Saitou N. (1992) In Takahata N. ed.,
Population Paleo-Genetics: Proceedings of the Seventeenth Taniguchi Symposium. Japan
Science Society Press, Tokyo, pp. 335-343

Evolutionary Rate of Insertions and Deletions in Nucleotide Sequences

NARUYA SAITOU

Laboratory of Evolutionary Genetics, National Institute of Genetics, Mishima 411, Japan

Evolutionary change of nucleotide sequences has been studied mainly from the viewpoint of nucleotide substitutions. Now we have ample knowledge of the pattern and rate of nucleotide substitutions for many organisms and genes (*e.g.*, Nei, 1987). Insertions and deletions of nucleotides are, however, not well studied. One reason is that they do not occur frequently in the coding regions because most of them are strongly deleterious.

Saitou (1991) constructed a molecular phylogeny of hominoids from DNA sequence data in which pseudogenes and spacer regions were included. In this study, we analyzed insertions and deletions in the nucleotide sequences of these non-coding regions for higher primates by utilizing the result of Saitou's (1991) study. We show that insertions and deletions occur rather frequently in the non-coding region of higher primate genomes, and estimated the evolutionary rate of insertions and deletions, which turned to be more or less uniform among different evolutionary lineages. This paper is based on a part of N. Saitou and S. Ueda's unpublished study.

MATERIALS AND METHODS

We used four different sets of nucleotide sequence data. These are as follows. (1) η-globin (short): a 2.0-kb fragment containing the η-globin pseudogene in the β-globin gene family. Aligned sequence data for human, chimpanzee, gorilla, orangutan, rhesus monkey, and owl monkey are taken from Koop *et al.* (1986). (2) β-globin: a 3.1-kb fragment containing a spacer DNA of the β-globin gene family. Aligned sequence data for human (T allele), chimpanzee, gorilla, orangutan, rhesus monkey, and spider monkey are taken from Maeda *et al.* (1986). (3) η-globin (long): a 7.1-kb fragment containing the η-globin pseudogene that includes η-globin (short) sequences above. Aligned sequence data for human, chimpanzee, gorilla, and orangutan are taken from Miyamoto *et al.* (1987). (4) Ig-ε3: a 2.3-kb fragment containing immunoglobulin ε3 processed pseudogene. Aligned sequence data for human, chimpanzee, gorilla, and orangutan are taken from Ueda *et al.* (1989).

Two methods were used for analyzing those data. One is Tajima and Nei's (1984) method for estimating evolutionary distance (D_{XY}) between sequences X and Y due to insertions and deletions. D_{XY} is given by

$$D_{XY} = -2 \log (n_{xy}/\sqrt{n_X n_Y}), \qquad (1)$$

where n_{XY} is the number of homologous nucleotides between sequences X and Y, and n_X, and n_Y are the total number of nucleotides in sequences X and Y, respectively (Tajima and Nei, 1984).

The other method is slightly complicated. We first assume the molecular phylogeny of sequences used. This tree is taken from Saitou (1991) (see Fig. 3). Based on this tree, the maximum parsimony method (see Sneath and Sokal, 1973) is applied to locate insertions and deletions to each branch of the phylogenetic tree. In this case, the number of nucleotides involved in insertions and deletions is irrelevant. A contiguous block of nucleotides or a gap is considered to be created by one event.

RESULTS AND DISCUSSION

1. Analysis Based on the Evolutionary Distance Due to Insertions and Deletions

D_{XY} matrices for η-globin (short) and β-globin sequences for six primate species are presented in Table I. D_{XY} values for η-globin (short) and β-globin are more or less similar for the same set of pairs of species. However, those between rhesus macaque and hominoid species (human, chimpanzee, gorilla, and orangutan) for η-globin (short) sequences are much larger than those for β-globin sequences. This is apparently because of the existence of two rather long gaps specific either to hominoids (1,610–1,633 bp) or to non-hominoids (245–282 bp) for η-globin (short) sequences.

Figure 1 shows two phylogenetic trees of higher primates based on the D_{XY}

TABLE I

Evolutionary Distance Matrices (Above Diagonal for β-Globin and Below Diagonal for η-globin Short) Due to Insertions and Deletions for Six Primate Species (taken from Saitou and Ueda, unpublished)

Human	—	0.0057	0.0070	0.0101	0.0215	0.0601
Chimpanzee	0.0047	—	0.0051	0.0070	0.0183	0.0584
Gorilla	0.0056	0.0056	—	0.0057	0.0164	0.0564
Orangutan	0.0240	0.0240	0.0192	—	0.0139	0.0538
Rhesus m.	0.0925	0.0927	0.0876	0.0763	—	0.0536
New World m.[a]	0.0559	0.0560	0.0502	0.0628	0.0636	—
	Human	Chim-panzee	Gorilla	Orang-utan	Rhesus macaque	New W. monkey

[a] Spider monkey and owl monkey in the case of β-globin and η-globin short sequences, respectively.

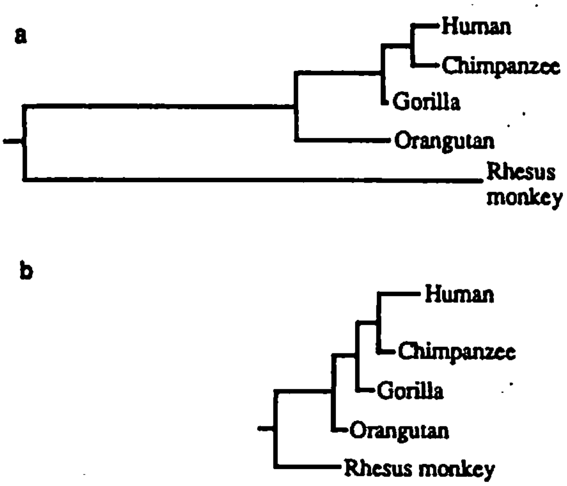

Fig. 1. Phylogenetic trees of higher primates reconstructed by using the neighbor-joining method from the evolutionary distances due to insertions and deletions. a: η-globin (short) gene region. b: β-globin spacer region (taken from Saitou and Ueda, unpublished).

matrices of Table I by using the neighbor-joining method (Saitou and Nei, 1987). New World monkey sequences are assumed to be outgroups to locate the root of trees. It is interesting to note that the branching pattern of the both trees is identical with each other, and that pattern is also identical with those of Fig. 1. Especially because of the large D_{XY} values between rhesus macaque

TABLE II

Evolutionary·Distance Matrices (Above Diagonal for β-Globin and Below Diagonal for η-Globin Short) Due to Nucleotide Substitutions for Human, Chimpanzee, and Orangutan (taken from Table I of Saitou, 1991).

Human	—	0.0140	0.0157	0.0307	0.0774	0.1169
Chimpanzee	0.0123	—	0.0123	0.0255	0.0742	0.1118
Gorilla	0.0144	0.0181	—	0.0269	0.0731	0.1135
Orangutan	0.0308	0.0346	0.0373	—	0.0701	0.1138
Rhesus m.	0.0734	0.0780	0.0781	0.0785	—	0.1331
NewWorld m.[a]	0.1133	0.1145	0.1170	0.1174	0.1369	—
	Human	Chimpanzee	Gorilla	Orangutan	Rhesus macaque	New W. monkey

[a] Spider monkey and owl monkey in the case of β-globin and η-globin short sequences, respectively.

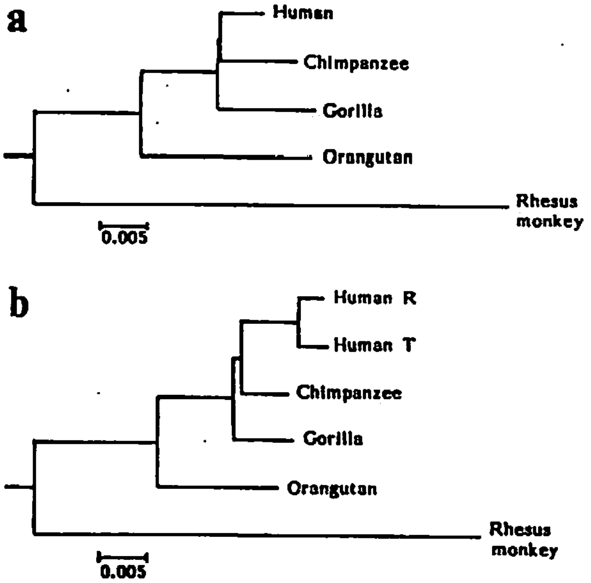

Fig. 2. Phylogenetic trees of higher primates reconstructed by using the neighbor-joining method from the evolutionary distances due to nucleotide substitutions. a: η-globin gene region. b: β-globin spacer region (taken from Fig. 1 of Saitou, 1991).

and hominoid species for η-globin (short) sequences, however, lengths of corresponding branches of these two trees are considerably different.

Compared to the large difference between η-globin (short) and β-globin

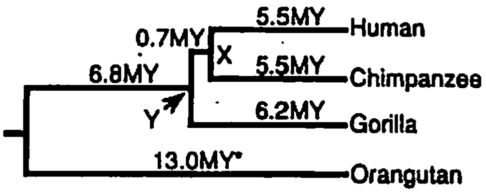

Fig. 3. A molecular phylogeny of hominoids estimated by using UPGMA from the nuclear DNA sequence data (modified from Fig. 2 of Saitou, 1991).

```
a              b
H:AC----GC     H:AAAAAAAT
C:AC----GC     C:AAAAAA-T
G:AC----GC     G:AAAAA--T
O:GCAAATGC     O:AAAG---T
```

Fig. 4. Two examples of the gaps in the aligned sequences. a: 731–738 bp of the Igε gene. b: 3,041–3,048 bp of the η-globin (long) gene.

sequences observed in Table I, evolutionary distances due to nucleotide substitutions are rather similar between these two data set (see Table II). Figure 2 shows two phylogenetic trees based on the distance matrices of Table II by using the neighbor-joining method. It is clear that the branching patterns are identical in both trees and that the lengths of corresponding branches of these two trees are more or less similar to each other.

2. *Analysis Based on the Maximum Parsimony Method*

We use Fig. 3 as the given phylogenetic tree. This tree was constructed by using UPGMA (Sneath and Sokal, 1973), since the approximate constancy of the rate of molecular evolution was observed for these four species (Saitou, 1991). The divergence time of 13 million years (MY) between human and orangutan is assumed to calibrate the molecular clock. Thus the branch between node X and human corresponds to 5.5MY, that between nodes X and Y corresponds to 0.7MY, and so on (see Fig. 3).

Each gap was then allocated to the appropriate branch of the tree by applying the maximum parsimony principle. For example, there is a gap at 731–738 bp of Ig-ε sequences (see Fig. 4a). In this case, one gap is unambiguously allocated to the branch between node Y and orangutan of Fig. 3. It should be noted that the root of the tree of Fig. 3 is ignored in this operation.

When there is more than one possibility for allocation of gaps to branches, these possible gaps are allocated to corresponding branches equally. For

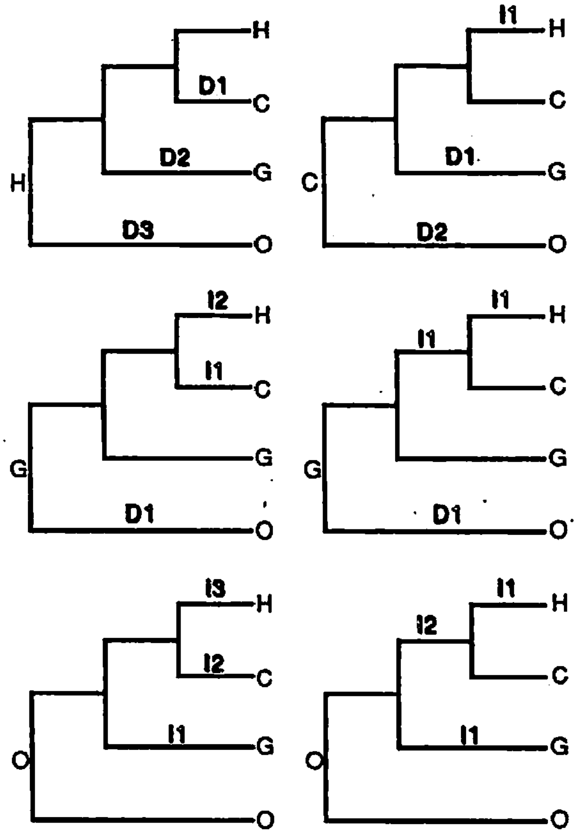

Fig. 5. Six possible ways to allocate insertion/deletion events from a multiple-aligned sequence data to each branch of a phylogenetic tree (taken from Saitou and Ueda, unpublished). H, C, G, and O stand for human, chimpanzee, gorilla, and orangutan, respectively. One of these four characters is given at the root of the tree to show the assumed sequence of the common ancestor. I and D denote insertion and deletion events, respectively. Numbers after I or D are the number of nucleotides involved in the corresponding insertion or deletion.

example, there are multiple gaps at 3,041–3,048 bp of η-globin (long) sequences (see Fig. 4b). In this case, there are six possible patterns of insertions and deletions (Fig. 5). We allocate all these possible insertions and deletions to the corresponding branches with a probability of 1/6.

Table III is a summary of the result. Sequence data for η-globin (long), β-globin, and Ig-ε sequences were used. Figures are the numbers of gaps for each branch that were determined unambiguously, and figures in parentheses

TABLE III

Numbers of Insertion/Deletion Events at Each Branch of the Phylogenetic Tree of Four Hominoids (taken from Saitou and Ueda, unpublished)

Branch	η-Globin 7,019 bp	β-Globin 3,167 bp	Ig-ε3 2,341 bp	Total 12,527 bp	Rate[a]
X-human	7(7.9)	4(4.7)	2(3.5)	13(16.1)	2.9
X-chim.	4(4.5)	4(4.7)	2(3.3)	10(12.4)	2.3
Y-gorilla	5(7.8)	2(2.5)	1(2.5)	8(12.8)	2.1
Y-orang.	25(26.6)	4(4.9)	5(6.5)	34(38.0)	1.9
X-Y	4(5.3)	1(1.3)	1(1.3)	6(7.8)	11.1
Total	45(52)	15(18)	11(17)	71(87)	2.3
Rate[b]	7.4	5.7	7.3	6.9	0.18

[a] Rate of gap generation per million years for 12,527 bp region.

[b] Rate of gap generation per 1,000 bp for the total of 37.7 million years.

are those with multiple possibilities are added. A total of 87 gaps were counted for all the three data sets. Seventy one gaps were unambiguously allocated to specific branches.

Evolutionary rate of gap creation is estimated for each branch and for each sequence region, and these rates are given at the last row and column of Table III, respectively. We first note that evolutionary rates for three different sequence regions are more or less the same, being around 6-7 gaps per 1,000 bp for the total of 37.7 million years for the phylogenetic tree of Fig. 3. This is a clear contrast to the result based on the evolutionary distance due to insertion and deletion.

Interestingly enough, the evolutionary rate for each branch of the tree is also more or less similar. The rate is around 2-3 gaps per million years for a total of 12,527 bp-long sequences. There is, however, an exception. The rate (11.1/MY/12,527 bp) for the branch between nodes X and Y of Fig. 3 is far greater than that of the remaining branches.

We have two explanations for this discrepancy. One seeks for an evolutionary significance. Because of some unknown mechanism the evolutionary rate of gap creation was higher for the branch X-Y than the other branches. The other explanation is concerned with the alignment process. Because the maximum parsimony principle is used in the alignment of sequences, gaps "observed" can merely be artefact. Two examples of gaps in the aligned sequences are shown in Fig. 6. In both cases, we can have alternative, less parsimonious alignments. If these alternative alignments reflect the true evolutionary history of these sequences, gaps found in the new alternative alignments are no longer allocated to the branch X-Y.

```
a
H:AC----AATC        H:AC----AATC
C:AC----AATC   ?    C:ACAAT----C
G:ACAATTAATC  --->  G:ACAATTAATC
O:ACAATTAATC        O:ACAATTAATC

b
H:TCTTTTTTAT        H:TCTTTTTTAT
C:TCTTTTTTAT   ?    C:TCTTTTTTAT
G:TCTTTTT-AT  --->  G:TCTTTTT-AT
O:TCTTTTT-AT        O:TCTTTT-TAT
```

Fig. 6. Two examples of alternative alignment. a: 1,489-1,498 bp of the η-globin (short) gene. b: 3,265-3,274 bp of the η-globin (long) gene. In both cases, alignments at the left are original, most-parsimonious ones, and those at the right are alternative, non-parsimonious alignments.

It is interesting to note that a repetition of AAT and T are observed in Fig. 6(a and b), respectively. Such repetitions can cause the frequent replication slippage, and the most parsimonious alignments may not be the correct representation of the evolutionary history.

In any case, the overall rate of gap creation per 1,000 bp per million years is estimated to be 0.18 (see Table III). This estimate of the rate seems to be the first one for gap creation, or for insertion/deletion events. The evolutionary rate of nucleotide substitution for the DNA region analyzed in the present study was estimated to be 1.25×10^{-9} per site per year, and the corresponding rate of gap creation is 0.18×10^{-9} per site per year. Therefore, nucleotide substitutions seem to occur about 7 times higher than gap creations in non-coding regions of our genomes.

SUMMARY

Gaps created through alignment are routinely eliminated when we compare nucleotide sequences, and nucleotide substitutions have been extensively studied. However, insertions and deletions that are responsible for gaps in aligned sequences seem to occur rather frequently especially in non-coding regions. We studied the rate of insertions and deletions in pseudogene sequences by using (1) Tajima and Nei's evolutionary distance due to insertions and deletions and (2) the maximum parsimony principle to locate insertion/deletion events on a phylogenetic tree. The rate of insertion/deletion events was found to be rather constant.

Acknowledgments

The author thanks Dr. Shintaroh Ueda for valuable discussions. This study was supported by grants-in-aid from the Ministry of Education, Science, and Culture of Japan.

REFERENCES

Koop, B.F., M. Goodman, P. Xu, K. Chan, and J.L. Slighton, 1986. *Nature* 319: 234-238.

Maeda, N., C.I. Wu, J. Bliska, and J. Reneke, 1988. *Mol. Biol. Evol.* 5: 1-20.

Miyamoto, M.M., J.L. Slighton, and M. Goodman, 1987. *Science* 238: 369-373.

Saitou, N., 1991. *Am. J. Phys. Anthropol.* 84: 75-85.

Saitou, N. and M. Nei, 1987. *Mol. Biol. Evol.* 4: 406-425.

Sneath, P. H. P. and R. Sokal, 1973. *Numerical Taxonomy.* W.H. Freeman, San Francisco.

Tajima, F. and M. Nei, 1984. *Mol. Biol. Evol.* 1: 269-285.

Ueda, S., Y. Watanabe, N. Saitou, K. Omoto, H. Hayashida, T. Miyata, H. Hisajima, and T. Honjo, 1989. *J. Mol. Biol.* 205: 85-90.