Bull. Fac. Agr., Saga Univ. No. 91 : 37～43 (2006)

In Silico Analysis of Repeat Sequences from the Porcine and Bovine Genome

Yasuhiko WADA and Takuto SHIGEMORI
(Laboratory of Animal Production)
Received September 30, 2005

Summary

The human and mouse genome research projects reported that the proportion of the length of interspersed repeats in the mouse genome is greater than 35% and that of humans is greater than 45%. The evolution and function of the interspersed repeats has not yet been clarified. In this study, the interspersed repeats were searched from the sampled genome sequences. The proportion of each class of repeats in the sequences was calculated and the age distribution was analyzed by the nucleotide substitution ratio obtained from each consensus sequence.

The LINE (long interspersed repeat element) was the largest part in the mammalian genome; however, in the porcine genome it was slightly lesser than those in the other mammalian genomes. The ratio of SINE (short interspersed repeat element) repeats in the human and porcine genomes was approximately 14%. The results suggest that in the genome of some mammalian species, the SINE repeats are copied and interspersed at a higher frequency.

The peaks of the nucleotide substitution ratio obtained from the consensus sequence of SINE, LINE, LTR element, and DNA element in the porcine genome were 10%－20%, 30%－40%, 20%－30%, and 20%－30%, respectively. The peak of the age distribution of the classes of repeats were almost the same in the mammalian genome. The BovB/Art2 family was only located in the cattle genome, and the L3/CR1, LINE1, and LINE2 families were located in the porcine and cattle genomes.

Recombination of genomic regions can be investigated using the SINEs, LINEs, and other repeats, which have different age distributions. Therefore, our results on the age distribution of SINEs, LINEs, and other repeats are useful in studies on genome evolution.

Key words: interspersed repeats, SINE, LINE, repetitive sequences, genome structure

Introduction

When the initial whole genome sequences of human and mouse genomes were published, the number of repetitive sequence was regarded to be one of their most interesting features[1,2]. In particular, the proportion of the length of interspersed repeats in the mouse genome is greater than 35% and that of humans is greater than 45%. However, the proportion in the fly genome is 3.1% and that in worm is 6.5%[1]. The presence of a large fraction of interspersed repeats is the characteristic feature of the mammalian genome.

Interspersed repeats represent the fossils of transposable elements and are a principal force

involved in reshaping the genome. For example, Esnault *et al.* (2000)[3] showed that human LINE (long interspersed repeat element) can mobilize the transcribed DNA not associated with a LINE sequence by a process involving the diversion of the LINE enzymatic machinery by the corresponding mRNA transcripts. Since most genes are produced by gene duplication, the interspersed repeats appear to play an important role in evolution.

Many SINEs (short interspersed repeat elements), such as PRE-1[4,5] (swine), CHR-1[6] (whale), and vic-1[7] (camel) were reported in other mammals; some LINEs are found to be widely distributed in mammals. It is important to examine the proportion and age distribution of the interspersed repeats in other mammalian genomes in order to study the evolution and function of the repetitive sequences. In this study, the genomic sequences with lengths greater than 5 kbp were sampled from the porcine and bovine genomes peposited in the GenBank DNA database. The interspersed repeats were searched from the sampled genome sequences by the RepeatMasker 2 program (http://www.repeatmasker.org/). The proportion of each class of repeats present in the sampled sequences was calculated and the age distribution was analyzed by the nucleotide substitution ratio obtained from each consensus sequence.

Materials and methods

Porcine and bovine genome sequences with a length greater than 5 kbp were selected from GenBank DNA database. The number of the sampled sequences, the average and standard deviation of the size of the sequences, and the average and standard deviation of the GC contents (%) are shown in Table 1. The number of sequences in this study was 70 for swine and 99 for cattle. The average size of the selected genome sequences was approximately 46 kbp for swine and 33 kbp for cattle. The average GC contents of the porcine and cattle sequences were 47.2% and 46.7%, respectively.

The repeat elements were searched from the genome sequences by the RepeatMasker2 program. The full-length and partial-length members of all known repeat families that were selected by RepeatMasker2 were collected for each class of repeats. The percentage of the total bases in the genome sequences for each class of repeats was computed in order to examine the importance of each class of repeats in the porcine and cattle genomes.

Each repeat element was compared with the consensus sequence of the subclass of repeats in RepBase 6.7[8] in order to analyze the age distribution of the different classes of repeats. The approximate age of each repeat can be inferred from the nucleotide substitution ratio obtained from the consensus sequence by exploiting the fact that each copy is derived from the consensus sequence and has accumulated mutations randomly and independently of other copies. The nucleotide substitution ratio obtained from the consensus sequence was estimated by the method used

Table 1. Sampling of the genomic sequence data used in this study

Animal	Number of the sequences	Average and S.D. of the size of the sequences	Average and S.D. of the GC contents (%)
Swine	70	45788±64279	47.2±7.4
Cattle	99	32836±56970	46.7±9.0

by Tajima and Nei (1984)[9].

Results

The percentage of total bases in the sampled genome sequences for each class of repeats is shown in Table 2. LINE was the largest part in the mammalian genome; however, the proportion of LINE in the porcine genome was slightly lesser than those in the other mammalian genomes. The ratio of SINE repeats in human and porcine genomes was approximately 14%, and it was equivalent to the ratio of LINE in the porcine genome. The proportions of the LTR elements in human and mouse genomes is 8.55% and 9.87%, respectively and those in the other mammalian

Table 2. Percentage of the total bases of the genome sequences in each class of the repeat sequences.

	SINE	LINE	LTR element	DNA element	Unclassified interspersed repeat	Small RNA	Satellite	Simple repeat
Human[1)]	13.64	20.99	8.55	3.03	0.15	0.04	0.34	0.87
Mouse[1)]	8.22	19.20	9.87	0.88	0.37	0.06	0.30	2.27
Swine	14.32	13.77	2.36	1.69	0.02	0.10	0.00	0.79
Cattle	9.32	20.99	2.04	1.39	0.01	0.08	0.00	0.41

1) These data are transcribed from the Mouse Genome Sequencing Consortium (2002).

Table 3. Distribution of the nucleotide substitution ratio from the consensus sequence in the sampled sequences of the porcine genome.

Substitution ratio from consensus sequence (%)	SINE	LINE	LTR element	DNA element	Unclassified interspersed repeat
0−10	277*	3	1	1	0
10−20	998	93	18	39	0
20−30	524	368	123	136	1
30−40	472	402	100	87	1
40−50	67	360	5	33	0
50−60	14	111	18	12	0
60−70	1	1	1	1	0

* Number of repeats

Table 4. Distribution of the nucleotide substitution ratio from the consensus sequence in the sampled sequences of the cattle genome.

Substitution ratio from consensus sequence (%)	SINE	LINE	LTR element	DNA element	Unclassified interspersed repeat
0−10	322*	0	3	2	0
10−20	687	62	18	34	0
20−30	276	260	86	108	0
30−40	298	361	71	38	1
40−50	49	201	11	5	0
50−60	11	70	3	1	0
60−70	1	0	0	0	0

* Number of repeats

Table 5. Number of sequences of the LINE family in the porcine and cattle genomes.

	BovB/Art2	L3/CR1	LINE1	LINE2
Porcine genome	0	49	623	338
Cattle genome	457	38	539	275

genomes were less than 2.5%. The ratio of the DNA elements was less than 2.5%, except in the case of human DNA elements. The ratio of the simple repeats was less than 1.0%, except in the case of mouse simple repeats.

The distribution of the nucleotide substitution ratio obtained from the consensus sequence in the sampled sequences of the porcine genome is shown in Table 3. The peaks of the distribution of SINE, LINE, LTR element, and DNA element in the porcine genome were 10%－20%, 30%－40%, 20%－30%, and 20%－30%, respectively. The distribution of the nucleotide substitution ratio obtained from the consensus sequence in the sampled sequences of cattle genome is shown in Table 4. The peak of the distribution of SINE, LINE, LTR element, and DNA element in the bovine genome was the same as in the porcine genome. The results from porcine and bovine genome analyses also agree with those from human and mouse genome analyses[1,2].

The number of LINE families present in the porcine and bovine genomes is shown in Table 5. The BovB/Art2 family was located only in the bovine genome, and L3/CR1, LINE1, and LINE2 families were located in porcine and bovine genomes.

Discussion

The sampled genome sequences from the complete porcine and bovine genome sequences might include a bias because of the greater number of sequences around the gene. However, the difference in the proportion of SINE between complete human and mouse genomes suggest that the difference between porcine and cattle sampling genomes is reflected in the difference between the whole genomes. The peaks of the age distribution of the different classes of repeats were almost the same in the mammalian genome; however, the percentage of the total bases of SINE repeats was different. The results suggest that some mammalian species have genomes in which SINE repeats are copied and interspersed at a higher frequency.

It is thought that the LINE1, LINE2, and LINE3 families are widely located in the mammalian genom[10]. Our results indicated that the BovB/Art2 family is younger than the widely distributed LINEs. We can use the LINE1, LINE2, LINE3, and BovB/Art2 families to study the evolution of appropriate regions in the mammalian genomes.

As an example, the genome structure of the TNFα gene in cattle and humans is shown in Fig.1. The exon-intron structure of the TNFα gene was highly conserved in the cattle and human genomes. Two LINE/L2 repeats were located in the first intron of the cattle TNFα gene and in the 3' outer region of the human TNFα gene. The results suggest that recombination has occurred between the intron region of the TNFα gene from cattle and the 3' outer region of the TNFα gene from the human genome.

Understanding the genome history is useful for analyzing the expression and function of each gene in comparative research. However, in a genome, the regions excluding those that con-

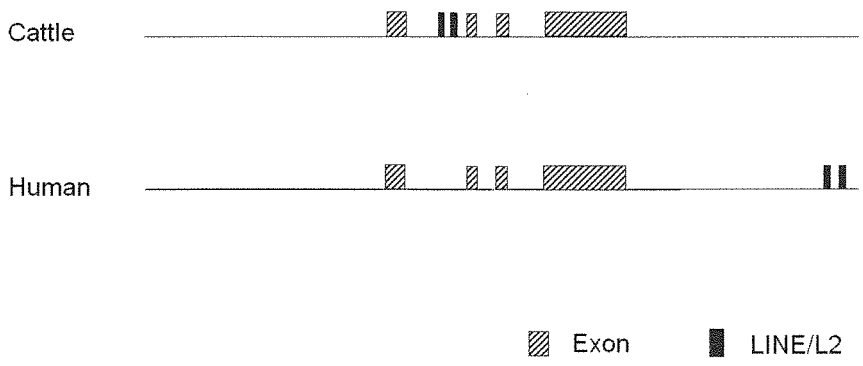

Fig. 1. Genome structure of the TNFα gene in the cattle and human genomes.

tain the genes are also important in evolution and gene function. For example, it is suggested that the difference in the promoter region of a gene results in different expression patterns between two animals.

Recombination of the genomic region can be investigated using SINEs, LINEs, and other repeats, which have different age distribution. If recombination is detected, we might be able to predict the change in the expression pattern of a gene in another animal in which the genomic region has not been sequenced. Therefore, our results on the age distribution of SINE, LINE, and other interspersed repeats are useful in the studies on genome evolution.

Acknowledgements

The part of this study was carried out under ISM Cooperative Research Program (2002-ISM-2038).

References

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
2. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.
3. Esnault C., Maestre J. & Heidmann T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics* **24**: 363-367.
4. Singer D.S., Parent L.J. & Ehrlich R. (1987) Identification and DNA sequence of an interspersed repetitive DNA element in the genome of the miniature swine. *Nucleic Acids Research* **15**: 2780
5. Shimamura M., Yasue H., Ohshima K., Abe H., Kato H., Kishiro T., Goto M., Munechika I. & Okada N (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* **388**: 666-670
6. Nomura O., Lin Z.H.,Muladno, Wada Y. & Yasue H. (1998) A SINE species from hippopotamus and its distribution among animal species. *Mamm. Genome* **9**: 550-555.
7. Lin Z., Nomura O., Hayashi T., Wada Y. & Yasue H. (2001) Characterization of a SINE species from vicuna and its distribution in animal species including the family *Camelidae. Mamm. Genome* **12**: 305-308.

8. Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**: 418-420.
9. Tajima, F. & Nei, M. (1984) Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* **1**: 269-285.
10. Okada N., Hamada M., Ogiwara I. & Ohshima K. (1997) SINEs and LINEs share common 3' sequences: a review. *Gene* **205**: 229-243

ブタおよびウシゲノムにおける反復配列の In Silico 分析

和田　康彦・重森　卓人
（動物生産学分野）
平成17年９月30日　受理

ヒトゲノムやマウスゲノムの解析によって，哺乳動物ゲノム配列の３割以上が反復配列によって占められていることがあきらかとなってきた．しかしながら，これらの膨大な数の反復配列の進化の歴史と生命体の中での機能についてはほとんど明らかになっていない．そこで，Genbank 塩基配列データベース中のブタおよびウシのゲノム配列から反復配列を網羅的に検索し，反復配列クラスごとの存在割合とコンセンサス配列からの塩基置換率の分布を調べた．

その結果，ブタを除いて哺乳動物ゲノム中の反復配列では LINE が最も多く存在しており，LINE は全ゲノム配列中の14－21％を占めていた．次に SINE が８％－14％，LTR が２％－10％であったが，動物種によってそれぞれの比率は大きく異なっていた．

コンセンサス配列からの塩基置換率のピークは SINE で10％－20％，LINE で30％－40％，LTR で20％－30％と，他の反復配列よりも SINE はより塩基置換率の小さな配列が多かった．ブタとウシで LINE の種類を比較したところ BovB/Art2はウシゲノムにのみ存在していたが，L3/CR1，LINE1および LINE2はブタゲノムにもウシゲノムにもともに存在していた．これらの研究結果は，SINE や LINE などの反復配列を用いてゲノムの組み換えを推測することを可能にするものであり，哺乳動物のゲノム進化の研究に大いに役立つものと考えられた．