Genes Genet. Syst. (2011) **86**, p. 123–129

Overestimation of nonsynonymous/synonymous rate ratio by reverse-translation of aligned amino acid sequences

Yoshiyuki Suzuki*

Graduate School of Natural Sciences, Nagoya City University, 1 Yamanohata, Mizuho-cho, Mizuho-ku, Nagoya-shi, Aichi-ken 467-8501, Japan

(Received 5 January 2011, accepted 17 March 2011)

In the analysis of protein-coding nucleotide sequences, the ratio of the number of nonsynonymous substitutions to that of synonymous substitutions (d_N/d_S) is used as an indicator for the direction and magnitude of natural selection operating at the amino acid sequence level. The d_S and d_N values are estimated based on the comparison of homologous codons, which are often identified by converting (reverse-translating) aligned amino acid sequences into codon sequences. In this method, however, homologous codons may be mis-identified when frame-shifts occurred or amino acid sequences were mis-aligned, which may lead to overestimation of the d_N/d_S ratio. Here the effect of reverse-translating aligned amino acid sequences on the estimation of d_N/d_S ratio was examined through a large-scale analysis of protein-coding nucleotide sequences from vertebrate species. Apparently, 1–9% of codon sites that were identified as homologous with reverse-translation contained non-homologous codons, where the d_N/d_S ratio was unduly high. By correcting the d_N/d_S ratio for these codon sites, it was inferred that the ratio was 5–43% overestimated with reverse-translation. These results suggest that caution should be exerted in the study of natural selection using the d_N/d_S ratio by reverse-translating aligned amino acid sequences.

Key words: nonsynonymous/synonymous rate ratio, reverse-translation, alignment, negative selection, positive selection

INTRODUCTION

Point mutations occurring in protein-coding nucleotide sequences are synonymous or nonsynonymous according to whether they retain or alter coding amino acids, respectively (Miyata and Yasunaga, 1980; Perler et al., 1980). Synonymous mutations are considered to be selectively neutral or nearly neutral, where the rate of synonymous substitution (r_S) may reflect the mutation rate. In contrast, nonsynonymous mutations are considered to be subject to natural selection operating at the amino acid sequence level. Since fixation probabilities of advantageous and deleterious mutations are higher and lower than that of neutral mutations, respectively, the rate of nonsynonymous substitution (r_N) may be greater and smaller than the mutation rate when positive and negative selection operates, respectively. Therefore, natural selection can be detected by comparing r_S and r_N, where $r_S < r_N$, $r_S > r_N$, and $r_S = r_N$ indicate positive, negative, and no selection, respectively (Kimura, 1977).

The comparison of r_S and r_N can be performed by comparing the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions that have accumulated during the same evolutionary time period (t), because it is expected that $d_S = r_S t$ and $d_N = r_N t$ (Hughes and Nei, 1988). The nonsynonymous/synonymous rate ratio (r_N/r_S) is estimated by d_N/d_S, which reflects the direction and magnitude of natural selection. In haploid organisms, $d_N/d_S = 2N_e s/(1 - e^{-2N_e s})$, where N_e and s denote the effective population size and the selection coefficient, respectively (Nielsen and Yang, 2003). The estimation of d_S and d_N is based on the comparison of homologous codons (Nei and Kumar, 2000; Suzuki and Gojobori, 2003). In the real data analysis, homologous codons are often identified by making a multiple alignment of amino acid sequences and converting it into codon sequences (Suyama et al., 2006; Wong et al., 2008; Schneider et al., 2009; Fletcher and Yang, 2010). This conversion process is called reverse-translation in this paper. The alignment of codon sequences obtained is usually treated as an observation without errors (Wong et al., 2008).

Edited by Hidenori Nishihara
* Corresponding author. E-mail: yossuzuk@nsc.nagoya-cu.ac.jp

The above approach, however, appears to contain problems that may result in mis-identification of homologous codons. First, it is implicitly assumed that the unit of insertions and deletions (indels) in protein-coding nucleotide sequences is the codon (Fletcher and Yang, 2010), which is not always the case in reality because frame-shifts can occur at least partially (Mills et al., 2006). When a frame-shift occurs at some codons in a sequence, they are no longer homologous to codons in other sequences. Yet, non-homologous codons may be identified as homologous in the above approach, because it is difficult to infer the occurrence of frame-shifts only from the comparison of amino acid sequences. Second, even when the unit of indels was the codon, aligning amino acid sequences itself is not always easy, and non-homologous amino acids may be aligned especially at variable sites, which may lead to mis-identification of homologous codons (Liu et al., 2009; Fletcher and Yang, 2010). It has been reported that excluding the sites with gaps from the alignment of amino acid sequences was insufficient to reduce alignment errors in the real data analysis, suggesting that mis-alignment of amino acid sequences may be common (Wong et al., 2008; Fletcher and Yang, 2010).

In general, majority of amino acid sites in proteins are under functional constraint with $d_N/d_S < 1$ (Suzuki and Gojobori, 2001; Suzuki, 2006). However, the d_N/d_S ratio is known to be inflated at mis-aligned codon sites (Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009). It has also been reported that as more non-homologous codons are aligned, more amino acid sites are falsely identified as positively selected (Vamathevan et al., 2008; Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009; Fletcher and Yang, 2010). The purpose of the present study was to examine the effect of reverse-translating aligned amino acid sequences on the estimation of d_N/d_S ratio, through a large-scale analysis of protein-coding nucleotide sequences from vertebrate species.

MATERIALS AND METHODS

Sequence data The entire sets of protein-coding nucleotide sequences for 10 vertebrate species (human [*Homo sapiens*; GRCh37], chimpanzee [*Pan troglodytes*; CHIMP2.1], orangutan [*Pongo pygmaeus abelii*; PPYG2], macaque [*Macaca mulatta*; MMUL_1.0], mouse [*Mus musculus*; NCBIM37], cow [*Bos taurus*; Btau_4.0], opossum [*Monodelphis domestica*; monDom5], chicken [*Gallus gallus*; WASHUC2], frog [*Xenopus tropicalis*; JGI4.1], and zebrafish [*Danio rerio*, Zv8]) (Nei et al., 2010) were retrieved from Ensembl Genes 57 through BioMart (Durinck et al., 2005). Distantly related species, such as chicken, frog, and zebrafish, were included in the analysis of d_N/d_S ratio in the present study for the following reasons. First, the d_S and d_N values did not appear to be saturated but increased linearly along with time for these species (Nei et al., 2010). Second, even for distantly related species, the d_S and d_N values are considered to be estimated reliably when the number of codon sites analyzed is relatively large (Nei and Kumar, 2000). Third, natural selection has been detected based on the d_N/d_S ratio even when distantly related species, such as chicken and frog, were included (Uddin et al., 2008; Goodman et al., 2009; Goodman and Sterner, 2010).

The possible orthology data for human sequences to sequences of other vertebrate species were also available in BioMart. It should be noted, however, that the possible orthology data in BioMart were generated based on the topology of the phylogenetic tree constructed from the multiple alignment of nucleotide sequences that was obtained by reverse-translating aligned amino acid sequences (Vilella et al., 2009). If homologous nucleotides were mis-identified in this process, the number of nucleotide substitutions may be overestimated for some pairs of sequences, which may result in construction of incorrect topology. It was therefore possible that orthologues were identified as paralogues, and *vice versa*. Although the identification of paralogues as orthologues may be problematic in the present study, the probability for the occurrence of mis-identification appears to be small, because it is unlikely that a particular topology (species tree) is generated by random effects. In addition, the probability may be further reduced by focusing only on one-to-one possible orthologues between species.

Data processing Using the possible orthology data, a list of one-to-one possible orthologues was generated between human and other vertebrate species. Nine lists of one-to-one possible orthologues obtained were combined using human sequences as the reference, to generate 4,313 sets of possible orthologues that were shared by 10 vertebrate species. The sets of possible orthologues whose member sequence contained a premature termination codon or an ambiguous nucleotide were discarded, and 3,878 sets of possible orthologues were retained for the next step.

For each of 3,878 sets of possible orthologues, multiple alignments of amino acid and nucleotide sequences for 10 vertebrate species were made by using the computer program CLUSTAL W (version 1.8) (Thompson et al., 1994) with the default parameter settings. The alignment of amino acid sequences was reverse-translated into codon sequences, and the alignment of codon sequences obtained was compared with the alignment of nucleotide sequences. The codon sites that were aligned consistently in these alignments for all of 10 vertebrate species (class-1 codon sites) were extracted to construct another alignment of codon sequences. It should be noted that the class-1 sites represent the codon sites that were aligned consistently using the amino acid and nucleotide

sequences. The sets of possible orthologues for which the number of class-1 codon sites was < 100 were discarded to reduce the possibility that they do not encode a real protein, and 3,325 sets of possible orthologues were retained for the next step.

For each of 3,325 sets of possible orthologues, two alignments of codon sequences generated above, by reverse-translating aligned amino acid sequences and by extracting class-1 sites, were used for estimating the d_S and d_N values between human and other vertebrate species by the method of Nei and Gojobori (1986) taking into account the transition/transversion rate ratio (Kondo et al., 1993; Zhang et al., 1998; Suzuki et al., 2009), which has been estimated to be 4 in mammals (Rosenberg et al., 2003; Jiang and Zhao, 2006; Zhang et al., 2007). The codon sites shared by all of 10 vertebrate species without gaps were used for the estimation. The sets of possible orthologues for which the d_S or d_N value between human and any of other vertebrate species was incalculable were discarded to reduce the possibility that they contained paralogous sequences. Finally, the remaining 3,222 sets of possible orthologues were considered as orthologues and used for the analysis of d_N/d_S ratio.

Analysis of d_N/d_S ratio Two alignments of codon sequences generated above, by reverse-translating aligned amino acid sequences and by extracting class-1 sites, were concatenated, after eliminating the codon sites with gaps, for 3,222 sets of orthologues to make the alignments of codon sequences with 1,318,081 codon sites and 1,128,326 codon sites, respectively. Using these alignments, the d_S and d_N values as well as the d_N/d_S ratio were estimated between human and other vertebrate species, as described above.

RESULTS AND DISCUSSION

d_N/d_S ratio by reverse-translating aligned amino acid sequences The d_S and d_N values as well as the d_N/d_S ratio estimated between human and other vertebrate species using the alignment of codon sequences constructed by reverse-translating aligned amino acid sequences are summarized in Table 1. The d_N/d_S ratio between human and non-human primates ranged from 0.260 to 0.272 (The Chimpanzee Sequencing and Analysis Consortium, 2005; Bakewell et al., 2007; Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007). In contrast, smaller d_N/d_S ratios, ranging from 0.131 to 0.217, were observed between human and non-primate mammals and non-mammalian vertebrates (Mouse Genome Sequencing Consortium, 2002; International Chicken Genome Sequencing Consortium, 2004; Rat Genome Sequencing Project Consortium, 2004).

It should be noted that the effect of natural selection is positively correlated with the effective population size of organisms (Kimura, 1983). Therefore, the difference in the d_N/d_S ratio observed above appears to reflect the fact that effective population sizes of primates are smaller than those of other vertebrate species analyzed in the present study. In fact, the effective population size has been estimated to be ~10,000 for human (Takahata, 1993), ~25,000 for chimpanzee (Won and Hey, 2005), ~15,000 for orangutan (Becquet and Przeworski, 2007), and ~25,000 for macaque (Bonhomme et al., 2009), which are smaller than the estimates of ~400,000 for mouse (Geraldes et al., 2008) and ~90,000 for cow (MacEachern

Table 1. The d_S and d_N values and the d_N/d_S ratio between human and other vertebrate species

Species	Reverse-translated[a]			Class-1[b]			Class-2-1[c]			Class-2-2[d]			Estimated[e]		
	d_S	d_N	d_N/d_S	d_S	d_N	d_N/d_S	d_S	d_N	d_N/d_S	d_S	d_N	d_N/d_S	d_S	d_N	d_N/d_S
Chimpanzee	0.0125[f]	0.00337	0.270	0.0114	0.00195	0.171	0.0122	0.00451	0.369	0.0989	0.101	1.02	0.0115	0.00232	0.201
Orangutan	0.0354	0.00921	0.260	0.0321	0.00496	0.154	0.0339	0.0114	0.337	0.265	0.277	1.05	0.0324	0.00589	0.182
Macaque	0.0678	0.0185	0.272	0.0616	0.0101	0.164	0.0635	0.0223	0.351	0.478	0.525	1.10	0.0619	0.0118	0.191
Mouse	0.436	0.0571	0.131	0.431	0.0430	0.0997	0.442	0.124	0.280	0.668	0.375	0.561	0.433	0.0541	0.125
Cow	0.311	0.0497	0.160	0.303	0.0359	0.118	0.320	0.0962	0.301	0.654	0.499	0.762	0.306	0.0443	0.145
Opossum	0.666	0.105	0.158	0.655	0.0783	0.120	0.675	0.209	0.310	1.04	0.797	0.764	0.658	0.0957	0.146
Chicken	0.909	0.158	0.173	0.899	0.121	0.135	0.907	0.286	0.315	1.18	0.992	0.844	0.900	0.143	0.159
Frog	1.27	0.232	0.183	1.27	0.177	0.140	1.20	0.382	0.319	1.34	1.34	1.00	1.26	0.203	0.162
Zebrafish	1.39	0.303	0.217	1.40	0.238	0.171	1.30	0.369	0.284	1.42	1.48	1.04	1.38	0.256	0.185

[a] Alignment of codon sequences was constructed by reverse-translating aligned amino acid sequences.
[b] Alignment of codon sequences was constructed by extracting the codon sites that were aligned consistently using the amino acid and nucleotide sequences for all of 10 vertebrate species.
[c] Codon sites that were aligned inconsistently using the amino acid and nucleotide sequences for any of 10 vertebrate species but consistently for the pairwise comparison of human and other vertebrate species.
[d] Codon sites that were aligned inconsistently using the amino acid and nucleotide sequences for the pairwise comparison of human and other vertebrate species.
[e] Estimated values of d_S and d_N as well as the d_N/d_S ratio by correcting the d_N/d_S ratio for the class-2-2 sites.
[f] Standard errors were mostly more than two orders of magnitude smaller than the estimates.

et al., 2009). These observations suggest that functional constraint has operated less effectively in primates compared to other vertebrate species.

d_N/d_S ratio by extracting the codon sites aligned consistently using the amino acid and nucleotide sequences In the above analysis, the alignment of codon sequences was constructed by reverse-translating aligned amino acid sequences. In this method, however, non-homologous codons may be aligned when frame-shifts occurred or amino acid sequences were mis-aligned, which may lead to overestimation of the d_N/d_S ratio, as discussed above. It may be difficult to measure the degree of overestimation accurately in the real data analysis, because the correct alignment of codon sequences is usually unknown. However, the codon sites that are

Table 2. The numbers of synonymous and nonsynonymous sites and differences, and the proportions of different sites between human and other vertebrate species

Species		Reverse-translated[a]			Class-1[b]			Class-2-1[c]			Class-2-2[d]		
		Site	Difference	Proportion	Site	Difference	Proportion	Site	Difference	Proportion	Site	Difference	Proportion
Chimpanzee	Synonymous	1,119,714	13,870	0.0124[f]	956,930	10,831	0.0113	149,568	1,814	0.0121	13,215	1,225	0.0927
		(1.00)[e]	(1.00)		(0.855)	(0.781)		(0.134)	(0.131)		(0.0118)	(0.0883)	
	Nonsynonymous	2,697,860	9,075	0.00336	2,310,381	4,503	0.00195	355,994	1,601	0.00450	31,485	2,972	0.0944
		(1.00)	(1.00)		(0.856)	(0.496)		(0.132)	(0.176)		(0.0117)	(0.327)	
Orangutan	Synonymous	1,119,651	38,722	0.0346	956,874	30,100	0.0315	145,807	4,839	0.0332	16,971	3,784	0.223
		(1.00)	(1.00)		(0.855)	(0.777)		(0.130)	(0.125)		(0.0152)	(0.0977)	
	Nonsynonymous	2,697,941	24,705	0.00916	2,310,419	11,430	0.00495	347,169	3,942	0.0114	40,353	9,332	0.231
		(1.00)	(1.00)		(0.856)	(0.463)		(0.129)	(0.160)		(0.0150)	(0.378)	
Macaque	Synonymous	1,119,917	72,613	0.0648	957,001	56,605	0.0591	142,108	8,649	0.0609	20,808	7,358	0.354
		(1.00)	(1.00)		(0.855)	(0.780)		(0.127)	(0.119)		(0.0186)	(0.101)	
	Nonsynonymous	2,697,811	49,235	0.0183	2,310,356	23,146	0.0100	338,014	7,421	0.022	49,442	18,669	0.378
		(1.00)	(1.00)		(0.856)	(0.47)		(0.125)	(0.151)		(0.0183)	(0.379)	
Mouse	Synonymous	1,121,273	370,520	0.330	958,166	314,142	0.328	145,223	48,473	0.334	17,884	7,906	0.442
		(1.00)	(1.00)		(0.855)	(0.848)		(0.130)	(0.131)		(0.0160)	(0.0213)	
	Nonsynonymous	2,697,505	148,365	0.055	2,310,105	96,527	0.0418	345,144	39,373	0.114	42,257	12,465	0.295
		(1.00)	(1.00)		(0.856)	(0.651)		(0.128)	(0.265)		(0.0157)	(0.0840)	
Cow	Synonymous	1,121,675	285,406	0.254	958,461	239,172	0.250	141,899	36,929	0.26	21,316	9,305	0.437
		(1.00)	(1.00)		(0.854)	(0.838)		(0.127)	(0.129)		(0.0190)	(0.0326)	
	Nonsynonymous	2,697,163	129,798	0.0481	2,309,825	81,054	0.0351	337,044	30,428	0.0903	50,294	18,316	0.364
		(1.00)	(1.00)		(0.856)	(0.624)		(0.125)	(0.234)		(0.0186)	(0.141)	
Opossum	Synonymous	1,116,628	492,847	0.441	954,113	416,835	0.437	131,183	58,364	0.445	31,333	17,647	0.563
		(1.00)	(1.00)		(0.854)	(0.846)		(0.117)	(0.118)		(0.0281)	(0.0358)	
	Nonsynonymous	2,699,964	265,535	0.0983	2,312,037	171,929	0.0744	313,841	57,252	0.182	74,086	36,355	0.491
		(1.00)	(1.00)		(0.856)	(0.647)		(0.116)	(0.216)		(0.0274)	(0.137)	
Chicken	Synonymous	1,117,171	588,412	0.527	954,779	499,984	0.524	118,277	62,244	0.526	44,115	26,184	0.594
		(1.00)	(1.00)		(0.855)	(0.85)		(0.106)	(0.106)		(0.0395)	(0.0445)	
	Nonsynonymous	2,701,419	384,051	0.142	2,313,190	259,366	0.112	284,564	67,644	0.238	103,666	57,040	0.550
		(1.00)	(1.00)		(0.856)	(0.675)		(0.105)	(0.176)		(0.0384)	(0.149)	
Frog	Synonymous	1,113,482	680,724	0.611	951,870	582,089	0.612	87,910	52,603	0.598	73,702	46,033	0.625
		(1.00)	(1.00)		(0.855)	(0.855)		(0.0790)	(0.0773)		(0.0662)	(0.0676)	
	Nonsynonymous	2,704,662	539,354	0.199	2,315,566	365,360	0.158	212,404	63,605	0.299	176,693	110,388	0.625
		(1.00)	(1.00)		(0.856)	(0.677)		(0.0785	(0.118)		(0.0653)	(0.205)	
Zebrafish	Synonymous	1,119,773	708,664	0.633	957,172	606,291	0.633	58,697	36,209	0.617	103,904	66,164	0.637
		(1.00)	(1.00)		(0.855)	(0.856)		(0.0524)	(0.0511)		(0.0928)	(0.0934)	
	Nonsynonymous	2,700,835	672,851	0.249	2,312,432	471,966	0.204	140,763	40,990	0.291	247,639	159,894	0.646
		(1.00)	(1.00)		(0.856)	(0.701)		(0.0521)	(0.0609)		(0.0917)	(0.238)	

[a] Alignment of codon sequences was constructed by reverse-translating aligned amino acid sequences.
[b] Alignment of codon sequences was constructed by extracting the codon sites that were aligned consistently using the amino acid and nucleotide sequences for all of 10 vertebrate species.
[c] Codon sites that were aligned inconsistently using the amino acid and nucleotide sequences for any of 10 vertebrate species but consistently for the pairwise comparison of human and other vertebrate species.
[d] Codon sites that were aligned inconsistently using the amino acid and nucleotide sequences for the pairwise comparison of human and other vertebrate species.
[e] Proportions of sites and differences in the alignment of codon sequences constructed by reverse-translating aligned amino acid sequences.
[f] Standard errors were mostly more than two orders of magnitude smaller than the estimates.

aligned consistently using the amino acid and nucleotide sequences for all of 10 vertebrate species (class-1 codon sites) may be more likely to be composed of homologous codons than those that are aligned inconsistently for any of 10 vertebrate species (class-2 codon sites).

Therefore, another alignment of codon sequences was constructed by extracting the class-1 sites, and the d_S and d_N values as well as the d_N/d_S ratio were estimated between human and other vertebrate species (Table 1). Compared to the case for reverse-translating aligned amino acid sequences, the d_N/d_S ratio for the class-1 sites dropped to 0.154–0.171 between human and other primates and 0.0997–0.171 between human and non-primate mammals and non-mammalian vertebrates. These results indicate that the d_N/d_S ratio was large for the class-2 sites. However, it should be noted that the codon sites under weak functional constraint or positive selection, where the d_N/d_S ratio is intrinsically high, are more difficult to be aligned compared to those under strong functional constraint, where the d_N/d_S ratio is low (Liu et al., 2009; Fletcher and Yang, 2010). Therefore, the large d_N/d_S ratio for the class-2 sites may be due to mis-alignment of homologous codons or intrinsically high d_N/d_S ratio.

Overestimation of d_N/d_S ratio by reverse-translation of aligned amino acid sequences To distinguish the above possibilities, the class-2 codon sites were further classified into those that were aligned consistently (class-2-1 codon sites) and inconsistently (class-2-2 codon sites) using the amino acid and nucleotide sequences for the pairwise comparison of human and other vertebrate species. The d_S and d_N values as well as the d_N/d_S ratio were estimated for these classes of sites separately (Table 1). For the class-2-1 sites, the d_S value was similar to that obtained for the class-1 sites, which were considered to be composed of homologous codons, suggesting that the codons in the class-2-1 sites were largely homologous. However, the d_N/d_S ratio for the class-2-1 sites was greater than that for the class-1 sites, indicating that the former and latter sites were relatively variable and conservative at the amino acid sequence level, respectively. In contrast, for the class-2-2 sites, the d_S value was much greater than that for the class-1 and class-2-1 sites, suggesting that non-homologous codons were included in this class of sites. The d_N/d_S ratio for the class-2-2 sites was also unduly high.

It should be noted, however, that the proportion of class-2-2 sites in the entire alignment of codon sequences constructed by reverse-translating aligned amino acid sequences was only 1–9% (Table 2). The proportion appeared to be positively correlated with the sequence divergence between vertebrate species, reflecting the fact that aligning sequences is more difficult when sequences are more variable (Liu et al., 2009; Fletcher and Yang, 2010). To examine the effect of class-2-2 sites on the estimation of d_N/d_S ratio, the actual ratio for the entire alignment was estimated by correcting the ratio for the class-2-2 sites, under the assumption that the ratio for this class of sites was similar to that for the class-2-1 sites. This assumption is based on the fact that class-2-1 sites in some pairs of vertebrate species may be classified as class-2-2 sites in other pairs, and *vice versa*. It was observed that the d_N/d_S ratio obtained without correction was 5–43% greater than that obtained with correction (Table 1). The uncorrected ratio was still 0.3–39% greater than the corrected ratio even when the d_N/d_S ratio for the class-2-2 sites was assumed to be twice as great as that for the class-2-1 sites (data not shown). The degree of overestimation for the d_N/d_S ratio appeared to be positively correlated with the ratios of the numbers of synonymous and nonsynonymous differences for the class-2-2 sites to those for other classes of sites (Table 2).

These results suggest that even if the proportion of mis-aligned codon sites is small, they cause significant overestimation of the d_N/d_S ratio for the entire alignment of codon sequences constructed by reverse-translating aligned amino acid sequences (Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009). These codon sites may also be falsely identified as positively selected (Vamathevan et al., 2008; Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009; Fletcher and Yang, 2010). Therefore, caution should be exerted in the study of natural selection using the d_N/d_S ratio by reverse-translating aligned amino acid sequences. It may be necessary to add information from nucleotide sequences to that from amino acid sequences for constructing reliable alignments of codon sequences (Fletcher and Yang, 2010). In addition, since the alignment of codon sequences is not an observation but an inference, it may also be useful to take into account alignment errors for obtaining reliable estimates of the d_N/d_S ratio (Wong et al., 2008).

The author thanks Masafumi Nozawa for technical comments on the retrieval and processing of protein-coding nucleotide sequences for preparing orthologues from 10 vertebrate species analyzed in the present study. The author is also indebted to Yuki Kobayashi and two anonymous reviewers for valuable comments. The present study was supported by KAKENHI 20570008.

REFERENCES

Bakewell, M. A., Shi, P., and Zhang, J. (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc. Natl. Acad. Sci. USA **104**, 7489–7494.

Becquet, C., and Przeworski, M. (2007) A new approach to estimate parameters of speciation models with application to apes. Genome Res. **17**, 1505–1519.

Bonhomme, M., Cuartero, S., Blancher, A., and Crouau-Roy, B. (2009) Assessing natural introgression in 2 biomedical model species, the rhesus macaque (*Macaca mulatta*) and

the long-tailed macaque (*Macaca fascicularis*). J. Hered. **100**, 158–169.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics **21**, 3439–3440.

Fletcher, W., and Yang, Z. (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol. **27**, 2257–2267.

Geraldes, A., Basset, P., Gibson, B., Smith, K. L., and Harr, B. (2008) Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol. Ecol. **17**, 5349–5363.

Goodman, M., and Sterner, K. N. (2010) Phylogenomic evidence of adaptive evolution in the ancestry of humans. Proc. Natl. Acad. Sci. USA **107**, 8918–8923.

Goodman, M., Sterner, K. N., Islam, M., Uddin, M., Sherwood, C. C., Hof, P. R., Hou, Z.-C., Lipovich, L., Jia, H., Grossman, L. I., and Wildman, D. E. (2009) Phylogenomic analyses reveal convergent patterns of adaptive evolution in elephant and human ancestries. Proc. Natl. Acad. Sci. USA **106**, 20824–20829.

Hughes, A. L., and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**, 167–170.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**, 659–716.

Jiang, C., and Zhao, Z. (2006) Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. Genomics **88**, 527–534.

Kimura, M. (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**, 275–276.

Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, New York, Melbourne.

Kondo, R., Horai, S., Satta, Y., and Takahata, N. (1993) Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. J. Mol. Evol. **36**, 517–531.

Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., and Warnow, T. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science **324**, 1561–1564.

MacEachern, S., Hayes, B., McEwan, J., and Goddard, M. (2009) An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics **10**, 181.

Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009) The difficulty of avoiding false positives in genome scans for natural selection. Genome Res. **19**, 922–933.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. **16**, 1182–1190.

Miyata, T., and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16**, 23–36.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature **420**, 520–562.

Nei, M., and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**, 418–426.

Nei, M., and Kumar, S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, Oxford, New York.

Nei, M., Suzuki, Y., and Nozawa, M. (2010) The neutral theory of molecular evolution in the genomic era. Annu. Rev. Genomics Hum. Genet. **11**, 265–289.

Nielsen, R., and Yang, Z. (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol. Biol. Evol. **20**, 1231–1239.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. (1980) The evolution of genes: the chicken preproinsulin gene. Cell **20**, 555–566.

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**, 493–521.

Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science **316**, 222–234.

Rosenberg, M. S., Subramanian, S., and Kumar, S. (2003) Patterns of transitional mutation biases within and among mammalian genomes. Mol. Biol. Evol. **20**, 988–993.

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., and Graur, D. (2009) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol. Evol. **1**, 114–118.

Suyama, M., Torrents, D., and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34**, W609–W612.

Suzuki, Y. (2006) Natural selection on the influenza virus genome. Mol. Biol. Evol. **23**, 1902–1911.

Suzuki, Y., and Gojobori, T. (2001) Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. Gene **276**, 83–87.

Suzuki, Y., and Gojobori, T. (2003) Analysis of coding sequences. In: The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny (eds.: M. Salemi, and A.-M. Vandamme), pp. 283–311. Cambridge University Press, Cambridge.

Suzuki, Y., Gojobori, T., and Kumar, S. (2009) Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection operating at the amino acid sequence level. Mol. Biol. Evol. **26**, 2275–2284.

Takahata, N. (1993) Allelic genealogy and human evolution. Mol. Biol. Evol. **10**, 2–22.

The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**, 69–87.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple-sequence alignment through sequence weighting, position-specific gap penalties, and weight-matrix choice. Nucleic Acids Res. **22**, 4673–4680.

Uddin, M., Goodman, M., Erez, O., Romero, R., Liu, G., Islam, M., Opazo, J. C., Sherwood, C. C., Grossman, L. I., and Wildman, D. E. (2008) Distinct genomic signatures of adaptation in pre- and postnatal environments during human evolution. Proc. Natl. Acad. Sci. USA **105**, 3215–3220.

Vamathevan, J. V., Hasan, S., Emes, R. D., Amrine-Madsen, H., Rajagopalan, D., Topp, S. D., Kumar, V., Word, M.,

Simmons, M. D., Foord, S. M., et al. (2008) The role of positive selection in determining the molecular cause of species differences in disease. BMC Evol. Biol. **8**, 273.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. **19**, 327–335.

Won, Y.-J., and Hey, J. (2005) Divergence population genetics of chimpanzees. Mol. Biol. Evol. **22**, 297–307.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008) Alignment uncertainty and genomic analysis. Science **319**, 473–476.

Zhang, J., Rosenberg, H. F., and Nei, M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**, 3708–3713.

Zhang, W., Bouffard, G. G., Wallace, S., Bond, J. P., and NISC Comparative Sequencing Program (2007) Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. J. Mol. Evol. **65**, 207–214.