, G., and Buted, J.S., *Nucleic Acids Res.*, 1989, vol. 17, pp. 699–722.
3. Korotkov, E.V., *Dokl. Akad. Nauk SSSR*, 1990, vol. 311, pp. 238–242.
4. Korotkov, E.V., *Mol. Biol.*, 1991, vol. 25, pp. 250–263.
5. Korotkov, E.V., *Izv. Akad. Nauk SSSR, Ser. Biol.*, 1992, no. 4, pp. 660–672.
6. Smit, A.F.A. and Riggs, A.D., *Nucleic Acids Res.*, 1995, vol. 23, pp. 98–102.
7. Jurka, J., Zietkiewicz, E., and Labuda, D., *Nucleic Acids Res.*, 1995, vol. 23, pp. 170–175.
8. Tulko, J.S., Korotkov, E.V., and Phoenix, D.A., *DNA Sequence*, 1997, vol. 8, pp. 31–38.
9. Pearson, W.R. and Lipman, D.J., *Proc. Natl. Acad. Sci. USA*, 1988, vol. 85, pp. 2444–2448.
10. Korotkov, E.V. and Korotkova, M.A., *DNA Research*, 1996, vol. 3, pp. 157–164.
11. Korotkov, E.V., *DNA Sequence*, 1994, vol. 4, pp. 411–413.
12. Gribskov, M., McLachlan, A.D., and Eisenberg, D., *Proc. Natl. Acad. Sci. USA*, 1987, vol. 84, pp. 4355–4358.
13. Waterman, M.S., *Introduction to Computational Biology. Map Sequences and Genomes*, London: Chapman and Hall Press, 1995.
14. Seledtsov, I.A., and Kolpakov, F.A. in *Proc. First Int. Conf. on Bioinformatics*, Novosibirsk: Inst. of Cytology and Genetics SO RAN Press, 1998, pp. 301–304.
15. Hamada, M., Kido, Y., Himberg, M., Reist. J., Cao, Y., Hasegawa, M., and Okada, N., *Genetics*, 1997, vol. 146, pp. 355–367.
16. Takahashi, K., Terai, Y., Nishida, M., and Okada, N., *Mol. Biol. Evol.*, 1998, vol. 15, pp. 391–368.
17. Terai, Y., Takahashi, K., and Okada, N., *Mol. Biol. Evol.*, 1998, vol. 15, pp. 1460–1471.
18. Xia, Z.F., Patino, R., Gale, W.L., Maule, A.G., and Densmore, L.D., *Gen. Comp. Endocrinol.*, 1998, vol. 113, pp. 360–368.
19. Kawamura, S., and Yokoyama, S., *Gene*, 1994, vol. 149, pp. 267–270.
20. Chang, M.S., Chang, G.D., Leu, J.H., Huang, F.L., Chou, C.K., Huang, C.J., and Lo, T.B., *DNA Cell Biol.*, 1996, vol. 15, pp. 827–844.
21. Kasturi, R., Chirala, S., Pazirandeh, M., and Wakil, S.J., *Biochemistry*, 1988, vol. 27, pp. 7778–7785.
22. Schulte, P.M., Gomez–Chiarri, M., and Powers, D.A., *Genetics*, 1997, vol. 145, pp. 759–769.
23. Lin, A.W., Chang, C.C., and McCormick, C.C J., *Biol. Chem.*, 1996, vol. 271, pp. 11911–11919.
24. Betz, U.A.K., Mayer, W.E., and Klein, J., *Proc. Natl. Acad. Sci, U.S.A.*, 1994, vol. 91, pp. 11065–11069.
25. Miller, K.M., and Withler, R.E., *Immunogenetics*, 1996, vol. 43, pp. 337–351.
26. Murata, S., Takasaki, N., Saitoh, M., Tachida, H., and Okada, N., *Genetics*, 1996, vol. 142, pp. 915–926.
27. Murata, S., Takasaki, N., Saitoh, M., and Okada, N., *Proc. Natl. Acad. Sci. USA*, 1993, vol. 90, pp. 6995–6999.
28. Takasaki, N., Park, L., Kaeriyama, M., Gharrett, A.J., and Okada, N., *J. Mol. Evol.*, 1996, vol. 42, pp. 103–116.
29. Ferbeyre, G., Smith, J.M., and Cedergren, R., *Mol. Cell. Biol.*, 1998, vol. 18, pp. 3880–3888.
30. Boeddrich, A., Burgtorf, C., Francis, F., and Lehrach, H., ftp://www.nbci.nih.gov/genbank/gbvrt.seq.Z
31. Gupta, S. and Sharma, Y.D., ftp://www.nbci.nlmnih.gov/genbank/gbvrt.seq.Z
32. Dhar, A., Gupta, S., and Sharma, Y.D., *FEBS Lett.*, 1998, vol. 423, pp. 193–197.
33. Karlin, S. and Altschul, S.F., *Proc. Natl. Acad. Sci. USA*, 1990, vol. 87, pp. 2264–2268.

Table 2. MIR nucleotide sequences found in various vertebrate and invertebrate genomes from GenBank (upper lines). Lower lines show the integrated MIR sequence with the bases of maximal weight at each position. The MIRs are numbered according to Table 1

N	Sequence
1	cagsgtgact–tacag–gagca–attag–ggttaagtgtcttgctcaagggcacatcaacatattcctcatctagtcagctcagggattcaaacc–agtaacctttcagttactg–gcccaatgctc gtct–actcctttgactc–cgagtc–tctccaattcattgaacgggttccagtgt–gtcg–at——ca–ttcacc–gtctcgaccctaagcttgggtccgtc–aga–ccgaggtctcgga–cacgag ttaacagctaggctat aattggtgatacgata
2	ggttcga–tccctgggct–ct–ctgtcctggtcgt–tgtgtcctgggcaagacactttacc–ctact–tgcct ggttcgaatcccagctctgccact–tactag–c–tgtgtgacctgggcaagttacttaacctct–ctgagcct
3	aggttgcc—ggttcgagccccggcttggacagtctcggtc–gttgtgtcctgggcaagacacttcacc–ctact—gcct agactgcccgggttcgaatcccagctctgccactttactagctgt–gtgacctgggcaagttacttaacctct–cttagcct
4	aaccagagggt–tggcgttcgaatccca–ctccatgccagccattttgccgttgtgtcctgggcaaggcactttacccacct–tgccttcagtgccccccacactggtgtatgaatgtgaatgaat agccagactgcctgg–gttcgaatcccagct–c–tgccacttacta–gctgt–gtgacctgggcaagttacttaacctctctgagcct–cagtttcct—ca–t–ctgtaa–aatggggat–aat
5	gttaag–gcgttgggttact–ga–tcagacggtccaggggttcaaaccccaacaccgccaagctgcca–ctgtcgggacttcgagcaaggcctttaatcctctccgccccagcagt gttaagagcac–aggct–ctggagccagactgcc—tgggttcgaatcccagctctgccac–ttactagctgt–gtgacctgggcaagttacttaa–cctctctgagcc–tcagt
6	ctgaggctggaaga–g–t—gtgacttgaccaaggtcacccagtgagtttctatggctgagttgggatttgaaccctggtctccagagtcctaatccaacactc–aaatcattacaccatgc gactccgagtctctccaatccattgaacgggttccagtgtgtcgatc—attcaccgtctcgaccctaagcttgggtccgtcagaccgaggtctcggacacgagaattggtgatacgatatg
7	ggcacttttccctaagtgcttaaagtgc–tt–cct–atatacatttcctttaataattct–tgtaacaaacccctg–tagtttaatcctctttattacagatgcgtatgagtgtgtctggggaggggga ccgtgaca–agattcgtgaaatgtacataattgagtaaat–taggagtgttgttgggatactccatccatgataat—aat–aggggtaa–aatgtctactcctt–tgactccga–gtctctccaat atggaga–ga–cagagacccccctaagcgaattcattgctgagcagagacttgaacc–aggctctct tcattgaacgggt–tccagtgtg–t–cgatcatt caccgtctcgaccctaagcttgggtccgtcaga
8	agtcatggcctagtggttagagagttt–gactcctaacccctaaggttg—tgggtttgagtctcgggccggcaaatccacga–ctgaggtgcccttgagcaaggcaccaaaccccactgtag–cttta agt–atagcatagtggttta–agagcacaggctctggagcc–a–gactggctgggttcgaatcccagctctgccact–tactagctg–tgtgacctgggcaagttacttaacctctctg–agcctca –tggacgtcatattta gttttcc–tcatctgta
9	aattctgattctggctc–cact–g–t–tgtggttttgtacaagt–ggtt—cgtgt–taccgcctcagttttccccatctgtaaaatggg–ac—taatagcatctacctca–agg–tcgttgagag aatcccagctctgccacttactagctgtgtgacctgggcaagttacttt acctctctg–agcctcagttttcctcatctgtaaaatggggattataatagtacctacctcatagggttgttgtgag
10	agtgggt–agagcattgcacttatgtcctgcaggtt—ctgggttcaactcccacacactgccact—ct–g—g–gtc–cc–tgagcaaaaccctt aacctcagat–tgcc–cact—gctcacc agtggttaagagcacag–gctc–tggagc–cagacttcctgggttcgaatccca–gctctgccacttactaggtgtgtgacctt gggcaagttacttaacctc–tctgagcctcagttttcctcatc aag–gagatgtgtta–aatgcagagaacaaattttcattggattgtatgttacaataaagatga–ttattattgtgtaaa–tgct t–gtaaaatg–gggataata–atagtacctacctcatagggttgt–tgtgaggattaa–atgagttaatac–atgtaaagtgct
11	ttaatattcac–ttcta–a–at–agcaaactgagacagagagaaaaaaaatacctt gcccaagcttgcccag–tagaaaagtggcaga aataataggggtaaaatgtctactcctttgactccgagtctctccaattcattgaacgggttccagtgtgtcgatca–ttcaccgtct
12	cactcactccctatgtaacctgggtgcaggacgttt ggcctccctgtgcctctcctt cctt cctccagcggtacaactggg cactt actagctgtgtgacctt gg–gcaagtt actt aacctctctgagcctc–agt——ttcctc–atctgta–aaatggg