Evolution of the Rh Blood Group Genes and Their Related Genes

Takashi Kitano

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Genetics
School of Life Science
The Graduate University for Advanced Studies

1998

ABSTRACT

Majority of the genes are evolving under the neutral mutation pressure. However, some genes are evolving through positive selection. Blood types were originally distinguished by the different molecular structure on erythrocytes. Therefore these products of blood group genes may cause interactions with other organisms, and there is possibility of positive selection on those genes. Because the Rh blood group gene products are membrane proteins, these products of blood group genes seemed to be affected by interactions with other organisms or cells on surface regions. It is known that the Rh blood group genes have homologous genes named Rh50, and hominoids have two or three Rh blood group genes. Therefore the Rh blood group genes and their related genes experienced a series of gene duplication events. Analyses of gene duplication events are also important to elucidate evolutionary rates and patterns of these genes. I thus analyzed the Rh blood group genes and their related genes from primates to fish to clarify the tempo and mode of evolution of these genes.

The human Rh blood type is one of the major blood group systems, and plays important roles in transfusion and clinical medicine, including haemolytic diseases of newborns, autoimmune diseases, and mild haemolytic anemia. Landsteiner and Wiener detected an antibody that agglutinates blood cells from rhesus macaques, and it was named Rh. Nucleotide sequences of Rh blood group genes in some primates were reported, and the phylogenetic relationship of primate Rh blood group genes have been conducted. However, the phylogenetic relationship of primate Rh blood group genes from these studies is not compatible with each other. Because hominoids have two or three loci of Rh blood group genes by gene duplication, gene conversion events (or some kind of convergent effects) may prevent to determine the true gene tree.

I examined the evolution of the Rh blood group genes of primates. Because we don't know the actual gene tree topology of primate Rh blood group genes, I assumed two plausible trees from nucleotide sequence data by using phylogenetic networks. I used the site by site reconstruction method under the maximum likelihood estimates to identify regions of gene conversion events assuming the two trees, and detected 9 or 11 converted regions. After eliminating the effect of gene conversions, I estimated numbers of nonsynonymous and synonymous substitutions for each branch of the both trees. Whichever we selected gene trees, the branch connecting hominoids and Old World monkeys showed significantly higher nonsynonymous than synonymous substitutions, that is, indication of positive selection by using a statistical test. Many other branches also showed higher nonsynonymous than synonymous substitutions, and this suggests that the Rh genes have experienced some kind of positive selection. In any case, we should be very careful when we analyse the evolutionary history of tandemly duplicated genes, for there is always possibility of gene conversions.

To examine evolutionary patterns of other mammalian Rh blood group genes, I determined complete coding regions of Rh blood group genes of five mouse subspecies and rat, and Rh50 genes of five mouse subspecies, rat, and crab-eating macaque, and examined these genes. Nucleotide and amino acid sequence similarities between Rh genes and Rh50 genes are 47.2-48.9 % and 34.4-37.8 %, respectively. Comparison of synonymous and nonsynonymous substitutions for the Rh50 gene also revealed a possibility of existence of positive selection for this gene in primates. Because primates showed more clear sign of positive selection than rodents both for Rh and Rh50 genes, it is possible that the pattern of host-parasite interaction is different between primates and rodents. Phylogenetic analyses of Rh and Rh50 amino acid sequences indicate that the Rh50 gene has been evolving about two times more slowly than the Rh blood group gene both in primates and rodents. This conservative nature of the Rh50 gene suggests its relative importance to the Rh blood group gene. From the comparison of synonymous

substitutions between Rh and Rh50 genes, it is suggested that the mutation rate of rodents is about three times higher than that of primates, and the divergence time between mouse and rat is estimated to be ca. 30 million years ago.

I also determined the Rh50-like genes of *Xenopus* and Japanese medaka and examined the long-term evolution of Rh, Rh50, and their related genes. The phylogenetic tree shows four clusters in this tree; Rh50 genes of mammals and the *Xenopus* Rh50-like gene, Rh genes of mammals, the Rh50-like gene of Japanese medaka, and two genes of *C. elegans*. Therefore, the *Xenopus* Rh50-like gene is probably orthologous to the Rh50 genes of mammals.

The topology of the phylogenetic tree suggests that the gene duplication of Rh and Rh50 genes occurred just before or after the divergence of teleost fish and other vertebrates. The branch lengths of Rh50 genes are much shorter than those of Rh genes, indicating a lower evolutionary rate in the Rh50 gene than in the Rh gene. Because its evolutionary rate is lower than that for the Rh protein gene, the Rh50 protein may be closer to the ancestral form before the gene duplication of Rh and Rh50 genes. The time of gene duplication that produced the Rh and Rh50 genes was estimated to be about 450-480 million years ago. This period roughly corresponds to the early Paleozoic, around the divergence between tetrapods and teleost fish lineages.

From database searches, it is suggested that the Rh blood group genes and their related genes are related to ammonium transporter genes of many organisms, especially trans-membrane domains. The phylogenetic tree for ammonium transporter proteins indicated two major groups for ammonium transporter proteins. I propose to call these two groups of ammonium transporter genes as α and β groups, and the Rh genes group is more similar to the amt β group than to the amt α group. It is suggested that the Rh blood group genes and their related genes have probably been existing as essential membrane proteins in many animal phyla.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Majority of the genes are evolving under the neutral mutation pressure (Kimura, 1983). However, some genes are evolving through positive selection. Blood types were originally distinguished by the different molecular structure on erythrocytes. Therefore these products of blood group genes may cause interactions with other organisms, and there is possibility of positive selection on those genes. The Rh blood group gene products are thought to be membrane proteins, but these actual function is not known. Analyses of the Rh blood group genes and their related genes are needed not only for seeking possibility of positive selection but also for elucidating the actual function of these genes. It is known that the Rh blood group genes have homologous genes named Rh50, and hominoids have two or three Rh blood group genes. Therefore the Rh blood group genes and their related genes experienced a series of gene duplication events. Analyses of gene duplication events are also important to elucidate evolutionary rates and patterns of these genes. I thus sequenced some homologous genes of the Rh blood group genes and analyzed those from primates to fish to clarify the tempo and mode of evolution of the Rh blood group genes and their related genes.

The history of the study of the Rh blood group genes

The human Rh blood type is one of the major blood group systems, and plays important roles in transfusion and clinical medicine, including haemolytic diseases of newborns, autoimmune diseases, and mild haemolytic anemia. Landsteiner and Wiener

(1940) detected an antibody that agglutinates blood cells from rhesus macaques, and it was named Rh. This antibody had similar features with the antibody discovered by Levine and Stetson (1939) from transfusion incompatibilities. Although the two antibodies had a similar specificity, these antibodies were shown to detect distinct antigens. Therefore the antibody detected by Landsteiner and Wiener was renamed LW. There were historically two hypotheses about the Rh blood group system; Wiener's (1943) one locus theory and Fisher-Race's (1944) three linked loci (C, D, E) theory. Therefore two different nomenclature for loci and alleles have been used.

Rh polypeptides were observed as phosphorylated 30-32 kD membrane proteins by using SDS-PAGE and immunoprecipitation (Moore, Woodrow, and McClelland 1982; Gahmberg 1982). Nucleotide sequences of Rh genes were determined independently by Cherif-Zahar et al. (1990) and Avent et al. (1990). The Rh blood group system was shown to be composed of two closely linked D and CE loci (Mouro et al. 1993) as predicted by Tippett (1986). In human, D and CE loci are constructed from ten exons, and are located on chromosome 1p34-p36 (Ruddle et al. 1972; Cherif-Zahar et al. 1991) (figure 1.1A). The physical length between D and CE loci is not known. Individuals are divided into Rh-positive and Rh-negative according to the presence or absence of the D antigen. C/c and E/e specificities are distinguished by four and one amino acid differences, respectively (Mouro et al. 1993). Rh gene products were estimated to have 12 trans-membrane domains (figure 1.2A) through hydropathy analysis (Avent et al. 1990) and immunological studies using an antipeptide antibody (Avent et al. 1992). It is suggested that the expression of Rh genes are restricted to tissues or cells exhibiting erythroid features from Northern blot analysis (Cherif-Zahar et al. 1990). However, Kajii et al. (1994) indicated that Rh genes are expressed not only in erythroid lineage but also in various leukocytes, though expressions in leukocytes are quite low from RT-PCR method.

Nucleotide sequences of Rh-like blood group genes in nonhuman primates were

also reported (Salvignol et al. 1994, 1995; Mouro et al. 1994a). Genomic DNA analysis by Southern blot using the human Rh genes as probes have shown that chimpanzee possesses three Rh-like loci (Salvignol et. al. 1993, 1994), though only two types of genes for chimpanzee were so far sequenced (Salvignol et al. 1995). Gorilla carries two Rh-like genes, while orangutans, gibbons, Old World monkeys, and New World monkeys carry a single Rh-like gene (Blancher, Calvas, and Ruffie 1992).

A protein was obtained together with the Rh gene product on immunoprecipitation with anti-Rh antibodies from human, and named as 50kD glycoprotein (Moore and Green 1987). This glycoprotein was considered to form heterotetramer with Rh blood group gene products and some other proteins (glycophorin B, LW antigen, Fy antigen, CD47, and ABH antigen) were added this heterotetramer on erythrocyte membranes (Eyers et al. 1994) (figure 1.2B). The nucleotide sequence of the human 50kD glycoprotein was determined, and its amino acid sequence was homologous with that of the human Rh gene (Ridgwell et al. 1992). Organization of the gene is similar to that of Rh genes and the locus is located on chromosome 6p21-qter (figure 1.1B). That protein was also predicted to have the 12 trans-membrane domains which are similar to those of the Rh blood group gene product. There are several names for this gene such as RHAG, but I call this gene as Rh50 and the Rh blood group gene as Rh hereafter for simplicity. It has been shown that the Rh_{null} regulator and the Rh_{mod} phenotypes are suppressed by the Rh50 product (Cherif-Zahar et al. 1996), and a splicing mutant of this gene was shown to cause an Rh_{null} phenotype (Kawano et al. 1998). These observations clearly indicate that the Rh50 gene is essential for expression of Rh antigens on erythrocytes. These Rh gene and Rh related gene products seem to play an important role for erythrocytes.

Positive selection

Because positive selection prefers nonsynonymous substitutions that cause adaptive amino acid changes rather than synonymous substitution that do not cause any amino acid changes on nucleotide sequence level, positively selected genes are considered to have nucleotide changes that the number of nonsynonymous substitutions is larger than that of synonymous substitutions.

Several genes are considered to be positively selected, such as antigen recognition sites of the major histocompatibility complex class I loci (Hughes and Nei 1988, 1989), alcohol dehydrogenese genes (Long and Langley 1993), α1-antitrypsin genes (Ohta 1994), hemagglutinin 1 gene of human influenza A virus (Ina and Gojobori 1994), abalone sperm lysin genes (Lee, Ota, and Vacquier 1995), primate lysozyme genes (Messier and Stewart 1997), and primate ribonuclease genes (Zhang, Rosenberg, and Nei 1998).

There are two types of mechanism of positive selection, that is, gain of function and interaction between host defense systems and pathogens. Examples of the former are primate lysozyme genes (Messier and Stewart 1997) and primate ribonuclease genes (Zhang, Rosenberg, and Nei 1998). In the case of these genes, particular branches of a gene tree usually experienced positive selection. As examples of the latter, there are studies of the major histocompatibility complex class I loci (Hughes and Nei 1988, 1989) and hemagglutinin 1 gene of human influenza A virus (Ina and Gojobori 1994). In those cases, because positive selection might occur by interaction between host defense systems and pathogens, positive selection may always operate on those genes, and those genes are thought to code cell surface proteins. Endo, Ikeo, and Gojobori (1996) searched the nucleotide sequence database and found that 17 gene groups were the candidates for the genes on which positive selection may operate. Nine of those 17 gene groups were surface antigens of parasites or viruses. Therefore, other blood group genes may also have possibility of experiencing positive selection.

Because the Rh blood group gene products are membrane proteins, these products

of blood group genes seemed to be affected by interactions with other organisms or cells on surface regions. Therefore there is possibility of positive selection on those genes as shown for the ABO blood group genes by Saitou and Yamamoto (1997).

Gene conversion

Gene conversion means the transfer of a gene segment from a donor gene to a homologous acceptor gene without the donor gene being changed in the process. Gene conversion was originally found from irregular segregations: i. e., departures from the 2+:2- ratio expected in tetrads from the Mendelian segregations of genes in heterozygous (+/-) condition in fungi (Holliday 1964; Roman and Ruzinski 1990). Since then, it has also been suggested that an analogous mechanism is responsible for transfer of gene segments in higher eukaryotes as well (Baltimore 1981; Slightom, Blechl, and Smithies 1980). In higher organisms, the term "gene conversion" has been used in the context of "templated segmental mutation". Gene conversion has been proposed to cause both wide homogenization (α-globin, Liebhaber, Goossens, and Kan 1981; murine serum amyloid A gene, Lowell et al. 1986) and polymorphism (MHC class I, Kuhner et al. 1990; MHC class I, Pease 1985; immunoglobulin, Thompson 1992) between homologous genes.

Hogstrand and Bohme (1998) examined gene conversion events with a PCR assay at the DNA level between the two MHC class II genes in mice sperm. They obtained values 1.2×10^{-6}-9.7×10^{-5} per locus per generation for those gene conversion frequency.

In primate Rh blood group genes, D and CE loci are tightly linked and they are quite similar (96.4 % nucleotide sequence identity). Therefore gene conversions may affect the phylogenetic relationships of those genes. Occurrences of gene conversions and unequal crossing-overs in the Rh loci have been detected (e. g., Cherif-Zahar et al.

1994; Mouro et al. 1994b; Beckers et al. 1996; Huang et al. 1996; Kemp, Poulter, and Carritt 1996; Carritt, Kemp, and Poulter 1997), but there has been no detailed analysis on those events.

Phylogenetic network

The evolutionary history of a gene should be presented as a tree. When real sequence data are analyzed, however, this tree structure may not be clearly observed. When two nucleotide positions show incongruent partition pattern, a discordancy diagram appears (Fitch 1977). Bandelt (1994) extended this idea and proposed the "phylogenetic network" method. A network structure is useful for delineating anomaly in the history of gene trees. For example, when two regions of a gene experienced recombination (and/or gene conversion), we may obtain a network, not a tree, if we analyze the sequence data by combining the two regions. We may also observe parallel substitutions among genes by using a phylogenetic network.

Figure 1.3 shows an explanation of network analyses. In the case of this sequence data (figure 1.3A), sites 1, 2, and 3 divide [a, b] and [c, d], while sites 4 and 5 divide [a, d] and [b, c]. Sites 7 and 8 are singular substitution sites. In this network (figure 1.3B), two topologies are contained. One topology (I and II) shows clusters ab and cd (this is the maximum parsimony tree), another one (II and IV) shows clusters ad and bc (this is only one additional change required compared from the maximum parsimony tree) (figure 1.3C). The network contains all possible topologies and we can identify sites experiencing parallel substitutions. If parallel substitution sites are contiguous, a gene conversion event over the region containing those sites is inferred. The phylogenetic network analysis is useful for closely related genes that experience evolutionary events resulting in non-tree structure.

Evolutionary studies for the Rh blood group genes

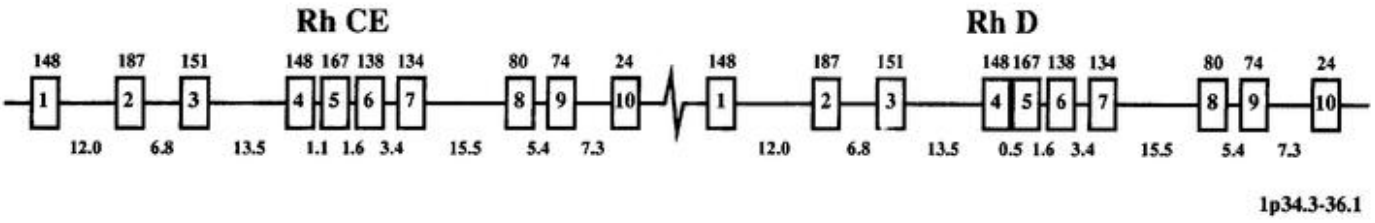
As I mentioned above, some nucleotide sequences of primate Rh blood group genes are reported. The phylogenetic relationship of primate Rh blood group genes have been conducted. Salvignol et al. (1995) constructed a neighbor-joining tree of primate Rh blood group genes. Blancher and Socha (1997) constructed a maximum likelihood tree of primate Rh blood group genes. Klein, O'hUigin, and Blancher (1997) constructed a UPGMA tree of primate Rh blood group genes. The phylogenetic relationship of primate Rh blood group genes from these studies is not compatible each other. Because hominoids have two or three loci of Rh blood group genes by gene duplication, gene conversion events (or some kind of convergent effects) may prevent to determine the true gene tree.

Questions to be addressed

In Chapter II, the first objective is a detailed analysis of the gene conversion events in hominid Rh blood group genes so as to infer their true phylogenetic relationship. The second objective is the examination of existence of positive selection on the Rh blood group genes to understand positive selection more clearly. In Chapter III, to compare evolutionary patterns of other mammalian Rh blood group genes, those genes for rodents are determined and phylogenetic analyses are carried out. In Chapter IV, to examine the long-term evolution of the Rh blood group genes and their related genes, those genes for Japanese medaka fish (*Oryzias latipes*) and African clawed frog (*Xenopus laevis*) are determined and phylogenetic analyses are carried out.

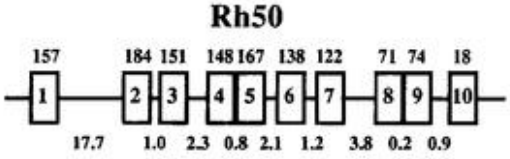
Figure 1.1 The two Rh loci located on 1p34.3-36.1 (A) and the Rh50 locus located on 6p21-qter (B) are schematically represented. Boxes with number mean exons. Numbers of above and below mean number of base pair for each exon and intron length (kb), respectively. These are based on Matassi et al. (1998).

(A)



1p34.3-36.1

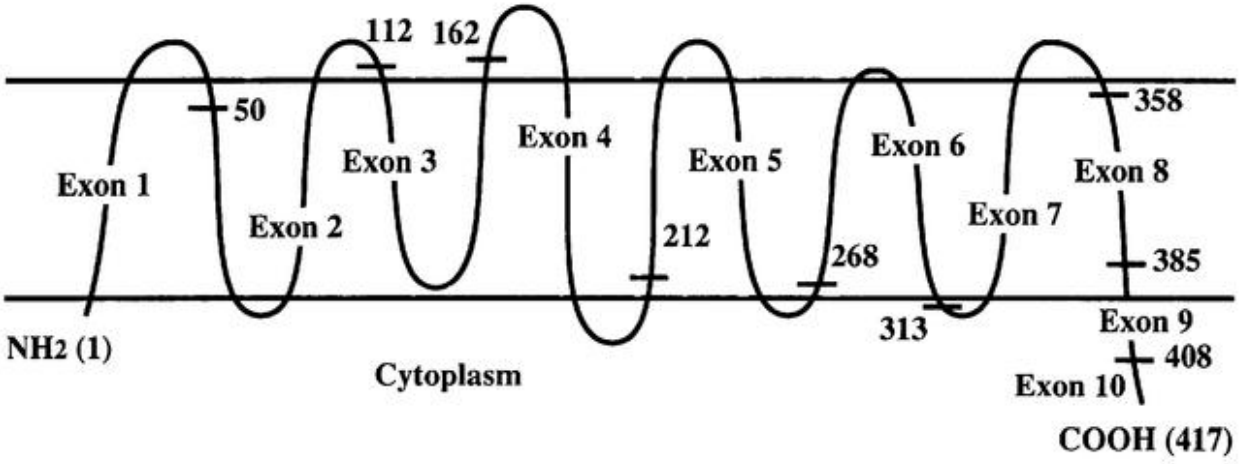
(B)



6p21-qter

Figure 1.2 (A) Topology of Rh protein deduced from hydropathy analysis. Bars with number mean boundaries of exons. (B) The Rh protein complex on the red cell membrane. These are based on Blancher and Socha (1997).

(A)


112
162
50
358
Exon 1
Exon 3
Exon 4
Exon 5
Exon 6
Exon 8
Exon 2
Exon 7
212
268
385
313
Exon 9
NH2 (1)
Cytoplasm
408
Exon 10
COOH (417)

(B)

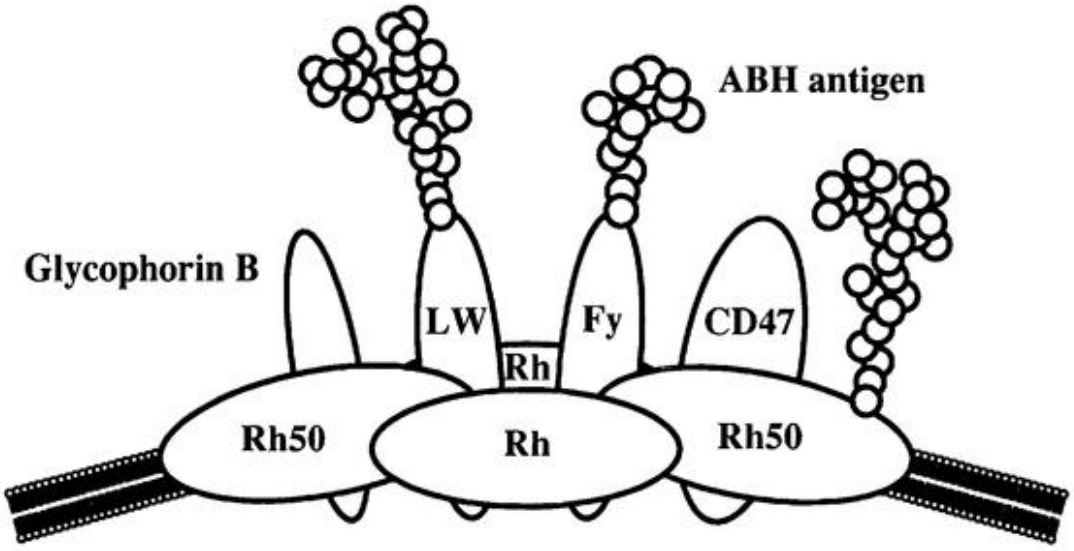
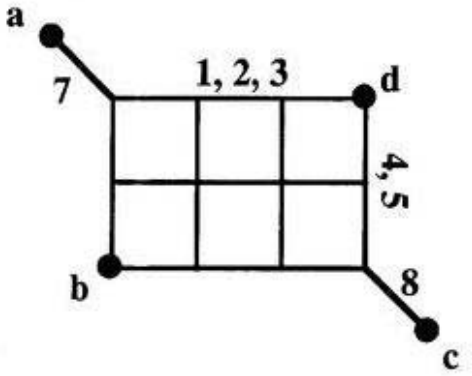

ABH antigen
Glycophorin B
LW
Fy
CD47
Rh
Rh50
Rh
Rh50

Figure 1.3 Comparisons of phylogenetic network and parsimonious trees. (A) Sequence data used for explanation of network analyses. (B) The phylogenetic network constructed from the data (A). Numbers are nucleotide positions responsible for corresponding edges and edge lengths are proportional to number of nucleotide differences. (C) Four possible trees embedded in the network.

(A)

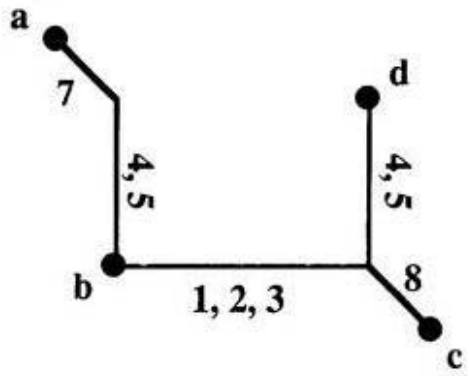
	12345678
Sequence a	TGCTAGTG
Sequence b	TGCCGGAG
Sequence c	CATCGGAA
Sequence d	CATTAGAG

(B)

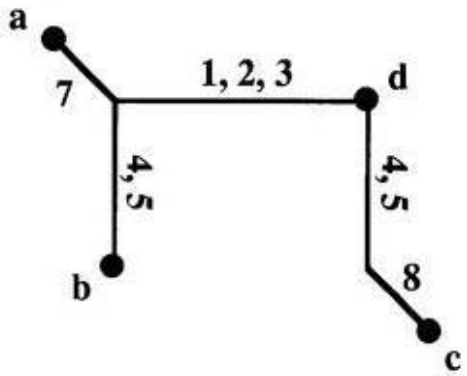


(C)

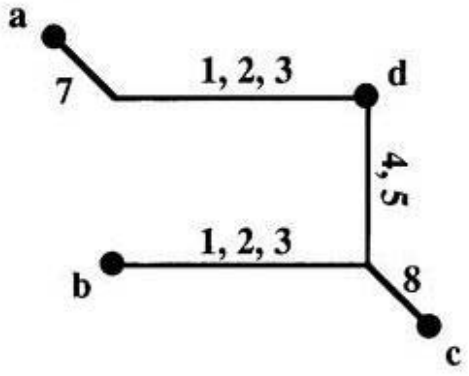
(I)



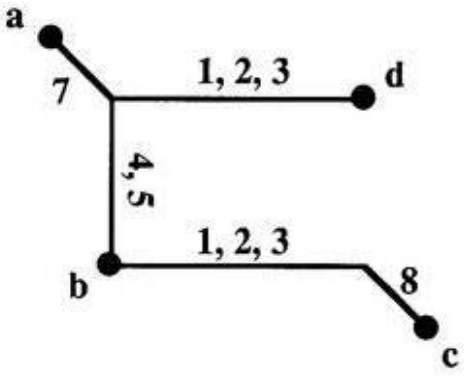
(II)



(III)



(IV)



CHAPTER II

EVOLUTION OF PRIMATE RH BLOOD GROUP GENES

cDNA sequences used

Nucleotide sequence data for Rh blood group genes were retrieved from the DDBJ/EMBL/GenBank international nucleotide sequence database and from published papers. Ten human (*Homo sapiens*) sequences, five chimpanzee (*Pan troglodytes*) sequences, three gorilla (*Gorilla gorilla*) sequences, two crab-eating macaque (*Macaca fascicularis*) sequences, and one rhesus macaque (*Macaca mulatta*) sequence were used in this study (table 2.1). All the sequences were complete cDNA with 1251 bp.

Gene conversions confuse the Rh gene tree

Because the two Rh loci are tightly linked, it is possible that they have experienced gene conversions or crossing-overs. Those events can confuse the phylogenetic relationship of the linked loci. I thus constructed phylogenetic networks for human, chimpanzee, and gorilla Rh blood group genes. Phylogenetic networks were constructed following the procedure of Bandelt (1994) and Saitou and Yamamoto (1997). Because the network for the whole sequence data have many dimensions, I constructed networks for five regions of the coding region of this gene. For simplicity, I selected one allele of each locus by examining phylogenetic networks of human (figure 2.1A) and chimpanzee (figure 2.1B) Rh blood group genes were selected. From these figures, I thus used human D-0 (a consensus sequence of human D alleles), human cE-1,

chimpanzee 1-3, and chimpanzee 2 genes. These alleles were used as representative of each locus in the following analyses. Nucleotide identities between human D and human CE, between chimpanzee 1 and chimpanzee 2, and between gorilla 1 and gorilla 2 are 96.4 %, 95.9 %, and 97.0 %, respectively.

Figure 2.2 shows phylogenetic networks for exons 1-3 (A), exons 4-5 (B), exon 6 (C), exon 7 (D), and exons 8-10 (E). All the networks contain parallelograms that suggest parallel substitutions or some kind of convergent changes. If the relationship of genes is not affected by gene conversion and/or crossing-over, the network may not contain so many parallelograms. Moreover, some sites of those parallelograms in figure 2.2 are contiguous (e.g. 391, 397, and 399 of figure 2.2A), suggesting the existence of conversion-like events.

Assumption of primate Rh blood group gene tree

Because chimpanzee possesses three Rh-like loci, at least two gene duplications occurred in the hominoid lineage. Three types of gene duplication patterns can be assumed. Figures 2.3A-C show these three possibilities. Tree 2.3A assumes that one gene duplication occurred in the common ancestor of human, chimpanzee, and gorilla, and second gene duplication occurred on one duplicated gene of the chimpanzee lineage. Tree 2.3B assumes that one gene duplication occurred in the common ancestor of human, chimpanzee, and gorilla, and second gene duplication occurred in one duplicated gene of the common ancestor of human and chimpanzee. Tree 2.3C assumes that two gene duplications occurred in the common ancestor of human, chimpanzee, and gorilla. Established phylogeny (human and chimpanzee are clustered first) for the three species (e.g., Horai et al. 1995) is adopted for each orthologous gene group.

To identify orthologous genes, I first classified sites based on phylogenetic

networks (table 2.2). For example, sites 380 and 383 of network A (figure 2.2) divide the human D-chimpanzee 1 pair from the remaining genes. Because two genes from the same species can no longer form a closest cluster in either trees, sites indicating those clusters are not shown.

If we assume tree 2A, two genes of chimpanzee (α-1 and α-2) are probably quite similar to each other, and it is not easy to identify them. Because the nucleotide identity between chimpanzee 1 and chimpanzee 2 sequences is similar to that between human D and human CE, and that between gorilla 1 and gorilla 2, I assume that these two genes probably correspond to chimpanzee α-1 (or α-2) and β, not chimpanzee α-1 and α-2. In this case orthologous trios can be extracted from phylogenetic networks. Because sites 514, 544, and 733 show clusters of human D-chimpanzee 1-gorilla 2 and human CE-chimpanzee 2-gorilla 1, and other 13 sites (380, 383, 916, 932, 985, 986, 541, 579, 581, 584, 1048, 1170, and 1025) are compatible with these clusters, I therefore determined the topology of hominoid Rh blood group genes as tree D in figure 2.2.

When we consider either trees B or C of figure 2.3, we have to assume two and three gene losses (or genes not yet identified), respectively. Because we don't know which genes are lost (or genes not yet identified), examination under trees 2.3B or 2.3C is more difficult than that under tree 2A. Because the cluster for human D and chimpanzee 1 is supported with 6 sites (380, 383, 916, 932, 985, and 986), and the cluster for chimpanzee 2 and gorilla 1 is supported with 4 sites (541, 579, 581, and 584), these clusters are plausible. Moreover 3 sites (514, 544, and 733) show clusters of human D-chimpanzee 1-gorilla 2 and human CE-chimpanzee 2-gorilla 1. Therefore I can determine the unrooted topology as follows:

(((human D,chimpanzee 1),gorilla 2),(human cE,(chimpanzee 2,gorilla 1))).

To determine the root I eliminated sites indicating two genes from the same species form a cluster from the multiply aligned sequence data and constructed a neighbor-joining tree (Saitou and Nei 1987, CLUSTAL W of Thompson, Gibson, and Higgins (1994) was

used). Three sequences of Old World monkeys were used as outgroups. The topology of the NJ tree was compatible with tree E in figure 2.3. The maximum likelihood analysis (NucML of Adachi and Hasegawa (1994) was used) also supported this topology. It is interesting to note that the trees D and E both showed the same cluster of human D-chimpanzee 1-gorilla 2. Only the position of human CE is different between the two trees. Tree E is more plausible because of 4 sites (541, 579, 581, and 584) of the network B (figure 2.2). Recently, Apoil and Blancher (personal communication) studied the evolutionary relationship of the primate Rh genes using the intron 4 sequences. They suggested that two gene duplications occurred in the common ancestor of human, chimpanzee, and gorilla. This scenario is compatible with my tree E of figure 2. However, tree E requires three gene losses (or genes not yet identified) compared to one gene loss for tree D, which is more parsimonious. I thus used those two assumed trees for the following analyses.

The site by site reconstruction method by using the maximum likelihood method

I used the site by site reconstruction method (Slightom et al. 1987) to identify regions of gene conversion events assuming the two trees (figures 2.3D and 2.3E). The original method is based on the parsimony method, however, we used the maximum likelihood method (Felsenstein 1981) to determine nucleotide sequences of ancestral nodes. Firstly, I obtained the maximum likelihood estimates of the nucleotide sequences of ancestral nodes by using PAML program version 1.3 (Yang 1997) and then substitution patterns were plotted on each branch of the assumed tree. Let us explain the actual procedures of the site by site reconstruction method using figure 2.4, where tree D of figure 2.3 was assumed. The two chimpanzee loci both had the same nucleotide (C)

in case A and the event causing this change is indicated by letter "P". They look like parallel substitutions, but this pattern could also be produced by a gene conversion after a substitution. In case B, two independent substitutions designated by "S" occurred in different species. In case C, a branch of one cluster experienced a substitution (indicated by letter "S") after the gene duplication, and no change in its descendants is indicated with "O". Case D is a next step of case C. Substitutions occurred in the descendants. These additional changes result in the same nucleotide with their duplicated genes, suggesting a gene conversion from the human CE to the human D genes and a gene conversion from chimpanzee 1 to chimpanzee 2 over this site. Therefore, the letter "Q" is given to the human D gene and the letter "R" is given to the chimpanzee 2 gene. If cases A or D are contiguous for two or more variant sites, a gene conversion event over the region containing those sites is inferred based on the maximum parsimony principle. In case D, I can infer the direction of gene conversion. I also performed the same procedure under the assumption of tree E of figure 2.3.

Let us compare rates of gene conversion with nucleotide substitutions to see if our parsimonious argument is valid. Hogstrand and Bohme (1998) estimated cis gene conversion frequency in mouse MHC class II genes and obtained values 1.2×10^{-6}-9.7×10^{-5} per locus (about 200 bp region) per generation. Rates of synonymous substitutions in various mammalian protein-coding genes is estimated to be 3.5×10^{-9} per site per year (Li 1997). If I assume that the generation time of wild mice is 0.5 year (Dr. Tsuyoshi Koide, personal communication), the rate of cis gene conversion frequency becomes 2.4×10^{-6}-1.9×10^{-4} per 200 nucleotide sites per year, while the corresponding rate of synonymous substitution is 7.0×10^{-7}. Therefore the rate of gene conversion events seem to be much higher than that of nucleotide substitutions. This justifies the site by site reconstruction procedure.

Table 2.3 shows substitution patterns of all the variable sites under tree D of figure 2.3 estimated by using the site by site reconstruction method. Ancestral sequences were

estimated by using the maximum likelihood method. According to likelihood values, patterns are arranged from the top to the bottom. I inferred regions of gene conversion events from these results. For example, sites 31-102 are inferred to have experienced a gene conversion because letters "P" and "Q" were contiguous. Human D and chimpanzee 1 sequences are identical on site 31 (indicated by "O") while human CE is indicated by "Q", thus the direction of this gene conversion is inferred from the human D gene to the human CE gene.

Two possible patterns were inferred for sites 579-584. One pattern is that three parallel substitutions occurred in chimpanzee 2 and gorilla 1 genes on sites 579, 581, and 584, and one substitution occurred in the gorilla 1 gene on site 580. Another pattern involves three substitutions in the ancestral CE gene of human, chimpanzee, and gorilla, followed by three backward substitutions in human CE on sites 579, 581, and 584 and one substitution in gorilla 1 gene on site 580. Either patterns require seven substitution events. If we assume a gene conversion event on this region on the latter case, however, only five events are necessary: three substitutions in the ancestral CE gene of human, chimpanzee, and gorilla, one gene conversion in human CE, and one substitution on the gorilla 1 gene on site 580. Therefore I inferred that a gene conversion occurred on this region.

If the total numbers of events (substitution and gene conversion) is decreased by taking into account the gene conversions, I selected that pattern in spite of lower likelihood estimation of ancestral nodes as shown in the above example (table 2.3). All other gene conversion regions were inferred in the same fashion. Table 2.4 shows substitution patterns of each variable sites under tree E of figure 2.3, by using the same procedure as in the case for table 2.3. Multiple alignments of non-converted and ancestral sequences are shown in appendix II.

Table 2.5 shows the inferred gene conversion events occurred in the hominoid phylogeny. Eleven and nine events were estimated for trees D and E of figure 2.3,

respectively. The direction of the gene conversion was not identified in some events. Because I cannot detect break points of gene conversions and analyzed only cDNA sequences, these ranges of gene conversions are minimum lengths.

The phylogenetic tree and synonymous/nonsynonymous substitutions

I reconstructed the two sets of non-converted sequences from the results shown in tables 2.3 and 2.4. In the case of the gene conversion event ID 1 for tree D (see table 2.5), for example, nucleotides of sites 31-102 of the human CE gene were substituted to those of the CE gene of the common ancestor of human and chimpanzee because of the direction of gene conversion (D → CE). As directions of gene conversion events IDs 5 and 9 for trees D and E and ID 6 for tree E were not determined, the sequence data of those sites were not used.

Figure 2.5 shows two phylogenetic trees (A and B) for the reconstructed primate Rh blood group gene sequences assuming tree D (figure 2.5A) and tree E (figure 2.5B), respectively. All the branch lengths were estimated from sequence data of extant nodes and ancestral nodes estimated above, and the mid-point rooting was used. Numbers of synonymous and nonsynonymous substitutions were estimated by Ina's (1995) method. In Ina's method, the proportion (R) of transition/transversion of the third codon is estimated from proportions of base changes that are observed in the entire phylogenetic tree. R values for trees D and tree E were estimated to be 1.79 and 1.87, respectively. Numbers of nonsynonyomous substitutions are higher than those of synonymous substitutions on almost all branches for both trees (tables 2.6 and 2.7). Application of the Fisher's exact test (Zhang, Kumar, and Nei 1997) showed that the nonsynonymous substitution is significantly higher than synonymous ones ($P = 0.0041$) in the branch (a-f) that connects hominoids and Old World monkeys for tree A of figure 2.5, where 79 nonsynonymous and 14 synonymous substitutions were estimated. I also considered an

alternative, less likely substitution patterns indicated in table 2.3 for this branch. Numbers of nonsynonyomous and synonymous substitutions for the branch become 76 and 14, respectively, but the difference is still highly significant (P = 0.0067). The same branch (a-f) that connects hominoids and Old World monkeys is also significant (P = 0.0031) for tree B of figure 2.5, where 80 nonsynonymous and 14 synonymous substitutions were estimated. I also considered an alternative, less likely substitution patterns indicated in table 2.4. Numbers of nonsynonyomous and synonymous substitutions for the branch become 73 and 14, respectively, and the difference is still statistically significant (P = 0.0100). In any case, whichever I selected gene trees, I could find the possibility of positive selection in the branch that connects hominoid and Old World monkey clusters.

I also estimated the total numbers of nonsynonyomous and synonymous substitutions for hominoid branches to examine the overall evolutionary pattern within hominoids. Nonsynonymous substitutions are not significantly higher than synonymous ones in hominoids for tree A (P = 0.1055), where 96 nonsynonymous and 27 synonymous substitutions were estimated, nor for tree B (P = 0.0889), where 97.5 nonsynonymous and 27.5 synonymous substitutions were estimated. Because these values were estimated from non-converted sequences data, however, I also considered the situation that all gene conversions occurred on the hominoid lineage to be parallel substitutions. The nonsynonymous substitutions (132) is significantly higher than synonymous ones (30) (P = 0.0046) within hominoids for tree A. The nonsynonymous substitutions (124.5) is also significantly higher than synonymous ones (30.5) in hominoids (P = 0.0117) for tree B. Because I eliminated effects of gene conversion events to estimate numbers of synonymous and nonsynonymous substitutions, it is noted that I carried out conservative estimation of these values in hominoids under the assumption of gene conversions, and those may be underestimates.

I also estimated the average rates of synonymous and nonsynonymous substitutions

for the Rh blood group genes, under the assumption of constancy of the evolutionary rate and the divergence between the Old World monkey and hominoid lineages to be 23 million years ago [MYA] (Kumar and Hedges 1998). Average numbers of synonymous and nonsynonymous substitutions per site between Rh blood group genes of Old World monkeys and of hominoids were estimated to be 0.068 and 0.124, respectively, for tree A of figure 2.5, applying Ishida et al.'s (1995) method. Average numbers of synonymous and nonsynonymous substitutions per site between Rh blood group genes of Old World monkeys and of hominoids for tree B of figure 2.5 were almost the same (0.067 and 0.125 for synonymous and nonsynonymous substitutions, respectively) as those of tree A of figure 2.5. Therefore, the rates of synonymous and nonsynonymous substitutions (/site/year) for the Rh blood group genes of Old World monkeys and of hominoids are estimated to be $1.46\text{-}1.48\times10^{-9}$ (=0.067-0.068/[2×23 MYA]) and $2.70\text{-}2.72\times10^{-9}$ (=0.124-0.125/[2× 23 MYA]), respectively. The evolutionary rate of synonymous substitution for the Rh gene is somewhat lower than that for other primate genes (2.3×10^{-9}: Li and Tanimura 1987). In any case, it is clear that the nonsynonymous substitution is in average higher than the synonymous one in primate Rh genes.

Table 2.1

Rh blood group genes and references used in this study

Genes (Accession Number)	References
Human D-1 (X63097)	Le Van Kim et al. (1992)
Human D-2 (X63094)	Le Van Kim et al. (1992)
Human D-3 (S57971)	Kajii et al. (1993)
Human D-4 (L08429)	Arce et al. (1993)
Human D-5 (S78509)	Huang et al. (1995)
Human D-6	Huang et al. (1996)
Human cE-1 (M34015)[a]	Cherif-Zahar et al. (1990)
Human cE-1 (X54534)[a]	Avent et al. (1990)
Human cE-2 (S57967)	Kajii et al. (1993)
Human Ce	Huang et al. (1996)
Chimpanzee 1-1 [317-IIR] (L37050)	Salvignol et al. (1995)
Chimpanzee 1-2 [394-2G]	"
Chimpanzee 1-3 [211-6E]	"
Chimpanzee 1-4 [317-IA] (L37049)	"
Chimpanzee 2 [211-IIF] (L37048)	"
Gorilla 1-1 [IC] (L37052)	"
Gorilla 1-2 [IIA2b]	"
Gorilla 2 [ID] (L37053)	"
Crab-eating macaque 1 (L37054)	"
Crab-eating macaque 2	"
Rhesus macaque (S70343)	Mouro et al. (1994a)

[a] Nucleotide sequences of M34015 and X54534 (both Human cE-1) are identical.

Names in square brackets are those used by Salvignol et al. (1995).

Table 2.2

Classification of sites to identify orthologous genes from phylogenetic networks

Pairs[a]	No. of sites	Sites	Network[b]
(Hu D,Ch 1,Go 2) - (Hu CE,Ch 2,Go 1)	3	514, 544, 733	B
(Hu D,Ch 2,Go 1) - (Hu CE,Ch 1,Go 2)	1	577	B
(Hu D,Ch 1) - (others)	6	380, 383	A
		916, 932	C
		985, 986	D
(Ch 2,Go 1) - (others)	4	541, 579, 581, 584	B
(Ch 1,Go 1) - (others)	2	852	C
		1122	E
(Hu D,Go 2) - (others)	1	1048	D
(Hu CE,Ch 1) - (others)	1	505	B
(Hu CE,Ch 2) - (others)	1	1170	E
(Hu CE,Go 1) - (others)	1	1025	D

[a]Hu: human, Ch: chimpanzee, Go: gorilla

[b]See figure 2.2 for networks A-E

Table 2.3

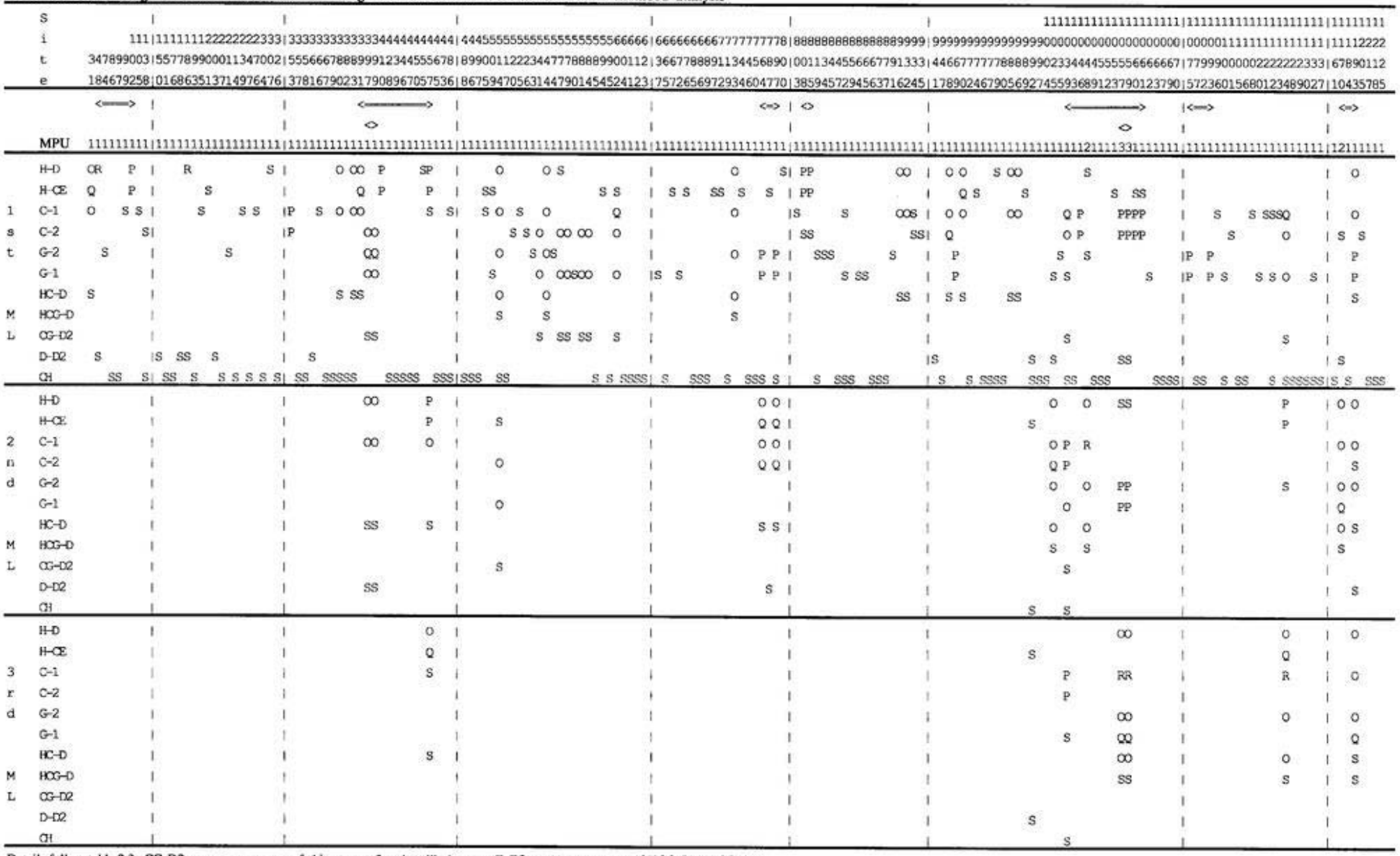
Patterns of site changes for each branches of assuming tree D estimated from the maximum likihood analysis

	S			11111111111111111111	11111111111111111111	11111111
	i	111	11111112222222333	33333333333344444444444	44455555555555555555566666	66666666677777777778	88888888888888889999	99999999999999900000000000000000000000	00000111111111111111	11112222
	t	347899003	55778990001134700 2	55566678889991234455567 8	89900112223447778888990011 2	36677888911344568 90	00112344455666779133 3	44667777778888990023344445555566666667	77999000002222222333	67890112
	e	184679258	01686351371497647 6	37816790231790896705753 6	86759470563144790145452412 3	75726569729346047 70	38591445729456371624 5	17890246790569274755936891237901237 90	57236015680123489027	10435785
		<——>		<——>	<——>	<=>	<——>	<——>	<==>	<=>	<=>
				<——>				<>			
	MPU[a]	11111111	11111111111111111	11111112121331111111111	11111111111111212211111111	11111111111111111	11111111121111111111	11111111111111111111112111133111111	11111111111121111111	12121111
	H-D	OP P	P S	O OOOOP SP	O S	O OS	QQ OO	O O S OO O S	P	O
	H-CE	QP P	S SP SS S QQQP P	SSO S S	S S SS S OSO	OO	S Q S SS SOOOO		P	O O
1	C-1	O S S	S S S	P S O OOOO S S	S S O Q	O O O	S S OOS	O O OO O P P QQQQ	S S SSS	O
s	C-2	S	S	P	O S S S SS SS O	O O	OO SS	Q Q P P OOOO	S	O OS
t	G-2	S			S OS	O	S S S S	P OS S	P P S	
	G-1				SO S SSSSS O	S S	S SS	P S S	P P S S S S	
	HC-D	S		S SSSS	O	O P P	SS	S S SS O		S
M	HC-CE				O O	P P	SS	SSSS		S S
L	HCG-D				S	S		S		
	HCG-CE				S S					
	OH	S SS S	SSSSSS SS S S S S	SSSSS SSSSS SS SSSSS SSS	SSS S S S SSSS	SS SSS S SS	SS S SSS SSS	SS S SSSS SSSS SS SSS SS SSSS	SS S SS S SSSSS	SSSS SSS
	H-D			P	S	O	O Q	O O SS	O Q O	O O
	H-CE			O O P	R SR RR	O	O O	Q	O O Q	
2	C-1			O		O	O	P R	O R	O O
n	C-2			O O SS	O OO OO	O	O O	P	O O	
d	G-2			Q Q PP		O	S	O O PP	O S O	O P
	G-1			O O PP	O OO OO		O	PP		Q P
	HC-D			S		O	Q	O O	O O	O S
M	HC-CE			O O	O OO OO	S	O S		Q S	
L	HCG-D					S		S S	S S	S
	HCG-CE			S S	S SS SS		S			
	OH									
	H-D			O			O	O OO	O O	O
	H-CE			RR Q			O	Q	O	
3	C-1			S			R	R RR	S	O
r	C-2			OO			O	O	O	
d	G-2			QQ			O	O OO	O	O
	G-1			OO				QQ		Q
	HC-D			S			O	O OO	O	S
M	HC-CE			OO			Q	S	Q	
L	HCG-D						S	S SS	S	S
	HCG-CE			SS						
	OH									

Vertical lines divide each exons. According to likelihood values, patterns are arranged from top to bottom. [a]MPU means ML patterns used. Letters (O, P, Q, R, and S) in this table are followed figure 2.4. Equal signs surrounded with angled brackets designate regions of gene conversions. H: human, C: chimpanzee, G: gorilla, HC: common ancestor of human and chimpanzee, HCG: common ancestor of HC and gorilla, OH: the branch between hominoid and Old World monkey.

Table 2.4

Patterns of site changes for each branches of assuming tree E estimated from the maximum likilihood analysis



Details follow table 2.3. CG-D2: common ancestor of chimpanzee 2 and gorilla 1 genes, D-D2: common ancestor of HCG-D and CG-D2

Table 2.5

Gene conversion events occurred in hominoid phylogeny

			Tree D		Tree E	
ID	Branch[a]	Exon	Sites	Direction	Sites	Direction
1	HC-Human	1	31-102	D ➡ CE	48-102	CE ➡ D
2	HCG-Gorilla	3	380-399	CE ➡ D	397-399	D2 ➡ D
3	HC-Human	3	391-457	D ➡ CE	391-457	D ➡ CE
4	HC-Human	4	579-584	D ➡ CE	-	-
5	HCG-HC	5	764-797	Undetermined	764-797	Undetermined
6	HC-Human	6	808-852	CE ➡ D	808-815	Undetermined
7	HC-Chimpanzee	7	1039-1061	CE ➡ D	1039-1061	D2 ➡ D
8	HCG-Gorilla	7	1057-1059	D ➡ CE	1057-1059	D ➡ D2
9	HCG-Gorilla	8	1075-1093	Undetermined	1075-1093	Undetermined
10	HC-Human	8	1122-1124	CE ➡ D	-	-
11	HCG-Gorilla	9	1170-1193	D ➡ CE	1170-1193	D ➡ D2

[a] HC: common ancestor of human and chimpanzee

HCG: common ancestor of HC and gorilla

Table 2.6

Estimation of number of substitutions between each nodes under tree D

Branch	$d_S \times 10^{-3}$	s	S	$d_N \times 10^{-3}$	n	N
HuD-c	3.04	1	329.70	10.24	9	885.30
HuCE-e	12.27	4	330.14	20.63	18	884.86
Ch1-c	21.63	7	330.68	12.55	11	884.32
Ch2-e	3.05	1	329.01	14.82	13	885.99
Go1-d	12.21	4	331.54	11.22	10	898.46
Go2-b	18.34	6	332.28	8.97	8	897.72
Cem1-g	3.01	1	333.14	5.47	5	917.86
Cem2-g	3.01	1	332.89	3.27	3	918.11
Rhm-f	12.12	4	333.14	4.37	4	917.86
a-b	5.96	2	337.66	5.49	5	913.34
a-d	0.00	0	337.42	9.92	9	913.58
a-f	43.29	14	335.09	91.94	79	915.91
b-c	3.05	1	328.71	9.08	8	886.29
d-e	3.05	1	328.35	5.66	5	886.65
f-g	3.01	1	332.89	5.47	5	918.11

d_S and d_N mean the number of the synonymous substitutions per synonymous sites and the number of the nonsynonymous substitutions per nonsynonymous sites, respectively. s and n mean the number of synonymous and nonsynonymous differences, respectively. S and N mean the numbers of synonymous and nonsynonymous sites for the sequences compared. Internal node designations used for defining branches follow those of figure 2.5A. Hu: human, Ch: chimpanzee, Go: gorilla, Cem: crab-eating macaque, Rhm: rhesus macaque

Table 2.7

Estimation of number of substitutions between each nodes for under tree E

Branch	$d_S \times 10^{-3}$	s	S	$d_N \times 10^{-3}$	n	N
HuD-d	0.00	0	341.31	8.94	8	900.69
HuCE-a	11.92	4	339.60	16.81	15	902.40
Ch1-d	25.41	8.5	342.55	13.90	12.5	908.45
Ch2-e	2.95	1	339.67	14.40	13	911.33
Go1-e	15.70	5	323.56	11.58	10	870.44
Go2-c	15.56	5	325.45	6.94	6	868.55
Cem1-g	2.99	1	335.15	5.48	5	915.85
Cem2-g	2.99	1	334.90	3.28	3	916.10
Rhm-f	12.05	4	335.15	4.38	4	915.85
b-c	5.91	2	340.62	5.51	5	910.38
b-e	0.00	0	339.50	11.05	10	911.50
b-a	2.95	1	340.05	8.83	8	910.95
c-d	2.94	1	341.22	11.08	10	909.78
a-f	43.00	14	337.27	93.57	80	913.73
f-g	2.99	1	334.90	5.48	5	916.10

d_S and d_N mean the number of the synonymous substitutions per synonymous sites and the number of the nonsynonymous substitutions per nonsynonymous sites, respectively. s and n mean the number of synonymous and nonsynonymous differences, respectively. S and N mean the numbers of synonymous and nonsynonymous sites for the sequences compared. Internal node designations used for defining branches follow those of figure 2.5B. Hu: human, Ch: chimpanzee, Go: gorilla, Cem: crab-eating macaque, Rhm: rhesus macaque

Figure 2.1 Phylogenetic networks for human (A) and chimpanzee (B) Rh blood group genes. Numbers are nucleotide positions responsible for corresponding edges and edge lengths are proportional to number of nucleotide differences. Sites with asterisk mean sites that change amino acids. Open circles denote sequences used in the following analyses.

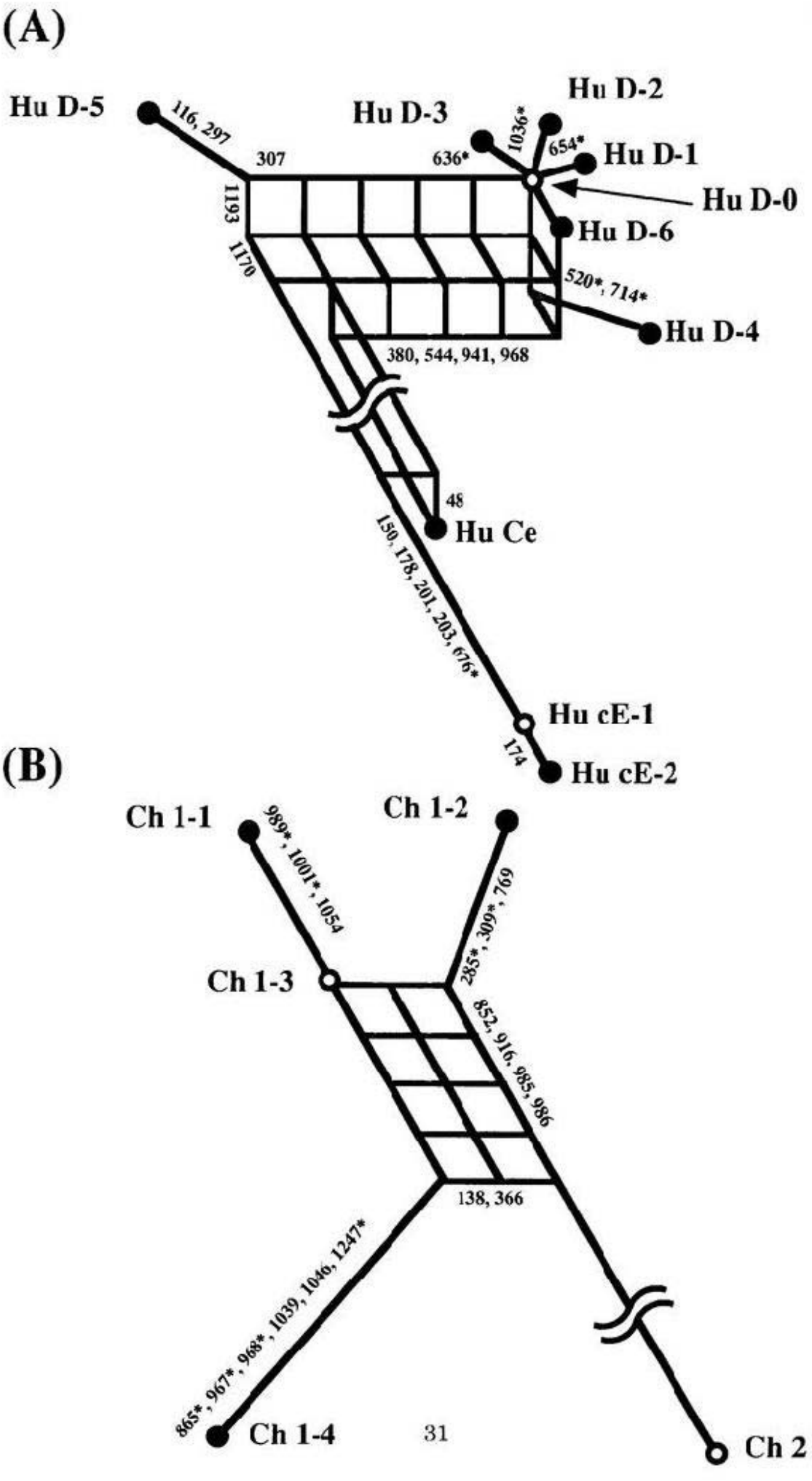

(A)
Hu D-5
116, 297
Hu D-3
Hu D-2
1036*
654*
Hu D-1
307
636*
1193
Hu D-0
Hu D-6
1170
520*, 714*
Hu D-4
380, 544, 941, 968
48
Hu Ce
150, 178, 201, 203, 676*
Hu cE-1
174
Hu cE-2
(B)
Ch 1-1
989*, 1001*, 1054
Ch 1-2
285*, 309*, 769
Ch 1-3
852, 916, 985, 986
138, 366
865*, 967*, 968*, 1039, 1046, 1247*
Ch 1-4
Ch 2

Figure 2.2 Phylogenetic networks for hominoid Rh blood group gene exons 1, 2, and 3 (A), exons 4 and 5 (B), exon 6 (C), exon 7 (D), and exons 8, 9, and 10 (E). Numbers are nucleotide positions responsible for corresponding edges and edge lengths are proportional to number of nucleotide differences. Full circles denote observed sequences. Hu, Ch, and Go mean human, chimpanzee, and gorilla, respectively.

Figure 2.2

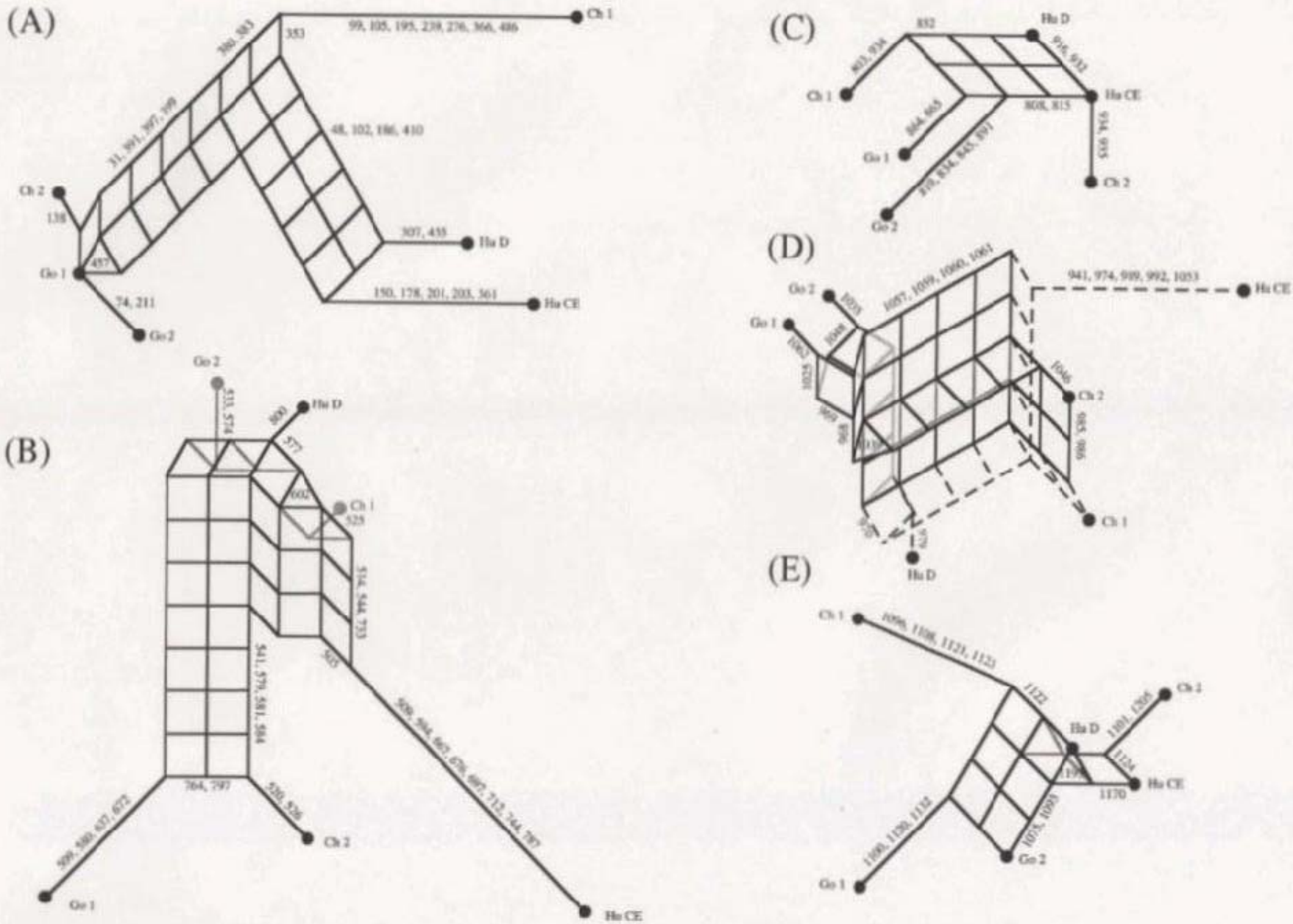

(A)
Ch 1
99, 105, 195, 239, 276, 366, 486
353
380, 383
31, 391, 397, 399
48, 102, 186, 410
Ch 2
138
457
Go 1
307, 435
Hu D
150, 178, 201, 203, 361
Hu CE
74, 211
Go 2
(B)
Go 2
533, 574
800
Hu D
577
602
Ch 1
525
514, 544, 733
541, 576, 581, 994
523
509, 594, 665, 676, 697, 713, 744, 797
764, 797
550, 556
Ch 2
509, 580, 637, 672
Go 1
Hu CE
(C)
832
Hu D
803, 934
916, 933
Ch 1
808, 815
Hu CE
864, 665
914, 955
Go 1
208, 834, 845, 891
Ch 2
Go 2
(D)
941, 974, 989, 992, 1053
Hu CE
Go 2
1057, 1059, 1060, 1061
Go 1
1046
Ch 2
986, 955
Ch 1
Hu D
(E)
Ch 1
1096, 1106, 1121, 1123
1122
Hu D
1181, 1205
Ch 1
1234
Hu CE
1170
1099, 1130, 1132
1075, 1093
Go 2
Go 1

Figure 2.3 Three model trees (A-C) and two assumed trees (D and E). Diamonds mean gene duplications. We can assume at least two gene duplications (represented by diamonds) occurred in the hominoid lineage because of chimpanzee possesses three Rh-like loci (trees A-C). Names of genes in these model trees are arbitrary. Two assumed trees (D and E) were used in the following analyses. Names of genes in these trees are actual gene names. Gene names in parentheses indicate undetected or deleted genes. The cluster including chimpanzee 2 and gorilla 1 is named D2 loci cluster in the case of tree E.

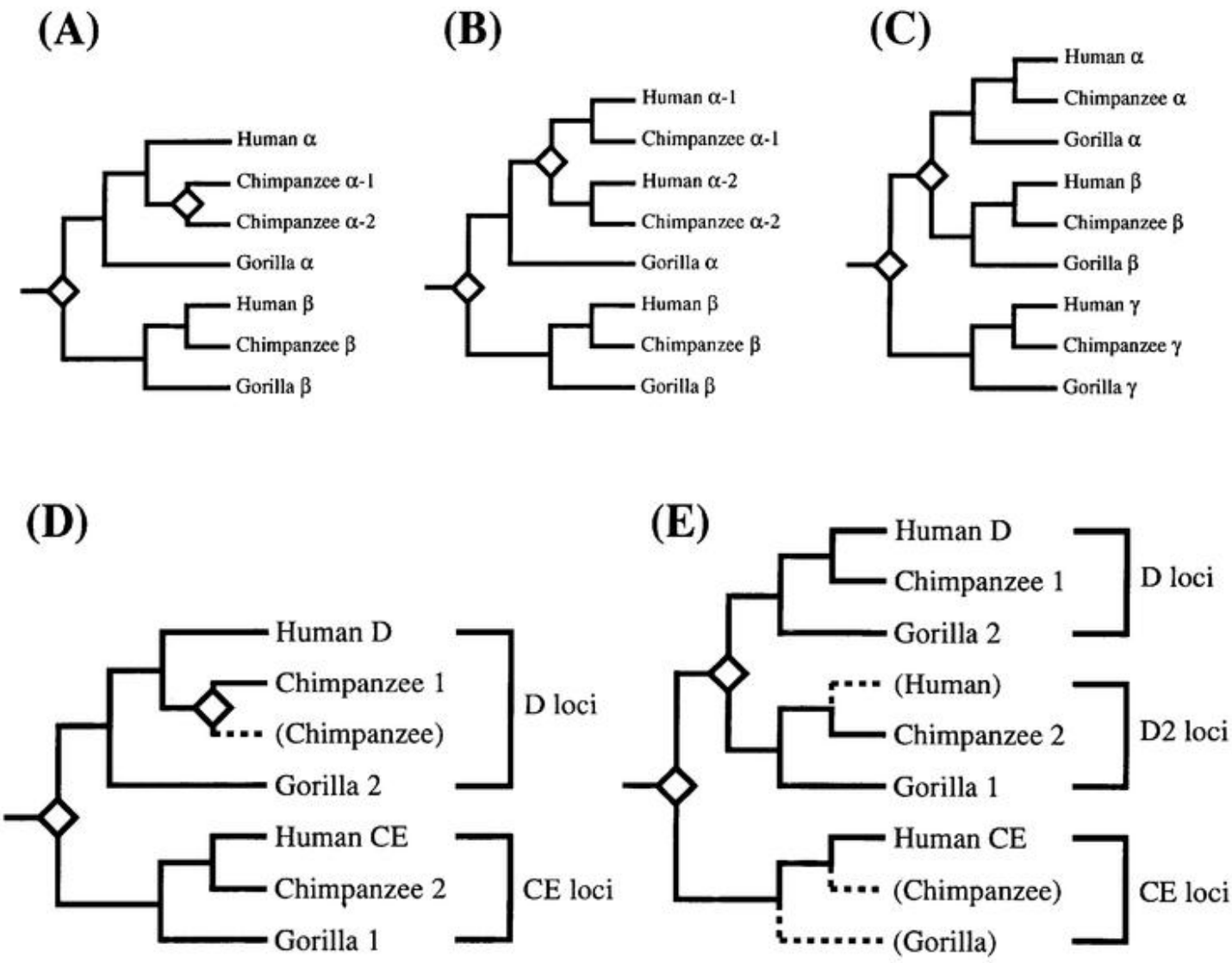

(A)
Human α
Chimpanzee α-1
Chimpanzee α-2
Gorilla α
Human β
Chimpanzee β
Gorilla β
(B)
Human α-1
Chimpanzee α-1
Human α-2
Chimpanzee α-2
Gorilla α
Human β
Chimpanzee β
Gorilla β
(C)
Human α
Chimpanzee α
Gorilla α
Human β
Chimpanzee β
Gorilla β
Human γ
Chimpanzee γ
Gorilla γ
(D)
Human D
Chimpanzee 1
(Chimpanzee)
Gorilla 2
D loci
Human CE
Chimpanzee 2
Gorilla 1
CE loci
(E)
Human D
Chimpanzee 1
Gorilla 2
D loci
(Human)
Chimpanzee 2
Gorilla 1
D2 loci
Human CE
(Chimpanzee)
(Gorilla)
CE loci

Figure 2.4 Explanation of the site by site reconstruction method. To classify variant sites, we used five symbols. The tree of figure 2D was assumed. In case (A), the root nucleotide is G, thus variants are chimpanzee 1 and 2. These variants are caused by shared nucleotide changes (form G to C) between the duplicated genes in a single species, suggesting a gene conversion in chimpanzee over this site, and they are presented by "P". In case (B), the root nucleotide is A and changes from A to T occurred in the chimpanzee 1 and the gorilla 1 genes. These nucleotide changes are not shared by the duplicated genes of the same species, they are thus true parallel substitutions, and they are designated by "S". In case (C), the root nucleotide is T and a change (from T to A) occurred in the common ancestor of orthologous human CE, chimpanzee 2, and gorilla 1. This change is not shared with its paralogous counterparts and is indicated by "S". There are no changes in its descendants, suggesting no gene conversion after the hominoid divergence over this site. Therefore, we give "O" to those branches. Case (D) is a kind of sequel to case (C), and additional nucleotide changes occurred in the human D and chimpanzee 2 genes. These additional changes resulted in the same nucleotide with their paralogous counterparts, suggesting the direction of gene conversion from the human CE to the human D genes and a gene conversion from chimpanzee 1 to chimpanzee 2 over this site, respectively. Therefore, "Q" is given to the human D gene and "R" is given to the chimpanzee 2 gene.

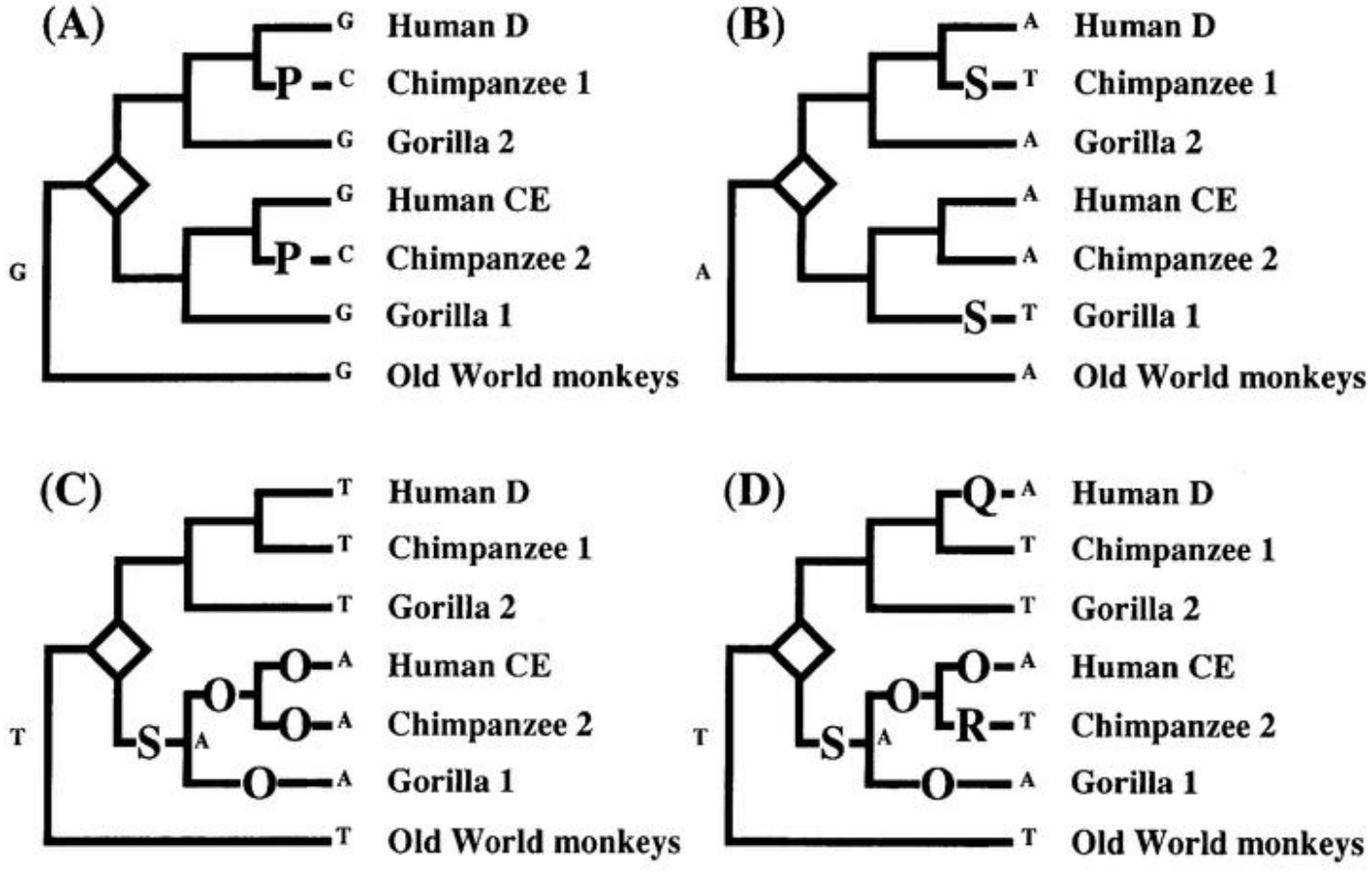
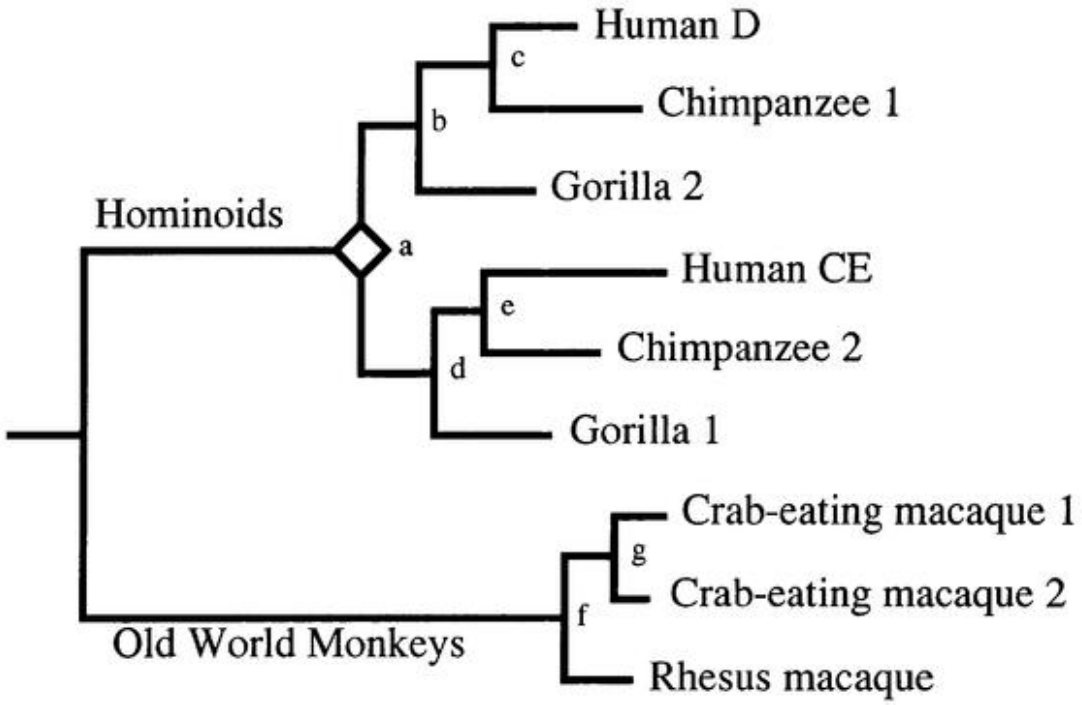
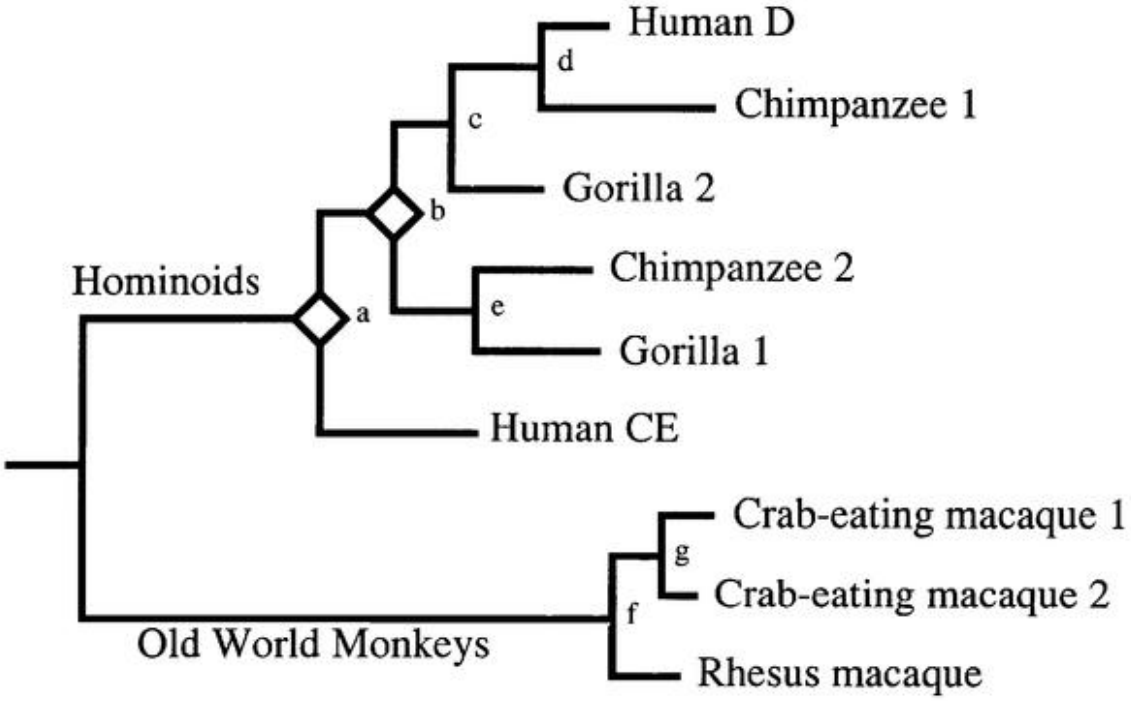

(A)
G
P
C
G
G
P
C
G
G
Human D
Chimpanzee 1
Gorilla 2
Human CE
Chimpanzee 2
Gorilla 1
Old World monkeys
(B)
A
S
T
A
A
A
S
T
A
Human D
Chimpanzee 1
Gorilla 2
Human CE
Chimpanzee 2
Gorilla 1
Old World monkeys
(C)
T
T
T
O
A
O
A
S
A
O
A
T
Human D
Chimpanzee 1
Gorilla 2
Human CE
Chimpanzee 2
Gorilla 1
Old World monkeys
(D)
Q
A
T
O
A
O
R
T
S
A
O
A
T
Human D
Chimpanzee 1
Gorilla 2
Human CE
Chimpanzee 2
Gorilla 1
Old World monkeys

Figure 2.5 The two possible gene trees (A and B) for the primate Rh blood group genes after eliminating the effect of gene conversions assuming tree D and tree E in figure 2, respectively. Diamonds designate gene duplications.

(A)


Human D
c
Chimpanzee 1
b
Gorilla 2
Hominoids
a
Human CE
e
Chimpanzee 2
d
Gorilla 1
Crab-eating macaque 1
g
Crab-eating macaque 2
f
Old World Monkeys
Rhesus macaque

(B)


Human D
d
Chimpanzee 1
c
Gorilla 2
b
Chimpanzee 2
Hominoids
a
e
Gorilla 1
Human CE
Crab-eating macaque 1
g
Crab-eating macaque 2
f
Old World Monkeys
Rhesus macaque

CHAPTER III

EVOLUTION OF RODENT RH BLOOD GROUP GENES AND THEIR RELATED GENES

PCR-direct sequencing of cDNA

Five mouse (*Mus musculus*) subspecies (*M. m. domesticus* (C57BL/10SnSlc), *M. m. brevirostris* (BFM/2MsfB6C3FI), *M. m. musculus* (BLG2/MsfB6C3FI), *M. m. castaneus* (CAST/Ei), and *M. m. molossinus* (MSM/Msf)), rat (*Rattus norvegicus* (Std: Wistar)), and crab-eating macaque (*Macaca fascicularis*) were used. Mice are provided from Laboratory of Mammalian Genetics (kindly provided by Drs. Toshihiko Shiroishi and Tomoko Sagai), National Institute of Genetics. Rat was bought from Nihon SLC. Samples of crab-eating macaque were kindly provided by Dr. Yuzuru Ikehara at Division of Cell Biology, Institute of Life Science, Soka University.

Bone marrow cells were extracted from femora by using RPMI Medium 1640 (Gibco-BRL) and 26G×1/2 syringe (TERUMO) and filtered 70 μm Cell Strainer (FALCON). Total RNAs were extracted by using the AGPC (Acid Guanidinium-Phenol-Chloroform) method. To remove DNA, DNase reactions were carried out by using Deoxyribonuclease (RT Grade) (Nippon Gene). Reverse transcription was performed by using AMV (Avian Myeloblastosis Virus) reverse transcriptase and oligo dT-adaptor primer of RNA PCR Kit AMV Ver. 2.1 (TaKaRa).

Degenerate PCR was performed and a partial product was obtained. I then performed 5' RACE (rapid amplification of the 5' cDNA ends) using 5'RACE System for Rapid Amplification of cDNA Ends version 2.0 (Gibco-BRL). 3'RACE is also curried out. To amplify the complete cDNA sequence, PCR was performed by using gene

specific primers. PCR was performed in a 20 μl reaction containing 0.5-1 μl of the first-strand cDNA, 1×Gene Taq Universal Buffer (Mg^{2+} free) (Nippon Gene), 1.5 mM $MgCl_2$, 0.2 mM dNTP, 10 pmol of each primer (designed on sites of 5' and 3' ends), and 1 units of AmpliTaq Gold (Perkin-Elmer). Amplification was carried out in DNA GeneAmp PCR System 2400 (Perkin-Elmer) with the following temperature parameters: 10 min at 95℃ followed by 40 cycles of 95℃ for 30 sec, 65℃ for 15 sec, and 72℃ for 1 min. PCR products were purified using MicroSpin Columns S-300 HR (Pharmacia Biotech). DNA sequencing was performed on PCR products using Dye Terminator Cycle Sequencing Kit and ABI prism 377 DNA sequencer (Perkin-Elmer). A progressive sequencing strategy was carried out with design of further primers to complete the sequence for coding region of both strands of the cDNA. Figure 3.1 shows the sequencing scheme for the Rh gene of *M. m. domesticus*. Other genes were also determined by using the same procedure.

Sequence comparisons of Rh and Rh50 cDNA coding regions

I sequenced Rh and Rh50 gene cDNAs for two mouse subspecies and rat. Those newly determined rodent sequences (DDBJ/EMBL/GenBank international nucleotide sequence database accession numbers are AB015189 - AB015194) were compared with human and crab-eating macaque Rh genes (Cherif-Zahar et al. 1990; Avent et al. 1990; Salvignol et al. 1995) and the human Rh50 gene (Ridgwell et al. 1992). Figure 3.2 shows the multiple alignment of nucleotide sequences of rodent Rh genes. CLUSTAL W version 1.6 (Thompson, Gibson, and Higgins 1994) was used for multiple alignment. Nucleotide sequence lengths of human, crab-eating macaque, mouse, and rat are 1254 bp, 1254 bp, 1257 bp, and 1269 bp, respectively. Four gaps (3, 15, 3, and 6 nucleotide long) were observed between primate and rodent sequences, and the rat Rh gene had extra

12 nucleotides (positions 337-348). Lengths of all gaps were multiplication of 3 and there is no frame shift. I also obtained an incomplete sequence for the rat Rh cDNA which lacks sites 149-661. These sites correspond to exons 2-4 of the human Rh gene, and this incomplete cDNA were probably produced by a splicing error.

Figure 3.3 shows the multiple alignment of nucleotide sequences of Rh50 genes. I also obtained a Rh50 gene cDNA for crab-eating macaque (DDBJ/EMBL/GenBank accession number is AB015467), and it was also compared. Nucleotide sequence lengths of human, crab-eating macaque, mouse, and rat are 1230 bp, 1287 bp, 1317 bp, and 1353 bp, respectively. The location of the stop codon of the human Rh50 gene is different from that of others, and its protein is 19 amino acids shorter corresponding to this region. There are repeats of 15 nucleotides around positions 100-150 (see figure 3.3), and its consensus sequence is AATGCTTCCCAGCAG. Rat and mouse have 5 and 3 repeats, respectively, while the two primate species have single repeat. Because all gaps were multiple of 3, they did not alter codon frames.

Sequence similarities (both for nucleotide and amino acid) are shown in table 3.1. Because nucleotide differences among five mouse subspecies did not differ each other, sequences for *M. m. domesticus* are used to estimate values. Nucleotide and amino acid sequence similarities between Rh genes and Rh50 genes are 47.2-48.9 % and 34.4-37.8 %, respectively. The GC contents of Rh and Rh50 genes were 52.5-55.2 % and 45.0-47.4 %, respectively (shown on the diagonal of table 3.1). These values were similar to those previously reported (Matassi et al. 1998), and may be related to gene locations on genomes; the Rh gene is located on chromosome 1p34-36 (Ruddle et al. 1972; Cherif-Zahar et al. 1991), while the Rh50 gene is on chromosome 6p21-qter (Ridgwell et al. 1992)

I constructed phylogenetic networks of rodent Rh (figure 3.4A) and Rh50 (figure 3.4B) genes. Two phylogenetic networks showed incompatibility of the phylogenetic relationship of mouse subspecies. The existence of ancestral polymorphism for these

genes is suggested.

Estimation of evolutionary rates and comparison with primate Rh blood group genes

I estimated numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions for Rh and Rh50 genes (table 3.2). Because nucleotide differences among five mouse subspecies did not differ each other, sequences for *M. m. domesticus* and *M. m. brevirostris* are used to estimate values. ODEN package (Ina 1994) was used to estimate numbers of synonymous and nonsynonymous substitutions (Nei and Gojobori 1986). d_S and d_N values between primates and rodents were estimated by averaging pairwise values. Numbers of synonymous substitutions (d_S) were similar between Rh and Rh50 genes, and they are more or less similar to those for other genes (Li and Tanimura 1987). Branching pattern of the Rh and Rh50 genes are also compatible with the established mammalian phylogeny. This indicates that I did orthologous comparison both for Rh and Rh50 genes.

Numbers of nonsynonymous substitutions (d_N) are about two times higher for the Rh gene than for the Rh50 gene; the ratios of Rh-d_N and Rh50-d_N are 2.0, 1.7, and 2.0 for human-macaque, mouse-rat, and primates-rodents comparisons, respectively (I neglected the comparison of the two mouse subspecies, for standard errors are so large). This evolutionary conservation of the Rh50 gene suggests that it may have more important function than the Rh gene. A relatively uniform ratio of Rh-d_N and Rh50-d_N for three different levels of divergence also suggests that a molecular clock (constancy of evolutionary rate) exists both for Rh and Rh50 genes.

Majority of genes are known to undergo neutral evolution, and number (d_S) of synonymous substitutions are expected to be higher than those (d_N) for nonsynonymous

substitutions under this situation (Kimura 1983). I compared d_S and d_N values to see if there is any unusual pattern deviated from neutrality in Rh and Rh50 genes. d_N of both Rh and Rh50 genes were higher than d_S when human and macaque sequences were compared, while the situation is reversed for other comparisons (table 3.2). I discussed in chapter two that many branches of a phylogenetic tree of primate Rh genes showed higher d_N than d_S, and this is compatible with a higher d_N for human and macaque Rh gene shown in table 3.2. It is interesting that the Rh50 gene also showed a similar evolutionary pattern for primates, but not for rodents. If the heterotetramer structure of the Rh and Rh50 gene products is correct, it is possible that this erythrocyte membrane protein complex is under some kind of positive selection in primates but not in rodents.

Evidence for higher rates of nucleotide substitution in rodents than in primates

Figure 3.5 shows comparison of d_S between Rh and Rh50 genes. Because d_S were similar between Rh and Rh50 genes, these values were almost plotted on the line of an angle of 45 degrees. It suggests that a molecular clock (constancy of evolutionary rate) exists both for Rh and Rh50 genes. Numbers in parentheses in this figure are relative evolutionary distances. It is suggested that mutation rates in rodents is higher than those in primates (Wu and Li 1985; Gu and Li 1992). I estimated relative rates between primate and rodent lineage and the divergence time between mouse and rat from these rates. Figure 3.6A shows the scheme for this procedure. The divergence time between primates and rodents was assumed to be 115-129 million years ago (MYA) (Easteal, Collet, and Betty 1995). For simplicity, I used the mean value (122 MYA) as this divergence time in this figure. Relative evolutionary distances between mouse and rat, and between human and macaque were 0.35 and 0.1, respectively. If I assume that the

relative rate (R) between primate and rodent lineages is equal (indicated by the arrow with number 1), and divergence times between human and macaque, and between mouse and rat were estimated to be 12.2 MYA {=(0.1/1)×[122(MYA)/2]×(2/1)} and 42.7 MYA {=(0.35/1)×[122(MYA)/2]×(2/1)}. If I assume R = 2 (two times higher in rodents than primates; indicated by the arrow 2 in figure 3.6A), divergence times between human and macaque, and between mouse and rat were estimated to be 18.3 MYA {=(0.1/1)×[122(MYA)/2]×(3/1)} and 32.025 MYA {=(0.35/1)×[122(MYA)/2]×(3/2)}. These values are given by

$$T_{hm} = (D_{hm}/D_{pr})\times(T_{pr}/2)\times(R+1) = (0.1/1)\times(122/2)\times(R+1), \quad (3.1)$$

$$T_{mr} = (D_{mr}/D_{pr})\times(T_{pr}/2)\times([R+1]/R) = (0.35/1)\times(122/2)\times([R+1]/R), \quad (3.2)$$

where T_{mr}, T_{hm}, and T_{pr} are divergence times (MYA) between mouse and rat, between human and macaque, and between primates and rodents, respectively, and D_{mr}, D_{hm}, and D_{pr} are relative evolutionary distances between mouse and rat (0.35), between human and macaque (0.1), and between primates and rodents (1), respectively. Figure 3.6B summarized the relationship between R and divergence times. From this figure, if I assume the divergence time between human and macaque is 23 MYA (Kumar and Hedges 1998), the R is about 2.8, then the divergence time between mouse and rat becomes ca. 30 MYA. This suggests that the mutation rate in rodents is about three times higher than that in primates. In this case synonymous substitutions (per site per year) of primate and rodent lineages were estimated to be about 1.6×10^{-9} and 4.4×10^{-9}. The divergence time between mouse and rat is a matter of argument. Jacobs and Pilbeam (1980) estimated it to be 8-14 MYA, but Wilson, Carlson, and White (1977) argued that it can be anywhere between 5-35 MYA. Recently, Kumar and Hedges (1998) estimated it to be 40.7 MYA based on 309 gene comparison. However the variance of the divergence time between mouse and rat is larger than that between human and macaque. In any case my result is compatible with Wilson, Carlson, and White 's (1977) traditional view.

Window analyses of synonymous and nonsynonymous substitutions

I performed window analyses for synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions to investigate their possible correlation with the protein structure (figure 3.7). The WINA program (Endo, Ikeo, and Gojobori 1996) was used for window analyses. The twelve predicted hydrophobic membrane-spanning regions are shown by black boxes with numbers. The PredictProtein server (EMBL) was used for analyses of transmembrane helix location.

There are several peaks (depicted by arrows) where nonsynonymous substitutions are higher than synonymous ones on putative outer membrane regions on primate Rh genes (figure 3.7A). One peak (designated as long arrows) is observed at the cell surface region between membrane-spanning regions 3 and 4 in all four comparisons (figures 3.7A-D). In the case of human Rh genes, one amino acid change (at position 103, P/S) on this region determine alleles c or C (Mouro et al. 1993), therefore this region is an actual outer membrane region. One peak (indicated by short arrow with asterisk) is observed at the cell surface region between membrane-spanning regions 7 and 8 in primate Rh (figure 3.7A) and Rh50 (figure 3.7C) comparisons. In the case of human Rh genes, one amino acid change (at position 226, A/P) on this region determine alleles e or E (Mouro et al. 1993), therefore this region is also an actual outer membrane region. One peak (indicated by short arrow with sharp) is observed at the cell surface region between membrane-spanning regions 11 and 12 in primate Rh (figure 3.7A) and Rh50 (figure 3.7C) comparisons. Human D protein differs from non-D proteins in 36 amino acid positions (Le Van Kim et al. 1992), but all the D antigen specific positions are not known. However, three amino acid changes (at positions 350, 353, and 354) on this region are thought to related to differences between D antigen and non-D antigens (Blancher and Socha 1997), therefore this region is also an actual outer membrane region.

One peak (indicated by white arrow) is also observed at the cytoplasmic region between membrane-spanning regions 10 and 11 in primate Rh (figure 3.7A), rodent Rh (figure 3.7B), and Rh50 (figure 3.7C) comparisons.

I also analyzed patterns of amino acid changes for each regions (trans-membrane, inner-membrane, and outer-membrane regions) of primates (human vs. crab-eating macaque) and rodents (mouse vs. rat) Rh and Rh50 genes (table 3.3). Numbers of amino acid substitutions were estimated by using Kimura's (1983) method. Numbers of amino acid substitutions were not much different from each regions in primate Rh genes. The number of amino acid substitutions of trans-membrane regions of primate Rh genes was higher than those of others. There is a possibility that primate Rh genes are released from selective constraint.

In primate Rh50 genes, the number of amino acid substitutions of outer-membrane regions is significantly greater than that of trans-membrane regions by using a one-sided t-test (at the 5 % level). In rodent Rh genes, the number of amino acid substitutions of outer-membrane regions is significantly (at the 1 % level) greater than those of inner- and trans-membrane regions. In rodent Rh50 genes the number of amino acid substitutions of outer-membrane regions is also significantly greater than those of other regions (5 % and 1 % for inner- and trans- membrane regions, respectively). It is also noted that outer- membrane regions of rodent Rh and Rh50 genes contain 4 gaps, and 1 and 11 gaps, respectively.

Effects of number of OTU

To examine effects of numbers of OTU (operational taxonomic unit) for estimation of the divergence time, I reconstructed four trees (figure 3.8) for Rh and Rh50 genes of primates and rodents by using the maximum likelihood method (go/0 program of OOta

and Saitou (1997) was used) and estimated the divergence time between Rh and Rh50 genes. Dr. Satoshi OOta kindly estimated values of branch length for maximum likelihood trees by using his program. The 223 amino acid sites for membrane-spanning regions were used for construction of trees (see figure 4.3 of Chapter IV). The root for each tree was located by assuming the Rh-like protein of sponge (Seack et al. 1997) as an outgroup. Numbers of amino acid substitutions of single-lineage were obtained applying Ishida et al.'s (1995) method from each tree (table 3.4). Because numbers of amino acid substitutions for Rh were consistently two - three times higher than those for Rh50, a rough molecular clock exists for both genes. Therefore, I estimated evolutionary rates of Rh and Rh50 genes by using the regression through origin. Divergence times between human and crab-eating macaque, between mouse and rat, and between primates and rodents were assumed to be 23 (Kumar and Hedges 1998), 30 (see above), and 122 (Easteal et al. 1995) MYAs, respectively, and they were used for calibration of the molecular clock.

In the case of tree A, the divergence time between Rh and Rh50 genes are estimated to be 130-170 MYAs. This period roughly corresponds to the early Mesozoic before the mammalian radiation. Because these values are close to the divergence time between primates and rodents, these values are probably underestimation. Acceleration of evolutionary rates for the primate lineage by positive selection may affected these values. In the case of tree B, the divergence time are estimated to be 290-320 MYAs. This period roughly corresponds to the late Paleozoic. These values are about two times higher that those of tree A. In the case of tree C, the divergence time are estimated to be 370-450 MYAs. In the case of tree D, the divergence time are estimated to be 340-380 MYAs. This period roughly corresponds to the middle Paleozoic, and the divergence between land vertebrates and amphibian lineages occurred around that period. The range of these values are smaller than those of tree C. It is suggested that numbers of calibration points for the molecular clock are needed to obtain better estimation of

divergence times. Figure 3.9 shows comparisons between amino acid substitutions and divergence times from tree D (figure 3.8). It indicates evolutionary constancy of Rh and Rh50 genes without primates. It is interested that evolutionary rates of Rh and Rh50 genes accelerate on the primate lineage. This result is consistent with the result of table 3.2. In any case, we should be very careful when we estimate the divergence time of genes.

Table 3.1

Similarities (%) of Rh and Rh50 nucleotide sequences (above diagonal) and amino acid sequences (below diagonal), and GC content (%) of each gene (on the diagonal in parentheses)

	1	2	3	4	5	6	7	8
1 Human RhcE	(53.7)	90.4	71.4	70.8	48.6	48.5	47.3	47.3
2 Macaque Rh	79.1	(52.5)	71.9	70.9	48.8	48.9	48.4	48.2
3 Mouse Rh	57.9	59.1	(55.2)	88.3	48.6	48.9	47.2	46.9
4 Rat Rh	56.7	58.6	81.6	(54.0)	48.8	48.5	47.7	47.4
5 Human Rh50	35.2	35.7	37.1	35.3	(46.7)	94.6	80.0	79.4
6 Macaque Rh50	35.4	37.7	37.8	36.3	88.8	(47.4)	79.6	78.6
7 Mouse Rh50	34.4	35.4	35.0	35.5	77.0	74.3	(45.0)	91.6
8 Rat Rh50	33.8	35.1	34.7	34.4	75.8	73.1	88.8	(45.4)

Table 3.2

Numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions

	MMD vs. MMB[a]	human vs. macaque[a]	mouse vs. rat[a]	primates vs. rodents[b]
d_S of Rh	0.013±0.007	0.071±0.016	0.226±0.031	0.595
d_S of Rh50	0.007±0.005	0.049±0.013	0.200±0.028	0.620
d_N of Rh	0.007±0.003	0.115±0.011	0.098±0.011	0.302
d_N of Rh50	0.001±0.001	0.057±0.008	0.058±0.008	0.153

[a] Pairwise values with standard errors. [b] Averages of pairwise values.

MMD and MMB designate *M. m. domesticus* and *M. m. brevirostris*, respectively.

Table 3.3

Comparisons of numbers of amino acid substitutions per site (d_A) and amino acid differences (P_A) for each regions of Rh and Rh50 genes

	Rh		Rh50	
	$d_A \pm$ S.E.	P_A	$d_A \pm$ S.E.	P_A
Primates				
trans-membrane	0.280±0.041	23.32(52/223)	0.081±0.020	7.6 (17/223)
inner-membrane	0.248±0.054	21.1 (23/109)	0.159±0.044	14.3 (14/98)
outer-membrane	0.233±0.059	20 (17/85)	0.194±0.052	17.1 (15/88)
Rodents				
trans-membrane	0.143±0.027	13 (29/223)	0.056±0.016	5.4 (12/223)
inner-membrane	0.170±0.043	15.2 (17/112)	0.131±0.036	12.0 (14/117)
outer-membrane	0.513±0.102	37.4 (31/83)	0.282±0.062	23.5 (23/98)

Numbers of different sites / numbers of sites compared are shown in parentheses of P_A.

Table 3.4

Numbers of amino acid substitutions and divergence times

Diverging node (MYA)	human/macaque (23)	mouse/rat (30)	primates/rodents (122)	Rh/Rh50	MYAs [a]
Rh (Tree A)	13.73	-	-	79.44	133
Rh (Tree B)	-	7.29	-	78.10	321
Rh (Tree C)	-	-	22.46	82.07	446
Rh (Tree D)	13.66	7.38	24.68	82.79	381
Rh50 (Tree A)	3.99	-	-	28.71	165
Rh50 (Tree B)	-	2.78	-	26.83	290
Rh50 (Tree C)	-	-	8.43	25.33	366
Rh50 (Tree D)	4.06	2.79	9.94	29.15	342

[a] Divergence time between Rh and Rh50 genes estimated from numbers of amino acid substitutions of each tree of figure 3.8.

Figure 3.1 The sequencing scheme for the Rh gene of *M. m. domesticus*. Boxes and arrows show PCR products and sequensing, respecitively. Primers are mentioned in Appendix III.

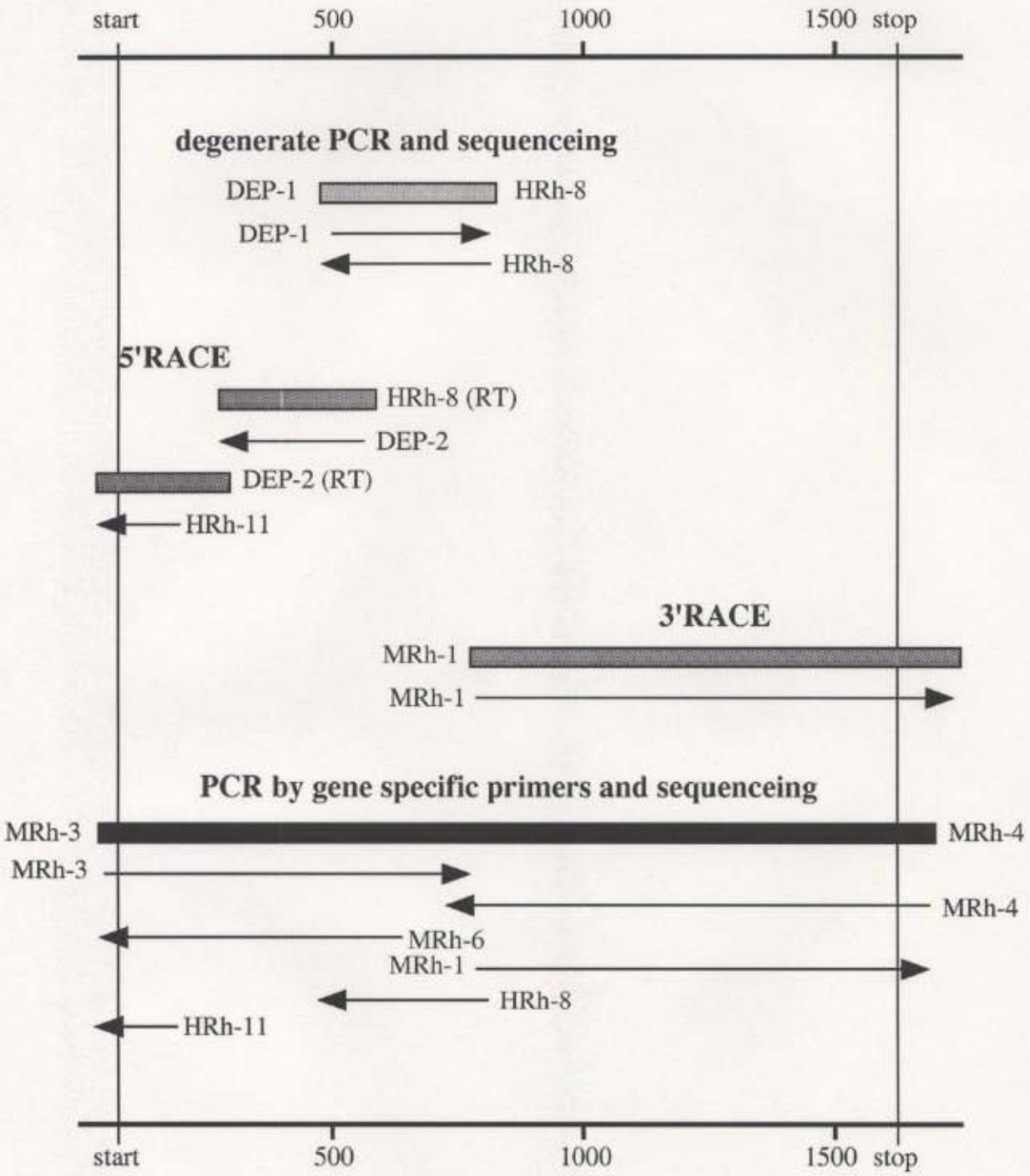

start
500
1000
1500
stop
degenerate PCR and sequenceing
DEP-1
HRh-8
DEP-1
HRh-8
5'RACE
HRh-8 (RT)
DEP-2
DEP-2 (RT)
HRh-11
3'RACE
MRh-1
MRh-1
PCR by gene specific primers and sequenceing
MRh-3
MRh-4
MRh-3
MRh-4
MRh-6
MRh-1
HRh-8
HRh-11
start
500
1000
1500
stop

Figure 3.2 The multiple alignment of nucleotide sequences of Rh genes. Nucleotide sequences for human RhcE (M34015 or X54534; they are identical) and for crab-eating macaque (L37054) were also included for comparison. Gaps are denoted by hyphens, and only nucleotides different from those of the human sequence are shown. MMD, MMB, MMMu, MMC, MMMo, and CEM denote *M. m. domesticus*, *M. m. brevirostris*, *M. m. musculus*, *M. m. castaneus*, *M. m. molossinus*, and crab-eating macaque, respectively.

Human_Rh
CEM_Rh
MND_Rh
MNB_Rh
MNMJ_Rh
MNC_Rh
MNMO_Rh
Rat_Rh

Figure 3.3 The multiple alignment of nucleotide sequences of Rh50 genes. Human Rh50 (X64594) was also included. Gaps are denoted by hyphens, and only nucleotides different from those of the human sequence are shown. Equal signs surrounded with angled brackets designate the repeat unit of 15 nucleotides. Abbreviations of species are the same as figure 3.2.

Human_Rh50
CEM_Rh50
MMD_Rh50
MMB_Rh50
MMHu_Rh50
MMC_Rh50
MMMo_Rh50
Rat_Rh50

Figure 3.4 The phylogenetic networks of Rh genes (A) and Rh50 genes (B) for five *Mus musculus* subspecies and rat. Numbers are nucleotide positions responsible for corresponding edges and edge lengths are proportional to number of nucleotide differences. Abbreviations of species are the same as figure 3.2.

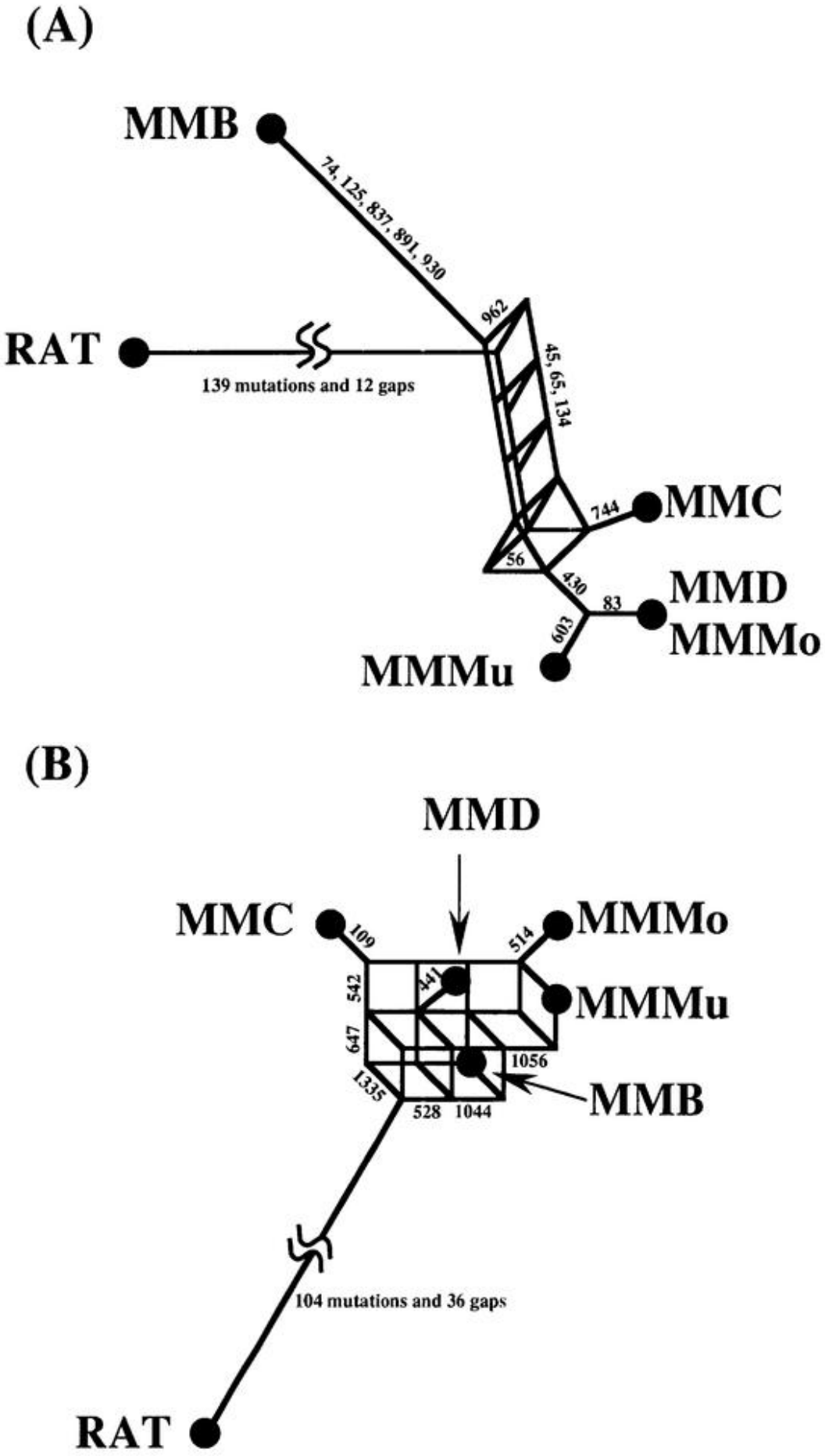

(A)
MMB
74, 125, 837, 891, 930
962
RAT
139 mutations and 12 gaps
45, 65, 134
744
MMC
56
430
83
603
MMD
MMMo
MMMu
(B)
MMD
MMC
109
514
MMMo
542
441
MMMu
647
1056
1335
MMB
528
1044
104 mutations and 36 gaps
RAT

Figure 3.5 Comparison of synonymous substitutions between Rh and Rh50 genes of primates (human and crab-eating macaque) and rodents (mouse and rat). Boxes show points of d_S with standard errors for Rh and Rh50 genes. Numbers in parentheses are relative rates.

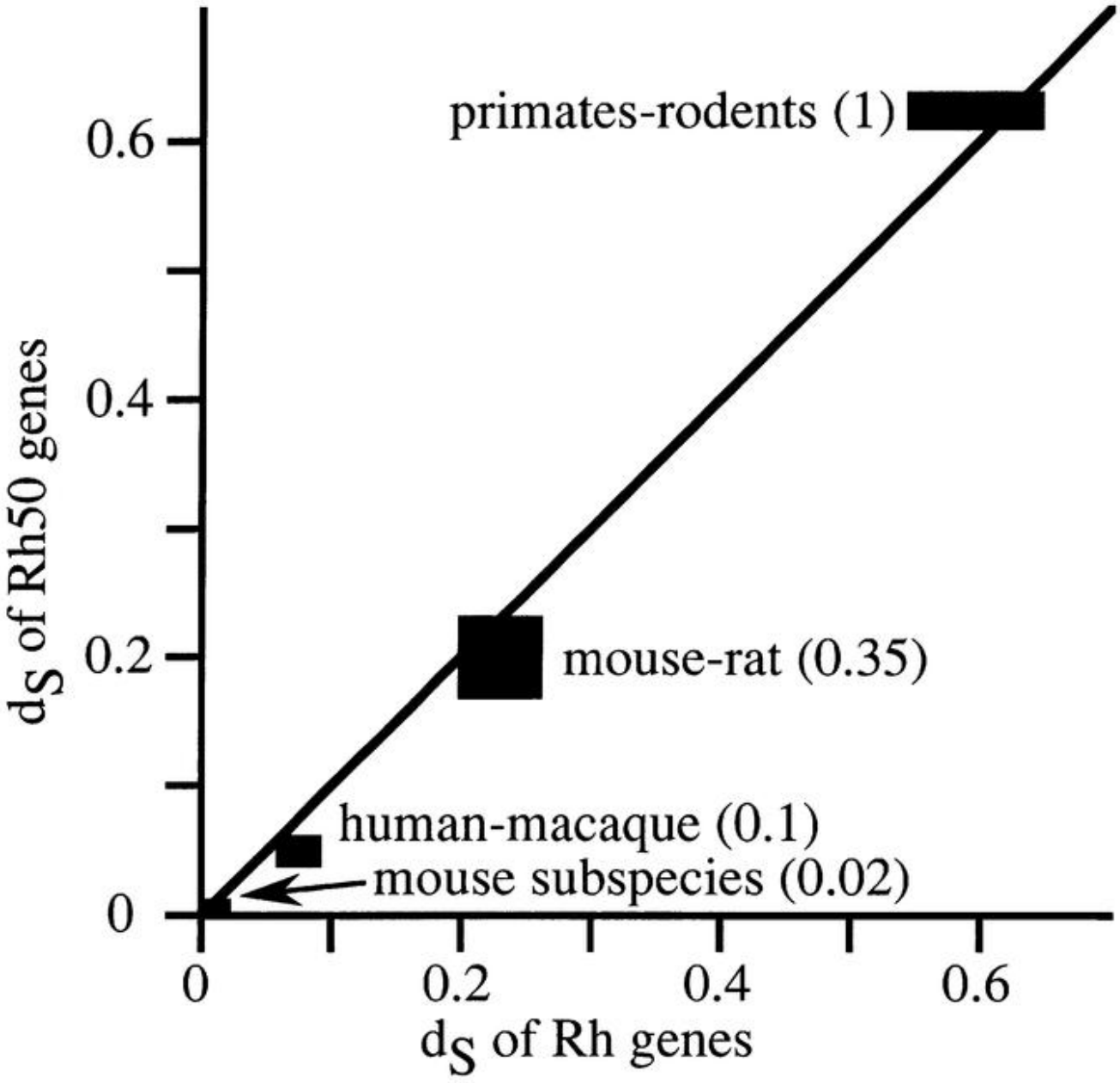
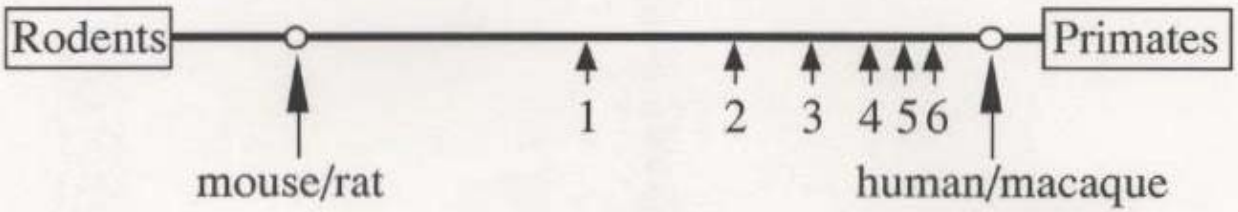

primates-rodents (1)
mouse-rat (0.35)
human-macaque (0.1)
mouse subspecies (0.02)
0.6
0.4
0.2
0
0
0.2
0.4
0.6
d_S of Rh50 genes
d_S of Rh genes

Figure 3.6 Examination of relative evolutionary rates (R) between primate and rodent lineages. (A) The scheme for this procedure. Numbers show R for each assuming divergence point between primate and rodent lineages. (B) The relationship between R and divergence times.

(A)


Rodents
Primates
1
2
3
4
5
6
mouse/rat
human/macaque

(B)

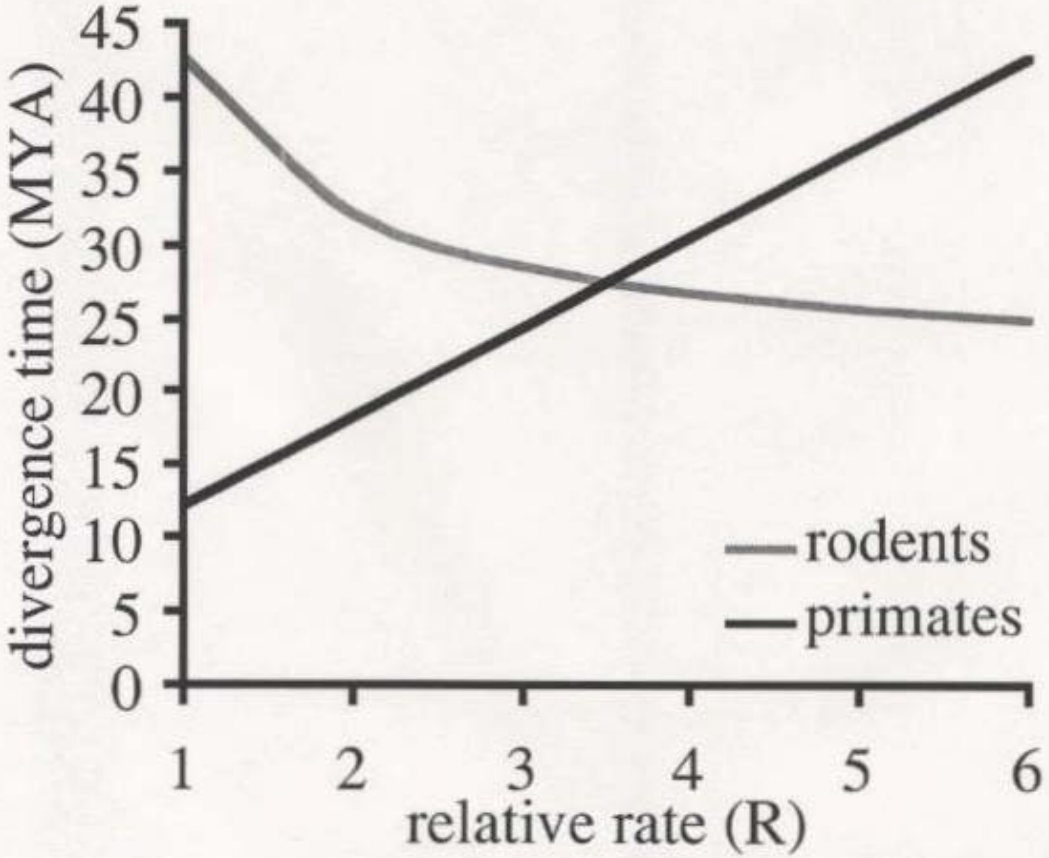

45
40
35
30
25
20
15
10
5
0
divergence time (MYA)
rodents
primates
1
2
3
4
5
6
relative rate (R)

Figure 3.7 Window analyses for synonymous (d_S: gray lines) and nonsynonymous (d_N: black lines) nucleotide substitutions for Rh genes between human and crab-eating macaque (A), for Rh genes between *M. m. domesticus* and rat (B), for Rh50 genes between human and crab-eating macaque (C), and for Rh50 genes between *M. m. domesticus* and rat (D). The 12 predicted hydrophobic membrane-spanning regions are shown by black boxes. Horizontal axes indicate numbers of nucleotide sites. See text for explanations of arrows.

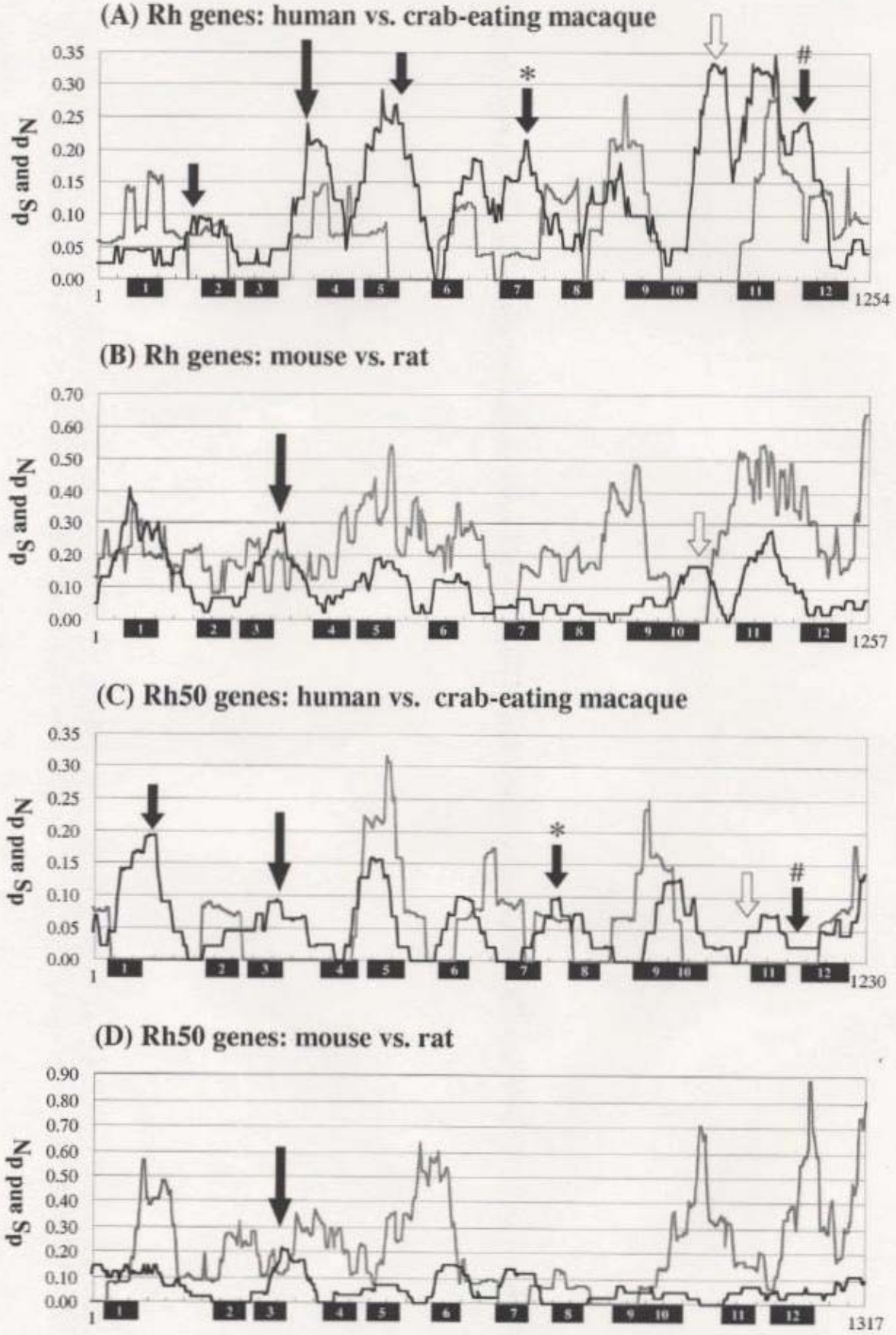

(A) Rh genes: human vs. crab-eating macaque
d_S and d_N
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
1254
(B) Rh genes: mouse vs. rat
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1257
(C) Rh50 genes: human vs. crab-eating macaque
1230
(D) Rh50 genes: mouse vs. rat
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1317

Figure 3.8 Four trees to examine effects of numbers of OTU for estimation of the divergence time. Numbers on branches show numbers of amino acid substitutions estimated from the maximum likelihood method. The root of each tree was located by assuming the Rh-like gene of sponge as an outgroup. (A) Rh and Rh50 genes for human and crab-eating macaque were used. (B) Rh and Rh50 genes for mouse and rat were used. (C) Rh and Rh50 genes for human and mouse were used. (D) Rh and Rh50 genes for human, crab-eating macaque, mouse, and rat were used.

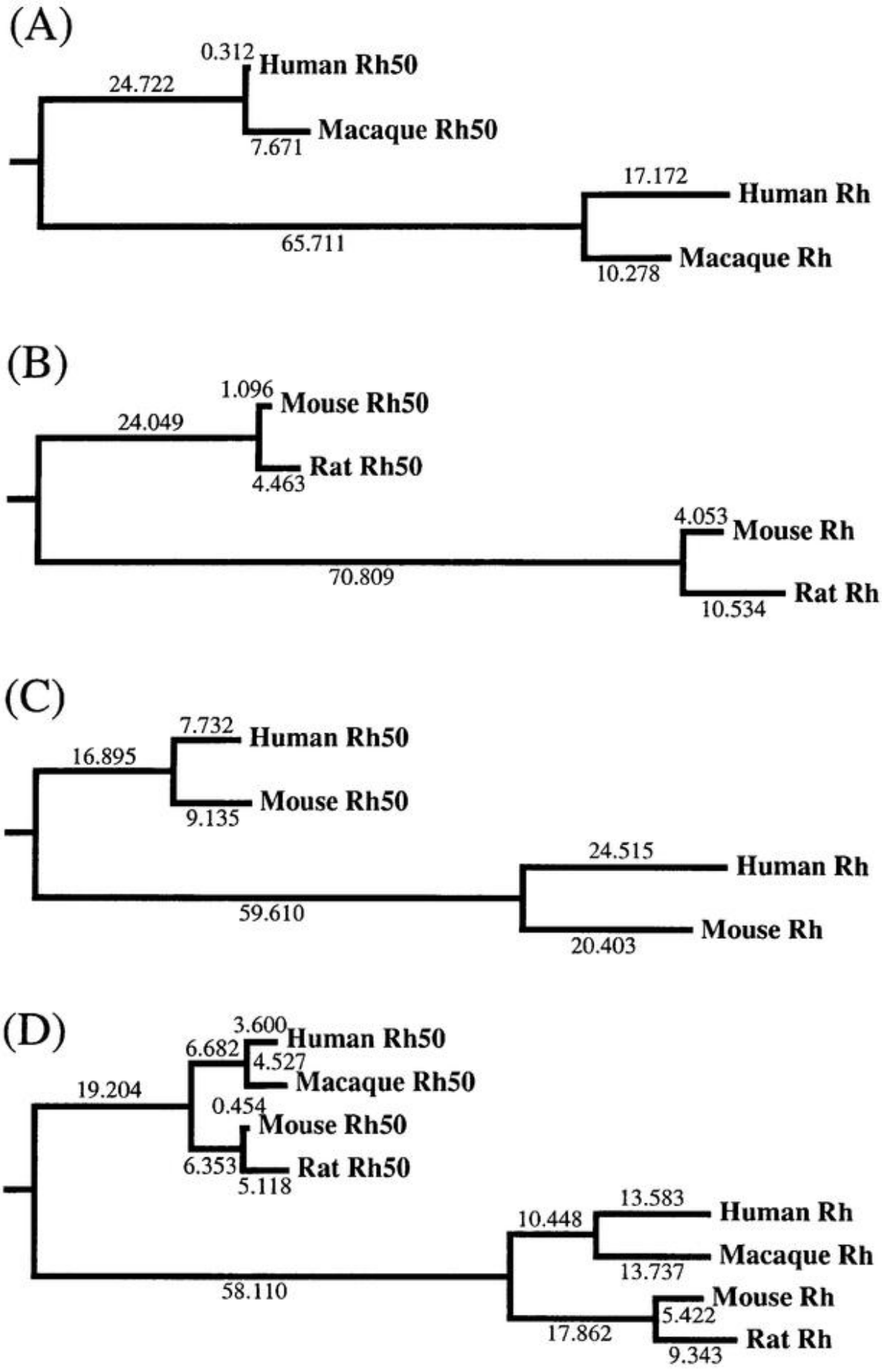
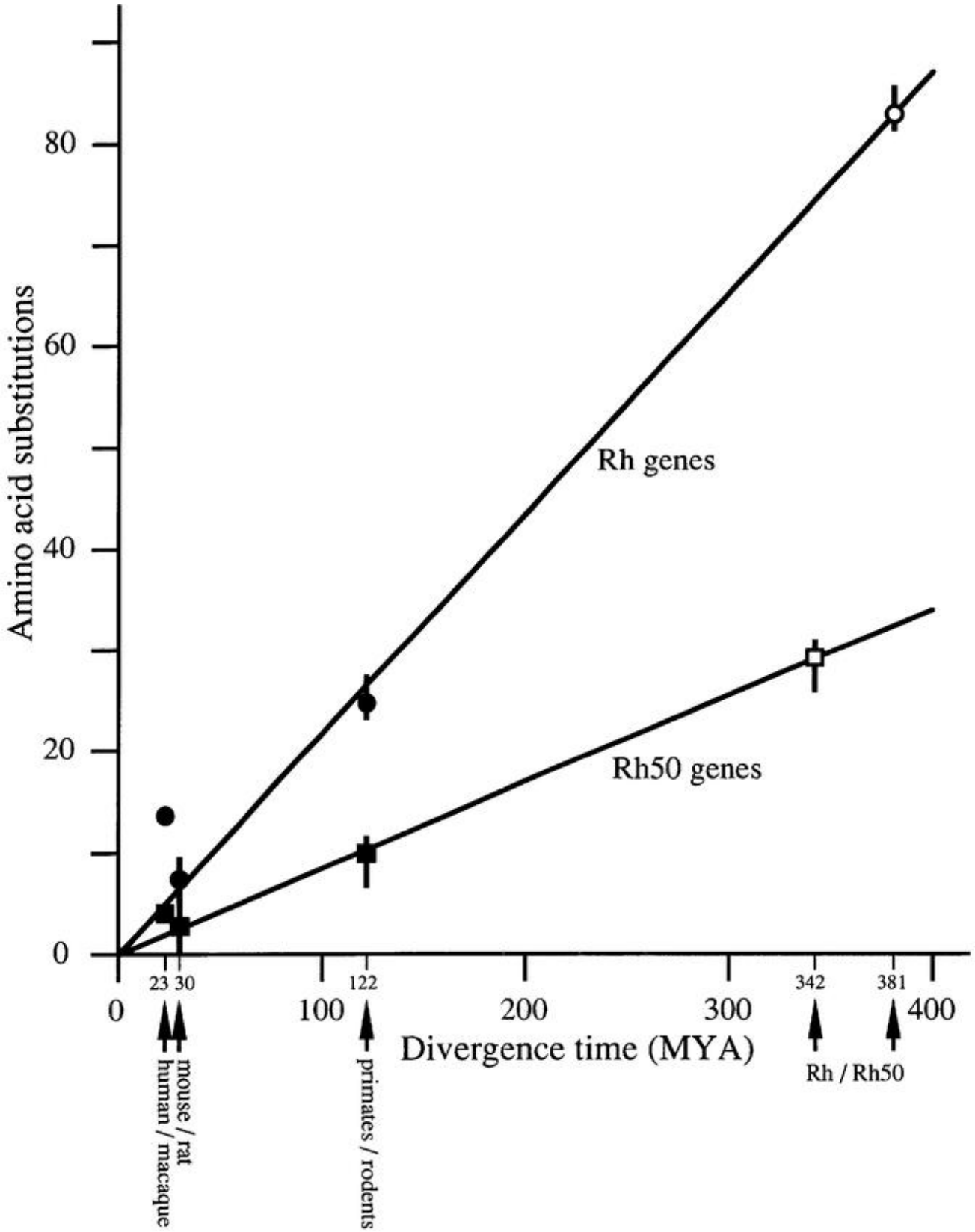

(A)
0.312
Human Rh50
24.722
Macaque Rh50
7.671
17.172
Human Rh
65.711
Macaque Rh
10.278
(B)
1.096
Mouse Rh50
24.049
Rat Rh50
4.463
4.053
Mouse Rh
70.809
Rat Rh
10.534
(C)
7.732
Human Rh50
16.895
Mouse Rh50
9.135
24.515
Human Rh
59.610
Mouse Rh
20.403
(D)
3.600
Human Rh50
6.682
4.527
Macaque Rh50
19.204
0.454
Mouse Rh50
6.353
Rat Rh50
5.118
13.583
Human Rh
10.448
Macaque Rh
13.737
58.110
Mouse Rh
5.422
17.862
Rat Rh
9.343

Figure 3.9 Comparisons between amino acid substitutions and divergence times from tree D (figure 3.8). Full circles and full squares show points for Rh and Rh50 genes, respectively. Divergence times between Rh (open circle) and Rh50 (open square) genes are estimated by using the regression through origin.


Amino acid substitutions
0
20
40
60
80
Rh genes
Rh50 genes
0
23
30
100
122
200
300
342
381
400
Divergence time (MYA)
mouse / rat
human / macaque
primates / rodents
Rh / Rh50

CHAPTER IV

LONG-TERM EVOLUTION OF THE RH BLOOD GROUP GENES AND THEIR RELATED GENES

PCR-direct sequencing of cDNA

African clawed frog (*Xenopus laevis*) and teleost fish Japanese medaka (*Oryzias latipes*) were bought from stores, "Half & Top" and "Home Assist", respectively, in Shizuoka prefecture. Total RNAs were extracted from femora of *Xenopus laevis* and from whole body of Japanese medaka, using ISOGEN (Nippon Gene). Reverse transcription was performed by using SuperScript™ II reverse transcriptase and oligo dT-adaptor primer.

Degenerate PCR for the Rh50 gene was performed and a partial product was obtained. I then performed 5' RACE (rapid amplification of the 5' cDNA ends) using 5'RACE System for Rapid Amplification of cDNA Ends version 2.0 (Gibco-BRL). 3'RACE is also curried out. To amplify the complete cDNA sequence, PCR was performed by using gene specific primers. PCR was performed in a 20 μl reaction containing 0.5-1 μl of the first-strand cDNA, 1×Gene Taq Universal Buffer (Mg^{2+} free) (Nippon Gene), 1.5 mM $MgCl_2$, 0.2 mM dNTP, 10 pmol of each primer (designed on sites of 5' and 3' ends), and 1 units of AmpliTaq Gold (Perkin-Elmer). Amplification was carried out in DNA GeneAmp PCR System 9700 (Perkin-Elmer) with the following temperature parameters: 10 min at 95℃ followed by 40 cycles of 95℃ for 30 sec, 65℃ for 15 sec, and 72℃ for 1 min. PCR products were purified using MicroSpin Columns S-300 HR (Pharmacia Biotech). DNA sequencing was performed on PCR products using Dye Terminator Cycle Sequencing Kit and ABI prism 377 DNA sequencer

(Perkin-Elmer). A progressive sequencing strategy was carried out with design of further primers to complete the sequence for coding region of both strands of the cDNA. Figure 4.1 shows the sequencing scheme for the Rh50-like gene of *Xenopus*. The Rh50-like gene of Japanese medaka was also determined by using the same procedure. Appendix III lists PCR primers used.

cDNA sequence of *Xenopus* Rh50-like gene

I sequenced the Rh50-like gene cDNA for *Xenopus*. The nucleotide sequence length of this gene is 1275 bp. This newly determined sequence was compared with Rh50 genes for human (Ridgwell et al. 1992), crab-eating macaque, mouse (*M. m. domesticus*), and rat. Figure 4.2 shows the multiple alignment of Rh50 genes (see also Chapter III). The location of the start codon of the *Xenopus* Rh50-like gene is identical to other Rh50 genes. The location of the stop codon of the *Xenopus* Rh50-like gene is almost similar to other Rh50 genes except for the human Rh50 gene. The *Xenopus* Rh50-like gene also has single 15 nucleotides repeat around positions 100-150 (see also Chapter III). Because all gaps were multiple of 3, they did not alter codon frames. Nucleotide sequence similarities between the *Xenopus* Rh50-like gene and other Rh50 genes are about 70 %. The GC contents of this genes was 45.87 %. This value was similar to other Rh50 genes (see table 3.1 in Chapter III).

cDNA and inferred protein sequence of Japanese medaka Rh50-like gene

I also sequenced the Japanese medaka (teleost fish) Rh50-like gene. Figure 4.3 shows its cDNA sequence and the inferred protein sequence of the Japanese medaka

Rh50-like gene. The nucleotide sequence length of this gene is 1467 bp. The cDNA sequence is numbered so that nucleotide 1 corresponds to the first codon position of the initiating methionine codon for the protein. The position of the initiating codon of this gene is different from those of known Rh and Rh50 genes. The stop codon (TAA) is located on nucleotide sites 1465-1467. The position of this gene is also different from those of known Rh and Rh50 genes. The GC contents of this genes was 48.93 %. This value was more similar to Rh50 genes than to Rh genes (see table 3.1 in Chapter III). Numbers of amino acid substitutions between the Japanese medaka Rh50-like gene and Rh genes, and between the Japanese medaka Rh50-like gene and Rh50 genes are 1.47-1.63 and 0.74-0.98, respectively, by using Kimura's (1983) method.

Amino acid sequence comparison of Rh, Rh50, and their related genes

Figure 4.4 shows the multiple alignment of amino acid sequences of Rh, Rh50, and their related genes. Two genes of *C. elegans* (Wilson et al. 1994) and an Rh-like gene of sponge (Seack et al. 1997) found by database searches by using BLAST (Altschul et al. 1990) were also included. CLUSTAL W version 1.6 (Thompson, Gibson, and Higgins 1994) was used for multiple alignment. The twelve predicted hydrophobic membrane-spanning regions are surrounded by boxes. The PredictProtein server (EMBL) was used for analyses of transmembrane helix location. These membrane-spanning regions did not include gaps and are relatively conserved.

The phylogenetic tree of Rh, Rh50, and their related genes

Because membrane-spanning regions did not include gaps and are relatively

conserved, I used only the 223 amino acid sites for membrane-spanning regions for tree construction. The program protml of MOLPHY version 2.2 (Adachi and Hasegawa 1994) and the computer package MEGA version 1.0 (Kumar, Tamura, and Nei 1993) were used for the maximum likelihood and the neighbor-joining analyses, respectively. Table 4.1 shows the results of the maximum likelihood and the neighbor-joining analyses for Rh, Rh50, and their related genes. Tree 1 showed the maximum likelihood value. The topology of tree 1 indicates that the *X enopus* Rh50-like gene forms a cluster with Rh50 genes and the Japanese medaka Rh50-like gene is located on the branch before the gene duplication between Rh and Rh50 genes (figure 4.5). Tree 2 showed the second maximum likelihood value. The topology of tree 2 indicates that the *X enopus* Rh50-like gene and the Japanese medaka Rh50-like gene form the cluster with Rh50 genes.

I also used the neighbor-joining method by using the gamma distance. Because we don't know the actual value of parameter "a", I used "a" values from 0.5 to 2.0. When the rate of amino acid substitution does not vary much from site to site (a = 1.7 - 2.0), tree 2 is chosen by using the neighbor-joining method. When the variation of the rate of amino acid substitution is intermediate (a = 0.8 - 1.6), tree 1 is chosen by using the neighbor-joining method. This is identical to the maximum likelihood tree. When the rate of amino acid substitution varies considerably (a = 0.5 - 0.6), tree 5 is chosen. When a = 0.7, Rh genes of mammals, Rh50 genes of mammals, and the *X enopus* Rh50-like gene showed the trichotomy (trees 1, 4, or 5 in table 4.1).

Figure 4.5A shows the maximum likelihood tree of Rh, Rh50, and their related genes by using twelve membrane-spanning regions. The root was located by assuming the Rh-like gene of sponge as an outgroup. Bootstrap values by using the neighbor-joining method with the gamma distance (a = 1.2: this is an average value for tree 1 in table 4.1) are shown on each branch. There are four clusters in this tree; Rh50 genes of mammals and the *X enopus* Rh50-like gene, Rh genes of mammals, the Rh50-like gene of Japanese medaka, and two genes of *C. elegans*.

Because the bootstrap value for the clustering of the *Xenopus* Rh50-like gene and mammalian Rh50 genes is not high, I also constructed the neighbor-joining tree of Rh blood group genes and their homologous genes by using all amino acid sites (figure 4.5B). Numbers of amino acid substitutions were estimated by using Kimura's (1983) method and the gaps were not used (see figure 4.4). This tree showed the same branching pattern with the maximum likelihood tree (figure 4.5A). I also analysed same data set by using the maximum likelihood method. Figure 4.6 shows top three maximum likelihood trees. The phylogenetic location of the medaka Rh50-like gene is not yet clear.

These tree suggests that the gene duplication (node D in figures 4.5A and 4.5B) of Rh and Rh50 genes occurred after speciation of fish and other vertebrates. The branch lengths of Rh50 genes is much shorter than those of Rh genes, indicating a lower evolutionary rate in the Rh50 gene than in the Rh gene. This pattern is consistent with the result of d_N in table 3.2 in Chapter III where all the coding regions were compared. It is interesting that after the gene duplication which produced Rh and Rh50 genes, the Rh gene lineage started to evolve more rapidly than the Rh50 lineage.

Comparison of numbers of amino acid substitutions

Numbers of amino acid substitutions are also estimated (table 4.2). The phylogenetic tree of figure 4.5 was used and single-lineage of amino acid substitution values were obtained applying Ishida et al.'s (1995) method. Because numbers of amino acid substitutions for Rh were consistently two - three times higher than those for Rh50, a rough molecular clock exists for both genes. This is consistent with the result of table 3.2 in Chapter III. Therefore, I estimated evolutionary rates of Rh and Rh50 genes by using the regression through origin. Divergence times between human and

macaque, between mouse and rat, between primates and rodents, and between mammals and amphibians were assumed to be 23 (Kumar and Hedges 1998), 30 (see Chapter III), 122 (Easteal, Collet, and Betty 1995), and 360 (Kumar and Hedges 1998) million years, respectively, and they were used for calibration of the molecular clock. Numbers of amino acid substitutions were thus obtained as 93.79 and 24.46 for Rh and Rh50 genes, respectively. If we use these rates, the time of gene duplication (node D in figure 4.5A) producing Rh and Rh50 genes was estimated to be about 448 or 479 million years ago from the data for Rh or Rh50, respectively. This period roughly corresponds to the early Paleozoic where before or after the divergence between tetrapods and teleost fish lineages.

The relationship to ammonium transporter proteins

Because products of Rh and Rh50 genes are predicted to have twelve transmembrane domains, it has been suggested that Rh and Rh50 proteins are related to ammonium transporter proteins (e. g., Marini et al. 1997b). I searched DDBJ amino acid sequence database (DAD) by using BLAST (Altschul et al. 1990) for finding homologous proteins. Two protein sequences of the human Rh50 gene and the Rh-like gene of sponge are used as query sequences. Table 4.3 shows the list of obtained protein sequences. Nineteen ammonium transporter proteins of organisms were found. No ammonium transporter gene was found by examining the complete genomic sequences of *Haemophilus infuluenzae* and *Mycobacterium genitalium*, two bacteria whose natural environment is human tissues (Marini et al. 1997a). Ammonium transporter proteins carry ammonium ion (NH_4^+) into cells (Marini et al. 1994). Ammonium ion is a nitrogen source supporting growth for primitive organisms such as yeast at an optimal rate.

Figure 4.7 shows the multiple alignment of trans-membrane domains for ammonium

transporter protein sequences compared with Rh, Rh50, and their related protein sequences. Because the first, second, third, and twelfth predicted membrane regions are not aligned well, these regions are not shown. This figure was made by using the BOXSHADE 3.21 server (http://www.isrec.isb-sib.ch/software/BOX_form.html). In this context, it may be interesting to note a similarity between the Rh blood group related proteins and ammonium transporter proteins.

Figure 4.8A shows the neighbor-joining tree of the Rh blood group related genes and ammonium transporter genes. This tree was constructed from the multiple alignment of figure 4.7 by using the gamma distance (a = 1.2: this is an average value for tree 1 in table 4.1). Bootstrap values greater than 90 % are shown on each branch. There are roughly three clusters in this tree; α group of ammonium transporter genes (amt α group), β group of ammonium transporter genes (amt β group), and the Rh blood group genes and their related genes (Rh genes group). Marini et al. (1997a) and Van Dommelen et al. (1998) showed phylogenetic trees for ammonium transporter proteins, and those trees also indicated two major groups for ammonium transporter proteins, but there are no appropriate names for those two groups. Therefore, I propose to call these two groups of ammonium transporter genes as α and β groups. Both α and β groups of ammonium transporter cluster include three domains of life (eubacteria, archaea, and eukaryota). Because most of the ammonium transporter genes are predicted from genome sequences, however, functional differences between gene products from α and β groups are not known.

I also performed the maximum likelihood method by using the JTT (Jones, Taylor, and Thornton 1992) model with data frequencies (the program protml of MOLPHY version 2.2 (Adachi and Hasegawa 1994) was used). For simplicity, I selected one sequence that has the shortest branch among each taxon of both amt α and amt β groups. The sequence of sponge was used as representative of the Rh blood genes and their related genes. Figure 4.8B shows top four maximum likelihood trees by the exhaustive

search. These likelihood values did not differ much with each other. All four trees indicated the same phylogenetic relationship for the amt α cluster, but differ on the relationship of amt β and Rh genes groups. Because the relationship among the Rh genes group and amt β group is not clear, it is difficult to define the divergence point of the Rh blood group related genes. In either case, however, it is suggested that the Rh blood group genes and their related genes have probably been existing as essential membrane proteins in many animal phyla.

The universal ancestor of all the organisms may already had an ammonium transporter gene, then the gene duplication occurred before divergences of the three domains (figure 4.9). Therefore most of the organisms may have amt α and/or amt β groups. For example, *Arabidopsis thaliana* has both amt α and amt β genes, while *Saccharomyces cerevisiae* has only amt α group genes. The Rh genes group is more similar to the amt β group than the amt α group. It suggests that after the gene duplication between amt α and amt β groups the Rh genes group diverged from the amt β group, or before the gene duplication of amt α and amt β groups the Rh genes group diverged and counterpart genes for Rh genes group in amt α group are not yet determined or were lost from the genomes. Because no Rh gene group counterpart in amt α group organisms was found by examining the complete genomic sequences (such as *Bacillus subtilis*, *Escherichia coli*, and *Saccharomyces cerevisiae*), the former is more plausible. *C. elegans* has both genes of amt β and the Rh genes groups, and there are possibilities that other animals also have amt β genes and plants have Rh genes.

Expressions of Rh and Rh50 genes are thought to restrict on erythrocyte, and erythrocytes are thought to developed on the vertebrate lineage. Therefore the location of expressions are thought to change from some other cell types to erythrocyte membrane before the gene duplication between Rh and Rh50 genes. Then the gene duplication between Rh and Rh50 genes occurred in the early Paleozoic, and the evolutionary rate of Rh50 genes are conserved because of selective constraint compared to Rh genes. In the

hominoid lineage, Rh genes are duplicated and might be released from selective constraint or are positively selected.

Table 4.1

The maximum liklihood and the neighbor-joining by gamma distance analyses of various kind of topologies for the Rh blood group related genes

Tree	Topology*	ΔlnL	NJ
1	(((((hR,cR),(mR,rR)),(((h50,c50),(m50,r50)),x50)),jm),(ce1,ce2),s)	0.0	0.7-1.6
2	((((hR,cR),(mR,rR)),((((h50,c50),(m50,r50)),x50),jm)),(ce1,ce2),s)	-0.1	1.7-2.0
3	(((((hR,cR),(mR,rR)),jm),(((h50,c50),(m50,r50)),x50)),(ce1,ce2),s)	-2.1	-
4	((((((hR,cR),(mR,rR)),((h50,c50),(m50,r50))),x50),jm),(ce1,ce2),s)	-2.9	0.7
5	((((((hR,cR),(mR,rR)),x50),((h50,c50),(m50,r50))),jm),(ce1,ce2),s)	-2.9	0.5-0.7
6	((((((hR,cR),(mR,rR)),jm),((h50,c50),(m50,r50))),x50),(ce1,ce2),s)	-12.4	-
7	(((((hR,cR),(mR,rR)),(((h50,c50),(m50,r50)),jm)),x50),(ce1,ce2),s)	-13.3	-
8	((((((hR,cR),(mR,rR)),x50),jm),((h50,c50),(m50,r50))),(ce1,ce2),s)	-14.5	-
9	(((((hR,cR),(mR,rR)),x50),(((h50,c50),(m50,r50)),jm)),(ce1,ce2),s)	-15.0	-

ΔlnL : The difference in log-likelihood from the ML tree (-3461.7) by using the JTT (Jones, Taylor, and Thornton 1992) model with data frequencies.

Numbers in the column of NJ mean parameters a in the gamma distance for the neighbor-joining method.

*Gene abbreviations are hR: human RhcE, cR: crab-eating macaque Rh, mR: mouse Rh, rR rat Rh, h50: human Rh50, c50: crab-eating macaque Rh50, m50: mouse Rh50, r50: rat Rh50, x50: *Xenopus* Rh50-like, jm: Japanese medaka Rh50-like, ce1: *C. elegans* 1, ce2: *C. elegans* 2, s: sponge Rh-like.

Table 4.2

Numbers of amino acid substitutions and divergence times

Diverging node	human/macaque	mouse/rat	primates/rodents	mammals/amphibians	Rh/Rh50
Rh[a]	11.87	7.33	23.97		93.79
Rh50[a]	4.07	2.76	10.17	16.77	24.46
MYA	23	30	122	360	448-479[b]

[a] Amino acid substituitons of single lineage based on phylogenetic tree of fugure 4.5A.

[b] Estimated from numbers of amino acid substitutions.

Table 4.3

The list of ammonium transporter proteins

Accessions	Organisms	Taxa
AE000674	*Aquifex aeolicus*	Bacteria
ACJ225126	*Azorhizobium caulinodans*	Bacteria
AF005275	*Azospirillum brasilense*	Bacteria
L03216	*Bacillus subtilis*	Bacteria
U82664	*Escherichia coli*	Bacteria
D90901	*Synechocystis* sp.	Bacteria
AE001036	*Archaeoglobus fulgidus*	Archaea
AE000846	*Methanobacterium thermoautotrophicum*	Archaea
U67463	*Methanococcus jannaschii*	Archaea
U67574	*Methanococcus jannaschii*	Archaea
AC003028	*Arabidopsis thaliana*	Eukaryota
X75879	*Arabidopsis thaliana*	Eukaryota
U53338 (C05E11.4)	*Caenorhabditis elegans*	Eukaryota
U53338 (C05E11.5)	*Caenorhabditis elegans*	Eukaryota
X95098	*Lycopersicon esculentum*	Eukaryota
AF001505	*Oryza sativa*	Eukaryota
Z72906	*Saccharomyces cerevisiae* (MEP1)	Eukaryota
Z71418	*Saccharomyces cerevisiae* (MEP2)	Eukaryota
U40829	*Saccharomyces cerevisiae* (MEP3)	Eukaryota

Figure 4.1 The sequencing scheme for the Rh50-like gene of *Xenopus laevis*. Boxes and arrows show PCR products and sequensing, respecitively. Primers are mentioned in Appendix III.

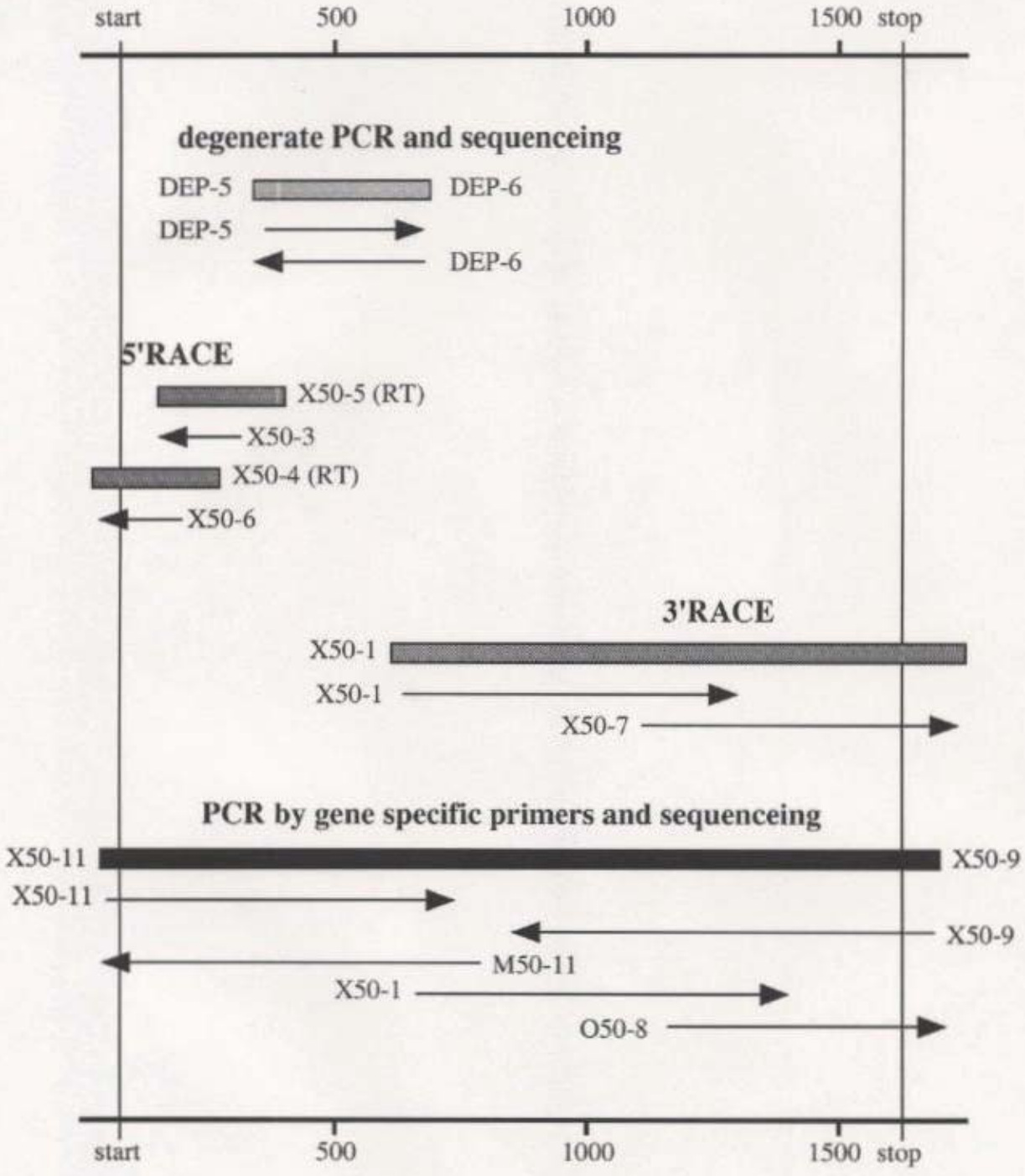

start
500
1000
1500
stop
degenerate PCR and sequenceing
DEP-5
DEP-6
DEP-5
DEP-6
5'RACE
X50-5 (RT)
X50-3
X50-4 (RT)
X50-6
3'RACE
X50-1
X50-1
X50-7
PCR by gene specific primers and sequenceing
X50-11
X50-9
X50-11
X50-9
M50-11
X50-1
O50-8
start
500
1000
1500
stop

Figure 4.2 The multiple alignment of nucleotide sequences of Rh50 genes for human, crab-eating macaque, mouse, rat, and the *Xenopus* Rh50-like gene. Gaps are denoted by hyphens, and only nucleotides different from those of the human sequence are shown.

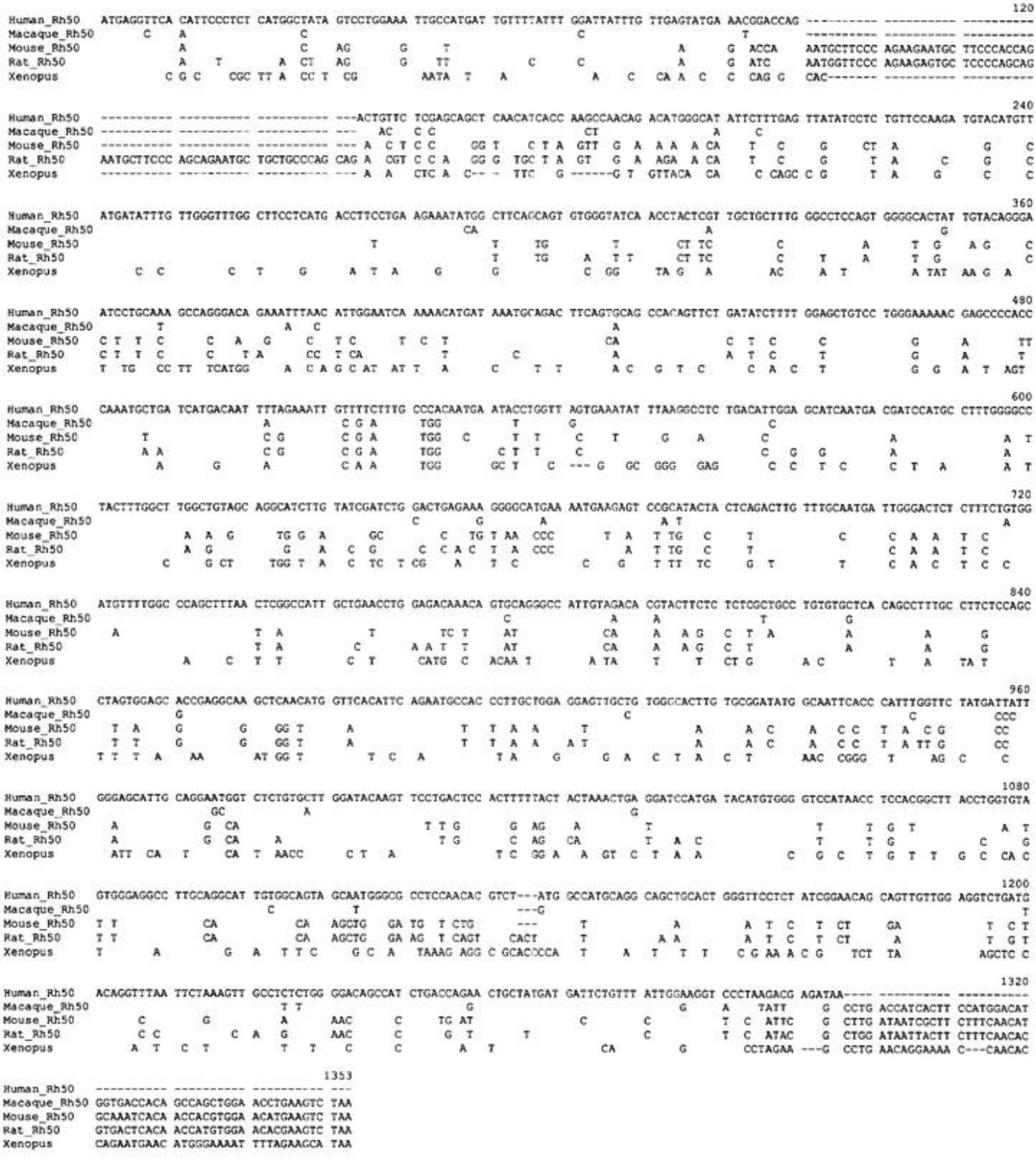


Figure 4.3 The cDNA sequence and inferred protein sequence of the Japanese medaka Rh50-like gene. The cDNA sequence is numbered so that nucleotide 1 corresponds with the initiating methionine codon for the protein. The stop codon (TAA) is shown by the asterisk.

```
   1 ATGGGCAACTGCTGTGAGAGCGCGTCCAACTTCTTTGGGCCCCAGAAGAACACAAACGTTCGTGTCAGCCTGCCTGCCGTCTGCTTCGTC   90
     M  G  N  C  C  E  S  A  S  N  F  F  G  P  Q  K  N  T  N  V  R  V  S  L  P  A  V  C  F  V
                               10                                  20                                  30

  91 TGGCAGATTGCTATGATTGTGCTGTTTGGGGTCTTCATCAGGTACGACGAGGAATCAGATGCTCACTGGGTAGAGTTAAAAAAGACTGAG  180
     W  Q  I  A  M  I  V  L  F  G  V  F  I  R  Y  D  E  E  S  D  A  H  W  V  E  L  K  K  T  E
                               40                                  50                                  60

 181 AACCTCACAGACCTCCAAAATGAATTCTACTTCAGATATCCAAGCTTCCAGGATGTCCACGTCATGATCTTTGTTGGCTTTGGTTTCCTC  270
     N  L  T  D  L  Q  N  E  F  Y  F  R  Y  P  S  F  Q  D  V  H  V  M  I  F  V  G  F  G  F  L
                               70                                  80                                  90

 271 ATGACGTTCCTAAAACGTTACAGCTTCAGTGCTGTGGGCTTTAACTTCCTGATCGCTGCCTTTGGCCTGCAGTGGGCTCTCCTCATGCAG  360
     M  T  F  L  K  R  Y  S  F  S  A  V  G  F  N  F  L  I  A  A  F  G  L  Q  W  A  L  L  M  Q
                              100                                 110                                 120

 361 GGCTGGTTCCACCACTTCGACTACTCTACTGGAAAAATCTACATAGGAATTGAAAGTTTGATAAATGCAGACTTCTGCTGTGCTGCCTCT  450
     G  W  F  H  H  F  D  Y  S  T  G  K  I  Y  I  G  I  E  S  L  I  N  A  D  F  C  C  A  A  S
                              130                                 140                                 150

 451 CTGATCGCCTATGGAGCCATCCTGGGTAAAGTCAGCCCTGTGCAGCTGATGGTTGTCACCTTGTTTGGTGTCACTCTGTTTGCTGTGGAG  540
     L  I  A  Y  G  A  I  L  G  K  V  S  P  V  Q  L  M  V  V  T  L  F  G  V  T  L  F  A  V  E
                              160                                 170                                 180

 541 GAGTATATCATCCTAGATCTCCTTCATTGCAGAGATTCTGGTGGCGCCATGGTCATTCACTGCTTTGGAGGCTACTATGGTTTGGCCATA  630
     E  Y  I  I  L  D  L  L  H  C  R  D  S  G  G  A  M  V  I  H  C  F  G  G  Y  Y  G  L  A  I
                              190                                 200                                 210

 631 TCCTGGGTGCTTTACCGACCAAATCTACATAGAAGTAAACGACTCAATGGATCCGTTTACCACTCTGATCTTTTTGCAATGATTGGCACA  720
     S  W  V  L  Y  R  P  N  L  H  R  S  K  R  L  N  G  S  V  Y  H  S  D  L  F  A  M  I  G  T
                              220                                 230                                 240

 721 TTGTTCCTGTGGATGTTCTGGCCCAGTTTCAATTCGGCCATCGCAAACCACGGCGATGGGCAGCACAGGACTGCAATGAACACCTACATC  810
     L  F  L  W  M  F  W  P  S  F  N  S  A  I  A  N  H  G  D  G  Q  H  R  T  A  M  N  T  Y  I
                              250                                 260                                 270

 811 GCTCTGGCTTCTTCTGTGCTCACTACTGTTGCCCTCTCAAGCATGTCCAAGAAGGAAGGAAAACTGGACATGGTACATATCCAGAATGCC  900
     A  L  A  S  S  V  L  T  T  V  A  L  S  S  M  S  K  K  E  G  K  L  D  M  V  H  I  Q  N  A
                              280                                 290                                 300

 901 ACTCTGGCAGGTGGTGTTGCCATGGGAACAGCAGCAGAGTTTATGATCACTCCTTACGGTTCGCTCATTGTGGGATTTTGCATCGGCATC  990
     T  L  A  G  G  V  A  M  G  T  A  A  E  F  M  I  T  P  Y  G  S  L  I  V  G  F  C  I  G  I
                              310                                 320                                 330

 991 ATCTCTACTTTTGGCTATTTGTACGTCACGCCTTTCTTAGAGAAGCGATTGAAGCTGCAGGATACATGTGGCATCCATAACCTGCATGCA 1080
     I  S  T  F  G  Y  L  Y  V  T  P  F  L  E  K  R  L  K  L  Q  D  T  C  G  I  H  N  L  H  A
                              340                                 350                                 360

1081 GTACCAGGCATGCTCGGTGGCTTCATAGGTGCCATCGTTGCAGCAACAGCAAGTGAATCGGTCTACAGCAAACAGGGGCTGATCGACACA 1170
     V  P  G  M  L  G  G  F  I  G  A  I  V  A  A  T  A  S  E  S  V  Y  S  K  Q  G  L  I  D  T
                              370                                 380                                 390

1171 TTTGGTTTTACTGGAAAGTACGAAAACAGATCACCGGGAACGCAGGGAGGCTATCAGGCTGCAGGAGTGTGCGTGGCCATGGCATTTGGG 1260
     F  G  F  T  G  K  Y  E  N  R  S  P  G  T  Q  G  G  Y  Q  A  A  G  V  C  V  A  M  A  F  G
                              400                                 410                                 420

1261 CTTGTTGGAGGAGCTATTGTTGGTTTCATCCTGAAGTTCCCAATCTGGGGCGATGCTGCTGATGACTACTGCTTTGATGATGAAGCCTAC 1350
     L  V  G  G  A  I  V  G  F  I  L  K  F  P  I  W  G  D  A  A  D  D  Y  C  F  D  D  E  A  Y
                              430                                 440                                 450

1351 TGGGAGCTTCCTGAAGAGGAAGAGACCATTCCTCCTGTCTTGGAGTACAACAATCACATGACACAACAAAAGCACCAGGAAACACCTGAG 1440
     W  E  L  P  E  E  E  E  T  I  P  P  V  L  E  Y  N  N  H  M  T  Q  Q  K  H  Q  E  T  P  E
                              460                                 470                                 480

1441 ACAAGCTTCTCTGTGGTAGAAAGCTAA                                                                   1467
     T  S  F  S  V  V  E  S  *
```

Figure 4.4 The multiple alignment of amino acid sequences of Rh blood group genes and their related genes. Amino acid sequence of Rh-like genes for *C. elegans* 1 (Z74026-B0240.1), *C. elegans* 2 (U64847-F08F3.3), and sponge are also used. Twelve membrane-spanning regions are shown by boxes.

Human_RhcE
Macaque_Rh
Mouse_Rh
Rat_Rh
Human_Rh50
Macaque_Rh50
Mouse_Rh50
Rat_Rh50
Xenopus
Medaka
C. elegans1
C. elegans2
Sponge

Figure 4.5 (A) The maximum likelihood tree of Rh blood group genes and their homologous genes by using twelve membrane-spanning regions. Bootstrap values by using the neighbor-joining method with the gamma distance (a = 1.2) are shown on each branch. (B) The neighbor-joining tree of Rh blood group genes and their homologous genes by using all regions without gap sites. Numbers of amino acid substitutions were estimated by using Kimura's (1983) method. Bootstrap values are shown on each branch. The diamond (node D) means the gene duplication between Rh and Rh50 genes in both trees. The root was located by assuming the Rh-like protein of sponge as an outgroup in both trees.

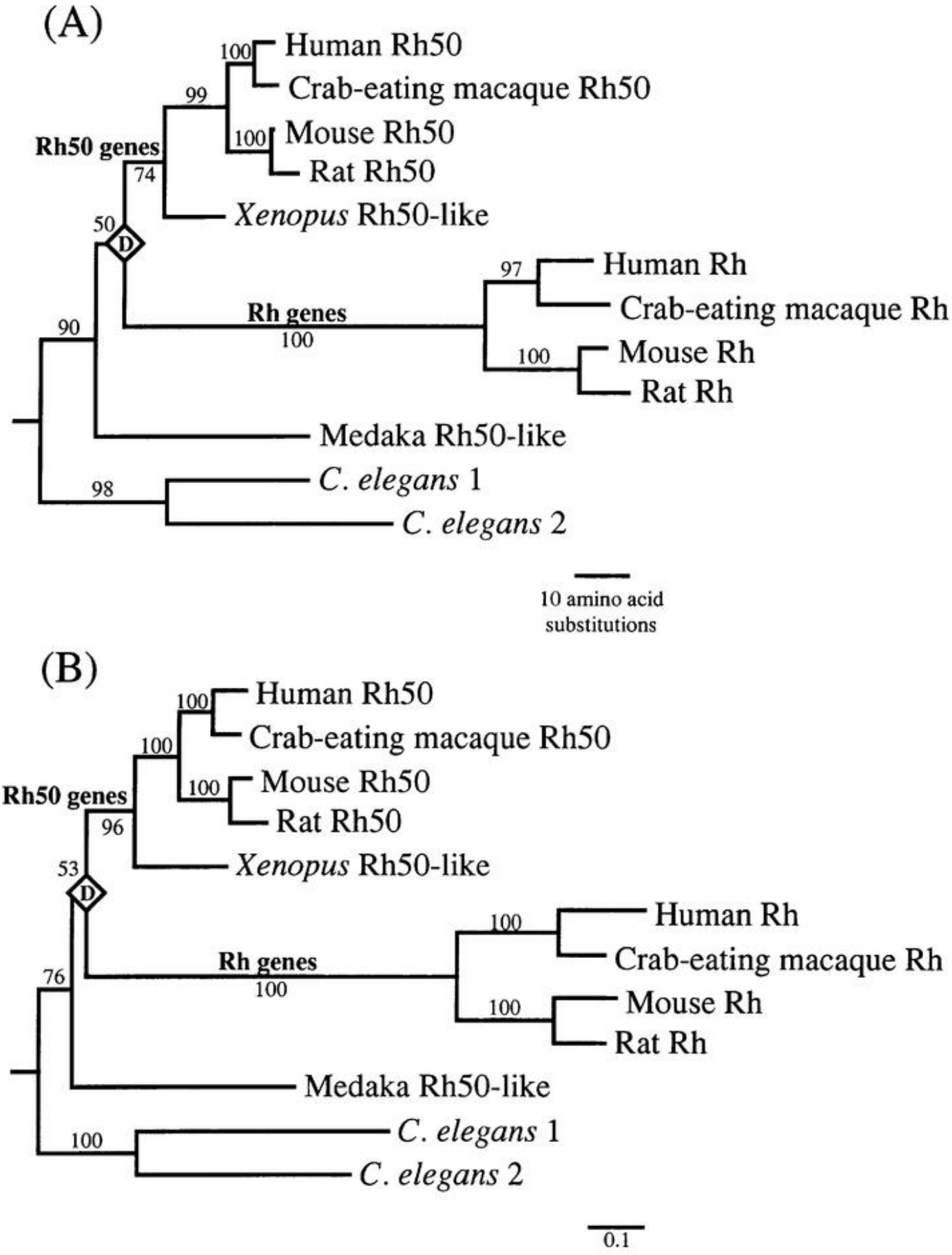
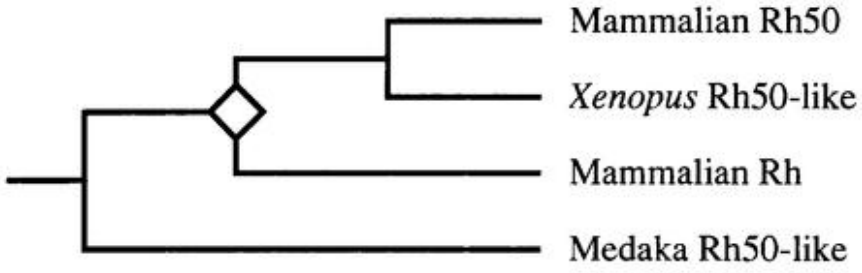
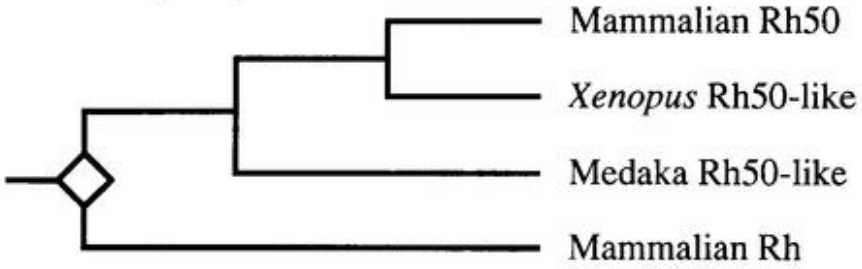

(A)
Human Rh50
Crab-eating macaque Rh50
Mouse Rh50
Rat Rh50
Xenopus Rh50-like
Rh50 genes
Human Rh
Crab-eating macaque Rh
Mouse Rh
Rat Rh
Rh genes
Medaka Rh50-like
C. elegans 1
C. elegans 2
10 amino acid substitutions
(B)
Human Rh50
Crab-eating macaque Rh50
Mouse Rh50
Rat Rh50
Xenopus Rh50-like
Rh50 genes
Human Rh
Crab-eating macaque Rh
Mouse Rh
Rat Rh
Rh genes
Medaka Rh50-like
C. elegans 1
C. elegans 2
0.1

Figure 4.6 Three possible tree topologies of Rh blood group genes and their homologous genes by using all amino acid sites. The program protml of MOLPHY version 2.2 (Adachi and Hasegawa 1994) was used. Likelihood differences from the ML (-9138.9) are shown in parentheses.

Tree 1 (-3.1)



Tree 2 (ML)



Tree 3 (-5.7)

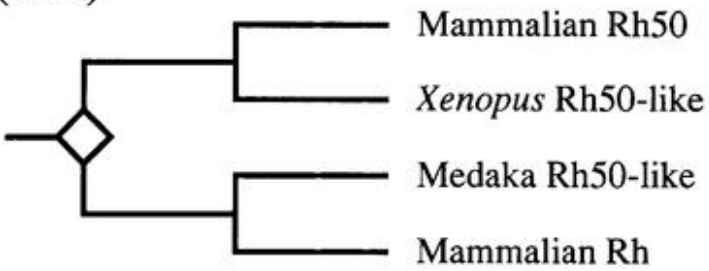


Figure 4.7 The multiple alignment of trans-membrane domains for ammonium transporter proteins compared with Rh, Rh50, and their related proteins. Equal signs surrounded with angled brackets designate trans-membrane domains.

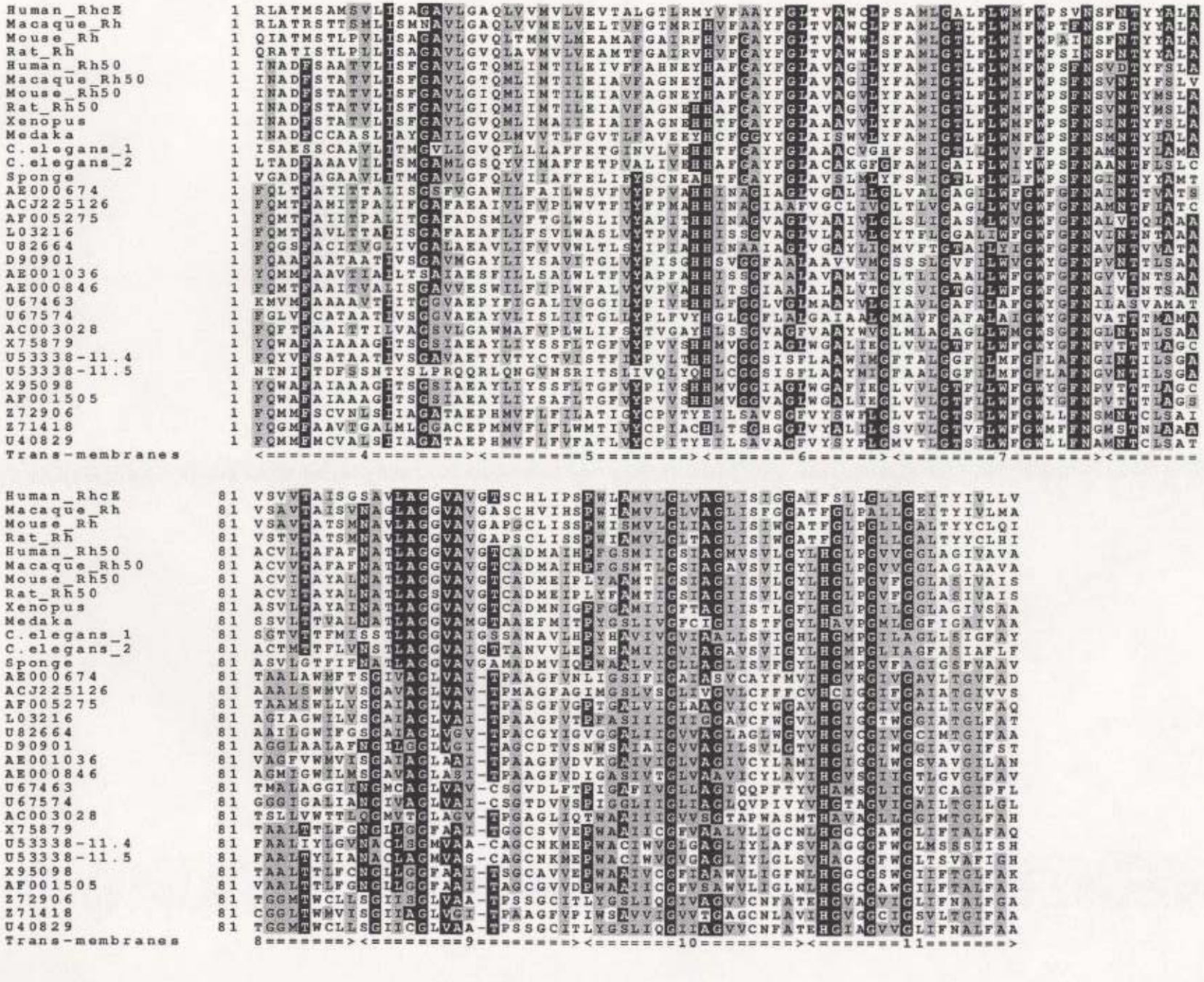
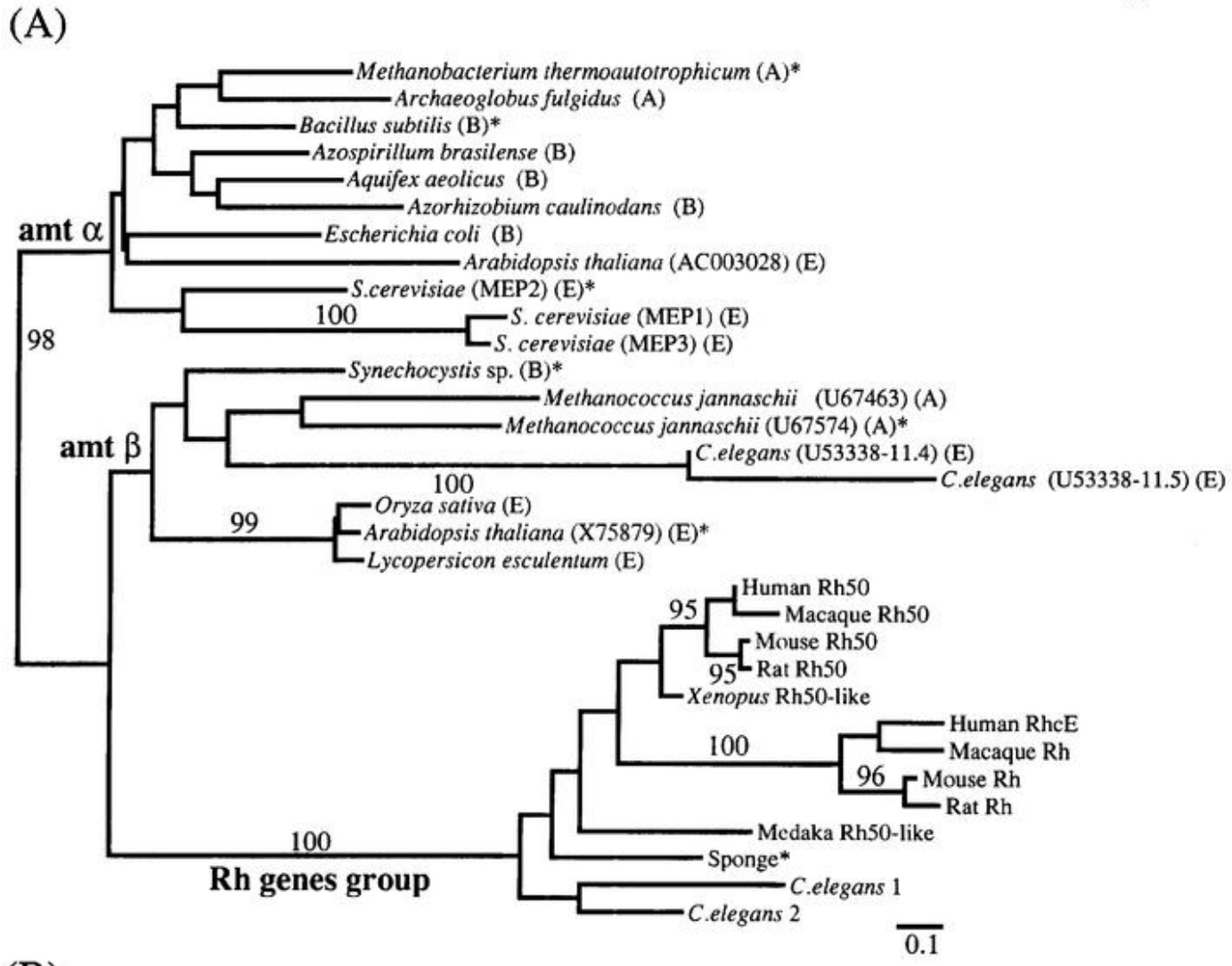

Human_RhcE
Macaque_Rh
Mouse_Rh
Rat_Rh
Human_Rh50
Macaque_Rh50
Mouse_Rh50
Rat_Rh50
Xenopus
Medaka
C.elegans_1
C.elegans_2
Sponge
AE000674
ACJ225126
AF005275
L03216
U82664
D90901
AE001036
AE000846
U67463
U67574
AC003028
X75879
U53338-11.4
U53338-11.5
X95098
AF001505
Z72906
Z71418
U40829
Trans-membranes

Figure 4.8 (A) The phylogenetic tree of ammonium transporter genes and the Rh blood group related genes. Bootstrap values greater than 90 % are shown on each branch. B, A, and E in parentheses mean taxa for bacteria, archaea, and eukaryota, respectively. Asterisks mean representative sequences of each group and are used for the maximum likelihood method. (B) Top four maximum likelihood trees used representative sequences. Likelihood differences from the ML (-1954.4) are shown in parentheses. A scale bar of each tree mean 10 amino acid substitutions.


(A)
Methanobacterium thermoautotrophicum (A)*
Archaeoglobus fulgidus (A)
Bacillus subtilis (B)*
Azospirillum brasilense (B)
Aquifex aeolicus (B)
Azorhizobium caulinodans (B)
amt α
Escherichia coli (B)
Arabidopsis thaliana (AC003028) (E)
S.cerevisiae (MEP2) (E)*
100
S. cerevisiae (MEP1) (E)
S. cerevisiae (MEP3) (E)
98
Synechocystis sp. (B)*
Methanococcus jannaschii (U67463) (A)
Methanococcus jannaschii (U67574) (A)*
amt β
C.elegans (U53338-11.4) (E)
C.elegans (U53338-11.5) (E)
100
Oryza sativa (E)
99
Arabidopsis thaliana (X75879) (E)*
Lycopersicon esculentum (E)
Human Rh50
95
Macaque Rh50
Mouse Rh50
95
Rat Rh50
Xenopus Rh50-like
Human RhcE
100
Macaque Rh
96
Mouse Rh
Rat Rh
Medaka Rh50-like
100
Sponge*
Rh genes group
C.elegans 1
C.elegans 2
0.1

(B)

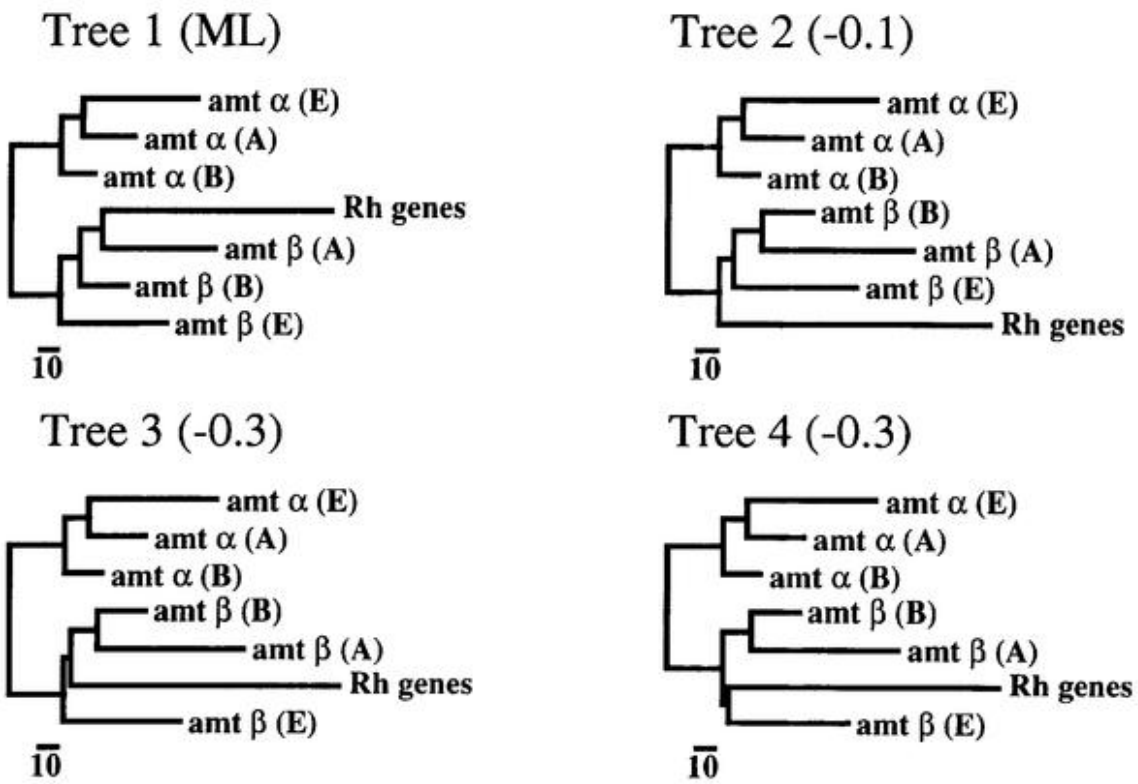
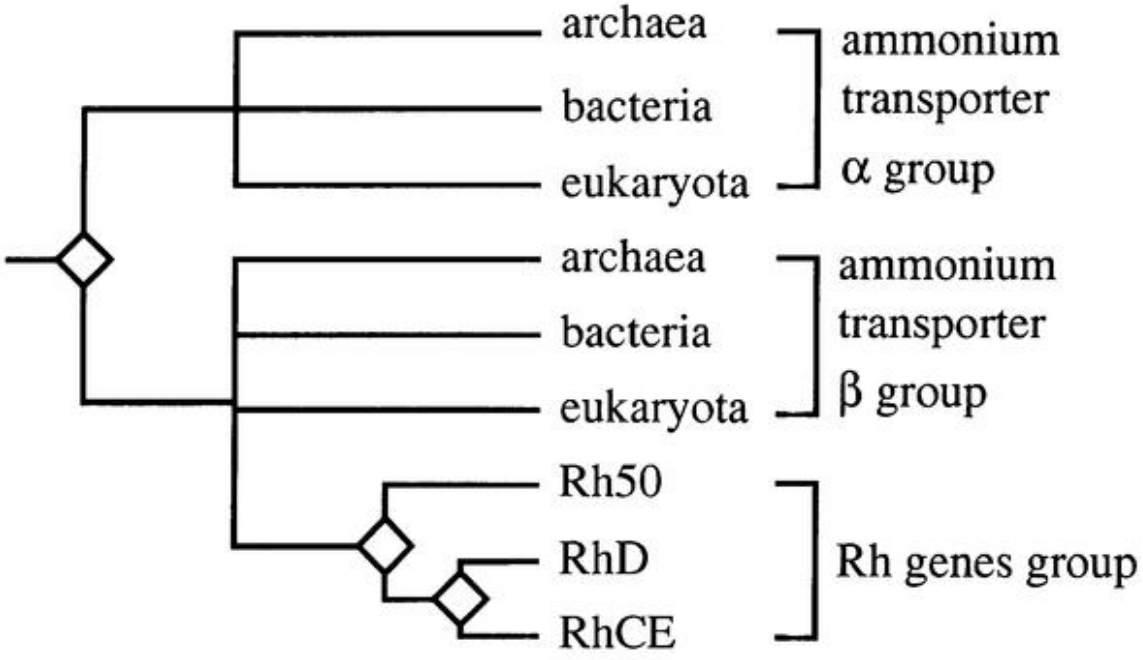

Tree 1 (ML)
amt α (E)
amt α (A)
amt α (B)
Rh genes
amt β (A)
amt β (B)
amt β (E)
10
Tree 2 (-0.1)
amt α (E)
amt α (A)
amt α (B)
amt β (B)
amt β (A)
amt β (E)
Rh genes
10
Tree 3 (-0.3)
amt α (E)
amt α (A)
amt α (B)
amt β (B)
amt β (A)
Rh genes
amt β (E)
10
Tree 4 (-0.3)
amt α (E)
amt α (A)
amt α (B)
amt β (B)
amt β (A)
Rh genes
amt β (E)
10

Figure 4.9 The scheme for evolution of ammonium transporter genes and the Rh blood group related genes. This tree was deduced from trees of figures 4.8A and 4.8B. Diamond mean gene duplications.


archaea
bacteria
eukaryota
ammonium
transporter
α group
archaea
bacteria
eukaryota
ammonium
transporter
β group
Rh50
RhD
RhCE
Rh genes group

CHAPTER V

CONCLUSIONS AND PROSPECTS

In this study, I examined the evolution of the Rh blood group genes of primates. Because I don't know the actual gene tree topology of primate Rh blood group genes, I assumed two plausible trees from nucleotide sequence data. Whichever I selected gene trees, I could find some gene conversion events on primate Rh blood group genes by using the modified site by site reconstruction method. I also showed the possibility of positive selection on the primate lineage of Rh blood group genes by using a statistical test. In any case, I should be very careful when we analyse the evolutionary history of tandemly duplicated genes, for there is always possibility of gene conversions.

To compare evolutionary patterns of primate Rh blood group genes with other mammals, I determined complete coding regions of Rh blood group genes of five mouse subspecies and rat, and Rh50 genes of five mouse subspecies, rat, and crab-eating macaque, and examined these genes. Phylogenetic analyses of Rh and Rh50 amino acid sequences indicate that the Rh50 gene has been evolving about two times more slowly than the Rh blood group gene both in primates and rodents. This conservative nature of the Rh50 gene suggests its relative importance to the Rh blood group gene. We published this study (a part of Chapter III) in the journal "Biochemical and Biophysical Research Communications" (Kitano et al. 1998).

From the comparison of synonymous substitutions between Rh and Rh50 genes, it is suggested that the mutation rate in rodents is about three times higher than that in primates. To examine effects of numbers of OTU for estimation of the divergence time, I reconstructed four trees for Rh and Rh50 genes of primates and rodents by using the maximum likelihood method and estimated the divergence time between Rh and Rh50 genes. It is suggested that numbers of calibration points for the molecular clock are

needed to obtain better estimation of divergence times. In any case, we should be very careful when we estimate the divergence time of genes.

Because the Rh blood group gene products are membrane proteins, these products of blood group genes seemed to be affected by interactions with other organisms or cells on surface regions, and there is a possibility of positive Darwinian selection. Endo, Ikeo, and Gojobori (1996) searched the nucleotide sequence database and found that 17 gene groups were the candidates for the genes on which positive selection may operate. Nine of those 17 gene groups were surface antigens of parasites or viruses. Eder and Spitalnik (1997) suggested that blood group antigens such as ABO, MN, and Lewis play a key role in pathogenesis of diseases. In fact, Saitou and Yamamoto (1997) found the evidence of positive selection in the ABO blood group genes of primates. The high rate of nonsynonymous substitutions for the primate Rh blood group suggests the existence of positive selection also on this gene, and this might be caused by some kind of interaction with pathogens. Comparison of synonymous and nonsynonymous substitutions for the Rh50 gene also revealed a possibility of existence of positive selection for this gene in primates. Because primates showed more clear sign of positive selection than rodents both for Rh and Rh50 genes, it is possible that the pattern of host-parasite interaction is different between primates and rodents.

I also determined the Rh50-like genes of *Xenopus* and Japanese medaka and analysed the long-term evolution of Rh, Rh50, and their related genes. The time of gene duplication that produced the Rh and Rh50 genes was estimated to be about 450-480 million years ago. This period roughly corresponds to the early Paleozoic just before or after the divergence between tetrapods and teleost fish lineages. From database searches, it is suggested that the Rh blood group genes and their related genes are related to ammonium transporter genes of many organisms, especially trans-membrane domains. The phylogenetic tree for ammonium transporter proteins indicated two major groups for ammonium transporter proteins. I propose to call these two groups of ammonium

transporter genes as α and β groups, and the Rh genes group is more similar to the amt β group than the amt α group. It is suggested that the Rh blood group genes and their related genes have probably been existing as essential membrane proteins in many animal phyla.

In this study, I elucidated the tempo and mode of evolution of the Rh blood group genes and their related genes. Several gene duplication events were observed in their evolutionary pathway. It is suggested that gene duplication events play an important role of evolution in genes. I think some new approaches used in this study (such as deduction of orthologous genes from phylogenetic networks, the site by site reconstruction method by using the maximum likelihood method, and estimations of divergence times by using single-lineages of a tree) are also useful for examination of other genes that experienced gene duplication events.

LITERATURE CITED

Adachi, J., and M. Hasegawa. 1994. MOLPHY. Programs for molecular phylogenetics, version 2.2. Institute of Statistical Mathematics, Tokyo.

Altschul, S. F., W. Gish, W. Miller, E.W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**: 403-410.

Arce, M., E. S. Thompson, S. Wagner, K. E. Coyne, B. A. Ferdman, and D. M. Lublin. 1993. Molecular cloning of Rh D cDNA derived from a gene present in Rh D positive but not Rh D-negative individuals. Blood **82**: 651-655.

Avent, N. D., S. K. Butcher, W. Liu, W. J. Mawby, G. Mallinson, S. F. Parsons, D. J. Anstee, and M. J. Tanner. 1992. Localization of the C termini of the Rh (rhesus) polypeptides to the cytoplasmic face of the human erythrocyte membrane. J. Biol. Chem. **267**: 15134-15139.

Avent, N. D., K. Ridgwell, M. J. Tanner, and D. J. Anstee. 1990. cDNA cloning of a 30 kDa erythrocyte membrane protein associated with Rh (Rhesus)-blood-group-antigen expression. Biochem. J. **271**: 821-825.

Baltimore, D. 1981. Gene conversion: some implication for immunoglobulin genes. Cell **24**: 592-594.

Bandelt, H.-J. 1994. Phylogenetic networks. Verh. Naturewiss. Ver. Hamburg (NF) **34**: 51-71.

Beckers, E. A., B. H. Faas, P. Ligthart, S. Simsek, M. A. Overbeeke, A. E. von dem Borne, D. J. van Rhenen, and C. E. van der Schoot. 1996. Characterization of the hybrid RHD gene leading to the partial D category IIIc phenotype. Transfusion **36**: 567-574.

Blancher, A., P. Calvas, and J. Ruffie. 1992. Etude des equivalents des antigenes Rhesus chez les primates non hominiens. CR. Soc. Biol. **186**: 682-695.

Blancher, A., and W. W. Socha. 1997. The Rhesus system. Pp. 147-218 *in* A.

Blancher, J. Klein, and W. W. Socha, eds. Molecular biology and evolution of blood group and MHC antigens in primates. Springer.

Carritt, B., T. J. Kemp, and M. Poulter. 1997. Evolution of the human RH (rhesus) blood group genes: a 50 year old prediction (partially) fulfilled. Hum. Mol. Genet. **6**: 843-850.

Cherif-Zahar, B., C. Bloy, C. Le Van Kim, D. Blanchard, P. Bailly, P. Hermand, C. Salmon, J.-P. Cartron, and Y. Colin. 1990. Molecular cloning and protein structure of a human blood group Rh polypeptide. Proc. Natl. Acad. Sci. U.S.A. **87**: 6243-6247.

Cherif-Zahar, B., M. G. Mattei, C. Le Van Kim, P. Bailly, J. -P. Cartron, and Y. Colin. 1991. Localization of the human Rh blood group gene structure to chromosome region 1p34.3-1p36.1 by in situ hybridization. Hum. Genet. **86**: 398-400.

Cherif-Zahar, B., V. Raynal, A. M. D'Ambrosio, J. -P. Cartron, and Y. Colin. 1994. Molecular analysis of the structure and expression of the RH locus in individuals with D--, Dc-, and DCw- gene complexes. Blood **84**: 4354-4360.

Cherif-Zahar, B., V. Raynal, P. Gane, M. G. Mattei, P. Bailly, B. Gibbs, Y. Colin, and J. -P. Cartron. 1996. Candidate gene acting as a suppressor of the RH locus in most cases of Rh-deficiency. Nature Genet. **12**: 168-173.

Easteal, S., C. Collet, and D. Betty. 1995. The mammalian molecular clock. Springer-Verlag, Heidelberg.

Eder, A. F., and S. L. Spitalnik. 1997. Blood group antigens as receptors for pathogens. Pp. 268-304 *in* A. Blancher, J. Klein, and W. W. Socha, eds. Molecular biology and evolution of blood group and MHC antigens in primates. Springer.

Endo, T., K. Ikeo, and T. Gojobori. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**: 685-690.

Eyers, S. A. C., K. Ridgwell, W. J. Mawby, and M. J. A. Tanner. 1994.

Topology and organization of human Rh (rhesus) blood group-related polypeptides. J. Biol. Chem. **269**: 6417-6423.

Felsenstein, J. 1981. Evolutionary tree from DNA sequence: A maximum likelihood approach. J. Mol. Biol. **17**: 368-376.

Fisher, R. A. cited by R. R. Race. 1944. An 'incomplete' antibody in human serum. Nature **153**: 771-772.

Fitch, W. M. 1977. On the problem of discovering the most parsimonious tree. Am. Nat. **111**: 223-257.

Gahmberg, C. G. 1982. Molecular identification of the human Rho (D) antigen. FEBS Lett. **140**: 93-97.

Gu, X., and W. H. Li. 1992. Higher rates of amino acid substitution in rodents than in humans. Mol. Phylogenet. Evol. **1**: 211-214.

Hogstrand, K., and J. Bohme. 1998. Intrachromosomal gene conversion frequency in the H2 differs between haplotypes. Immunogenetics **48**: 47-55.

Holliday, R. 1964. A mechanism for gene conversion in fungi. Genet. Res. **5**:282-304.

Horai, S., K. Hayasaka, R. Kondo, K. Tsugane, and N. Takahata. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. U.S.A. **92**: 532-536.

Huang, C. H., Y. Chen, M. Reid, and S. Ghosh. 1996. Genetic recombination at the human RH locus: A family study of the red-cell Evans phenotype reveals a transfer of exons 2-6 from the RHD to the RHCE gene. Am. J. Hum. Genet. **59**: 825-833.

Huang, C. H., M. E. Reid, and Y. Chen. 1995. Identification of a partial internal deletion in the RH locus causing the human erythrocyte D--phenotype. Blood **86**: 784-790.

Hughes, A. L., and M. Nei. 1988. Pattern of nucleotide substitution at major

histocompatibility complex class I loci reveals overdominant selection. Nature **335**: 167-170.

Hughes, A. L., and M. Nei. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. U.S.A. **86**: 958-962.

Ina, Y. 1994. ODEN: a program package for molecular evolutionary analysis and database search of DNA and amino acid sequences. Comput. Appl. Biosci. **10**: 11-12.

Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40**: 190-226.

Ina, Y., and T. Gojobori. 1994. Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. Proc. Natl. Acad. Sci. U.S.A. **91**: 8388-8392.

Ishida, N., T. Oyunsuren, S. Mashima, H. Mukoyama, and N. Saitou. 1995. Mitochondrial DNA sequences of various species of the genus *Equus* with a special reference to the phylogenetic relationship between Przewalskii's wild horse and domestic horse. J. Mol. Evol. **41**: 180-188.

Jacobs, L. L., and D. Pilbeam. 1980. Of mice and men" Fossil-basedndivergence dates and molecular "clocks". J. Hum. Evol. **9**: 551-555.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**: 275-282.

Kajii, E., F. Umenishi, S. Iwamoto, and S. Ikemoto. 1993. Isolation of a new cDNA clone encoding an Rh polypeptide associated with the Rh blood group system. Hum. Genet. **91**: 157-162.

Kajii, E., F. Umenishi, H. Nakauchi, and S. Ikemoto. 1994. Expression of Rh blood group gene transcripts in human leukocytes. Biochem. Biophys. Res. Commun.

202: 1497-1504.

Kawano, M., S. Iwamoto, H. Okuda, S. Fukuda, N. Hasegawa, and E. Kajii. 1998. A splicing mutation of the RHAG gene associated with the Rhnull phenotype. Ann. Hum. Genet. **62**: 107-113.

Kemp, T. J., M. Poulter, and B. Carritt. 1996. A recombination hot spot in the Rh genes revealed by analysis of unrelated donors with the rare D-- phenotype. Am. J. Hum. Genet. **59**: 1066-1073.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kitano, T., K. Sumiyama, T. Shiroishi, and N. Saitou. 1998. Conserved evolution of the Rh50 gene compared to its homologous Rh blood group gene. Biochem. Biophys. Res. Commun. **249**: 78-85.

Klein, J., C. O'hUigin, and A. Blancher. 1997. Evolution of blood group antigen polymorphism. Pp. 305-321 *in* A. Blancher, J. Klein, and W. W. Socha, eds. Molecular biology and evolution of blood group and MHC antigens in primates. Springer.

Kuhner, M., S. Watts, W. Klitz, G. Thomson, and R. S. Goodenow. 1990. Gene conversion in the evolution of both the H-2 and Qa class I genes of the murine major histocompatibility complex. Genetics **126**: 1115-1126.

Kumar, S. and S. B. Hedges. 1998. A molecular timescale for vertebrate evolution. Nature **392**: 917-920.

Kumar, S., K. Tamura, and M. Nei. 1993. MEGA: Molecular Evolutionary Genetics Analysis, version 1.0. The Pennsylvania State University, University Park, PA.

Landsteiner, K., and A. S. Wiener. 1940. An agglutinable factor in human blood recognized by immune sera for blood. Proc. Soc. Exp. Med. **43**: 223.

Lee, Y. H., T. Ota, and V. D. Vacquier. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. **12**: 231-

238.

Le Van Kim, C., I. Mouro, B. Cherif-Zahar, V. Raynal, C. Cherrier, J. -P. Cartron, and Y. Colin. 1992. Molecular cloning and primary structure of the human blood group RhD polypeptide. Proc. Natl. Acad. Sci. U.S.A. **89**: 10925-10929.

Levine, P., and R. E. Stetoson. 1939. An unusual case of intragroup agglutination. J. Amer. Med. Ass. **133**: 126-127.

Li, W. H. 1997. Molecular evolution. Sinauer Associates, Inc., Sunderland, MA.

Li, W. H., and M. Tanimura. 1987. The molecular clock runs more slowly in man than in apes and monkeys. Nature **326**: 93-96.

Liebhaber, S. A., M. Goossens, and Y. W. Kan. 1981. Homology and concerted evolution at the alpha 1 and alpha 2 loci of human alpha-globin. Nature **290**: 26-29.

Long, M., and C. H. Langley. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science **260**: 91-95.

Lowell, C. A., D. A. Potter, R. S. Stearman, and J. F. Morrow. 1986. Structure of the murine serum amyloid A gene family. Gene conversion. J. Biol. Chem. **261**: 8442-8452.

Marini, A. M., S. Soussi-Boudekou, S. Vissers, and B. Andre. 1997a. A family of ammonium transporters in *Saccharomyces cerevisiae*. Mol. Cell. Biol. **17**: 4282-4293.

Marini, A. M., A. Urrestarazu, R. Beauwens, and B. Andre. 1997b. The Rh (rhesus) blood group polypeptides are related to NH_4^+ transporters. Trends Biochem. Sci. **22**: 460-461.

Marini, A. M., S. Vissers, A. Urrestarazu, and B. Andre. 1994. Cloning and expression of the *MEP1* gene encoding an ammonium transporter in *Saccharomyces cerevisiae*. EMBO J. **13**: 3456-3463.

Matassi, G., B. Cherif-Zahar, V. Raynal, P. Rouger, and J. -P. Cartron. 1998.

Organization of the human RH50A gene (RHAG) and evolution of base composition of the RH gene family. Genomics **47**: 286-293.

Messier, W., and C. B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**: 151-154.

Moore, S., and C. Green. 1987. The identification of specific Rhesus-polypeptide-blood-group-ABH-active-glycoprotein complexes in the human red-cell membrane. Biochem. J. **244**: 735-741.

Moore, S., C. F. Woodrow, and D. B. McClelland. 1982. Isolation of membrane components associated with human red cell antigens Rh(D), (c), (E) and Fy. Nature **295**: 529-531.

Mouro, I., Y. Colin, B. Cherif-Zahar, J. -P. Cartron, and C. Le Van Kim. 1993. Molecular genetic basis of the human Rhesus blood group system. Nat. Genet. **5**: 62-65.

Mouro, I., C. Le Van Kim, B. Cherif-Zahar, I. Salvignol, A. Blancher, J. -P. Cartron, and Y. Colin. 1994a. Molecular characterization of the Rh-like locus and gene transcripts from the rhesus monkey (*Macaca mulatta*). J. Mol. Evol. **38**: 169-176.

Mouro, I., C. Le Van Kim, C. Rouillac, D. J. van Rhenen, P. Y. Le Pennec, P. Bailly, J. -P. Cartron, and Y. Colin. 1994b. Rearrangements of the blood group RhD gene associated with the DVI category phenotype. Blood **83**: 1129-1135.

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

Ohta, T. 1994. On hypervariability at the reactive center of proteolytic enzymes and their inhibitors. J. Mol. Evol. **39**: 614-619.

OOta, S., and N. Saitou. 1997. Reconstruction of maximum likelihood phylogenetic trees in parallel environment using logic programming. Pp. 326-327 *in* S.

Miyano, and T. Takagi, eds. Genome Informatics, Universal Academy Press, Tokyo.

Pease, L. R. 1985. Diversity in H-2 genes encoding antigen-presenting molecules is generated by interactions between members of the major histocompatibility complex gene family. Transplantation **39**: 227-231.

Ridgwell, K., N. K. Spurr, B. Laguda, C. MacGeoch, N. D. Avent, and M. J. Tanner. 1992. Isolation of cDNA clones for a 50 kDa glycoprotein of the human erythrocyte membrane associated with Rh (rhesus) blood-group antigen expression. Biochem. J. **287**: 223-228.

Roman, H., and M. M. Ruzinski. 1990. Mechanisms of gene conversion in *Saccharomyces cerevisiae*. Genetics **124**: 7-25.

Roubinet, F., A. Blancher, W. W. Socha, and J. Ruffie. 1993. Quantitative study of chimpanzee and gorilla counterparts of the human D antigen. J. Med. Primatol. **22**: 29-35.

Ruddle, F., F. Ricciuti, F. A. McMorris, J. Tischfield, R. Creagan, G. Darlington, and T. Chen. 1972. Somatic cell genetic assignment of peptidase C and the Rh linkage group to chromosome A-1 in man. Science **176**: 1429-1431.

Saitou, N. 1991. Reconstruction of molecular phylogeny of extant hominoids from DNA sequence data. Am. J. Phys. Anthropol. **84**: 75-85.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**: 406-425.

Saitou, N., and F. Yamamoto. 1997. Evolution of primate ABO blood group genes and their homologous genes. Mol. Biol. Evol. **14**: 399-411.

Salvignol, I., A. Blancher, P. Calvas, J. Clayton, W. W. Socha, Y. Colin, and J. Ruffie. 1994. Molecular genetics of chimpanzee Rh-related genes: their relationship with the R-C-E-F blood group system, the chimpanzee counterpart of the human rhesus system. Biochem. Genet. **32**: 201-221.

Salvignol, I., A. Blancher, P. Calvas, W. W. Socha, Y. Colin, J. -P. Cartron, and J. Ruffie. 1993. Relationship between chimpanzee Rh-like genes and the R-C-E-F blood group system. J. Med. Primatol. **22**: 19-28.

Salvignol, I., P. Calvas, W. W. Socha, Y. Colin, C. Le Van Kim, P. Bailly, J. Ruffie, J. -P. Cartron, and A. Blancher. 1995. Structural analysis of the RH-like blood group gene products in nonhuman primates. Immunogenetics **41**: 271-281.

Seack, J., Z. Pancer, I. M. Mueller, and W. E. G. Mueller. 1997. Molecular cloning and primary structure of a Rhesus (Rh)-like protein from the marine sponge *Geodia cydonium*. Immunogenetics **47**: 493-498.

Slightom, J. L., A. E. Blechl, and O. Smithies. 1980. Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell **21**: 627-638.

Slightom, J. L., T. W. Theisen, B. F. Koop, and M. Goodman. 1987. Orangutan fetal globin genes. J. Biol. Chem. **262**: 7472-7483.

Thompson, C. B. 1992. Creation of immunoglobulin diversity by intrachromosomal gene conversion. Trends Genet. **8**: 416-422.

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

Tippett, P. 1986. A speculative model for the Rh blood groups. Ann. Hum. Genet. **50**: 241-247.

Van Dommelen, A., V. Keijers, J. Vanderleyden, and M. de Zamaroczy. 1998. (Methyl) ammonium transport in the nitrogen-fixing bacterium *Azospirillum brasilense*. J. Bacteriol. **180**: 2652-2659.

Westhoff, C. M., and D. E. Wylie. 1996. Investigation of the Rh locus in gorillas and chimpanzees. J. Mol. Evol. **42**: 658-668.

Wiener, A. S. 1943. Genetic theory of the Rh blood types. Proc. Soc. Exp. Biol. N. Y. **54**: 316-319.

Wilson, A. C., S. S. Carlson, and T. J. White. 1977. Biochemical evolution. Annu. Rev. Biochem. **46**: 573-639.

Wilson, R., R. Ainscough, K. Anderson. et al. (53 co-authors). 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. Nature **368**: 32-38.

Wu, C.I., and W. H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. U.S.A. **82**: 1741-1745.

Yang, Z. 1997. Phylogenetic Analysis by Maximum Likelihood (PAML), Version 1.3. Department of Integrative Biology, University of California at Berkeley.

Zhang, J., S. Kumar, and M. Nei. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol. Biol. Evol. **14**: 1335-1338.

Zhang, J., H. F. Rosenberg, and M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. U.S.A. **95**: 3708-3713.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to Dr. Naruya Saitou for his invaluable guidance throughout this work. Grateful acknowledgement is extended to Profs. Takashi Gojobori, Toshimichi Ikemura, Yuji Kohara, Toshihiko Shiroishi, and Shintaro Ueda for serving on my advisory committee. I owe thanks to Drs. Tomoko Sagai and Yuzuru Ikehara for providing samples for mice and crab-eating macaque, and Drs. Kenta Sumiyama and Tetsushi Yamagata for their help and advice at this work, and Dr. Yasuo Ina for providing his programs, and Dr. Tsuyoshi Koide for discussing the generation time of wild mice, and Dr. Satoshi OOta for estimating values of branch length for maximum likelihood trees (figure 3.8 in Chapter III) by using his program. I also thank colleagues Miss Mika Kaneko and Miss Reiko Noda of the Laboratory of Evolutionary Genetics, and Mrs. Tatsuko Kwamoto, Mrs. Yumiko Yoshida, Miss Eri Sugimura, Miss Akane Suzuki, and Miss Keiko Ichikawa for their help and encouragement during this work. I would like to thank my parents, Masaki and Kyoko Kitano. They gave me this opportunity and support have been the motives to finish this work. I also would like to express my thanks to my friends Masahiro Iwasa, Kiyoshi Saigusa, Ryoji Hatano, and Michiko Sagawa for their encouragement during this work. Finally, I thank "LUCKY STRIKE" and "Maker's Mark" for relaxing my mind throughout this writing.

BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS 249, 78–85 (1998)
ARTICLE NO. RC989074

Conserved Evolution of the Rh50 Gene Compared to Its Homologous Rh Blood Group Gene

Takashi Kitano,* Kenta Sumiyama,*,[1] Toshihiko Shiroishi,† and Naruya Saitou*
*Laboratory of Evolutionary Genetics and †Laboratory of Mammalian Genetics, National Institute of Genetics, Mishima 411-8540, Japan

Received July 3, 1998

We have sequenced the complete coding region of the Rh blood group gene for mouse and rat and that of Rh-related 50 kD glycoprotein (Rh50) for mouse, rat, and crab-eating macaque. Phylogenetic analyses of Rh and Rh50 amino acid sequences indicate that the Rh50 gene has been evolving about two times more slowly than the Rh blood group gene in both primates and rodents. This conservative nature of the Rh50 gene suggests its relative importance to the Rh blood group gene. The time of gene duplication that produced the Rh and Rh50 genes was estimated to be about 240-310 million years ago. We also conducted window analyses of synonymous and nonsynonymous nucleotide substitutions for those two genes. Some peaks where nonsynonymous substitutions are higher than synonymous ones were located on outer membrane regions. This suggests the existence of positive Darwinian selection on Rh and Rh50 genes through host-parasite interactions.

Key Words: **evolutionary rate; phylogenetic tree; window analysis; positive selection.**

The human Rh blood group plays important roles in transfusion and clinical medicine, including haemolytic diseases of newborns, autoimmune diseases, and mild haemolytic anemia. Nucleotide sequences of human Rh genes were determined (1-5), and their products were estimated to have 12 transmembrane domains through hydropathy analysis (2) and immunological studies using an antipeptide antibody (6). Nucleotide sequences of Rh blood group genes in nonhuman primates were also reported (7-9). We recently analyzed those published Rh blood group genes of primates, and found a higher nonsynonymous substitutions than synonymous ones, a clear evidence of positive Darwinian evolution (Kitano and Saitou, unpublished). Therefore, it is interesting if a similar evolutionary pattern also exists in other mammalian groups.

A protein was obtained together with the Rh gene product on immunoprecipitation with anti-Rh antibodies from human, and named as 50 kD glycoprotein (10). This glycoprotein was considered to form heterotetramer with Rh blood group gene products on erythrocyte membranes (11). The nucleotide sequence of the human 50 kD glycoprotein was determined, and its amino acid sequence was homologous with that of the human Rh gene (12). That protein was also predicted to have the 12 trans-membrane domains which are similar to those of the Rh blood group gene product. There are several names for this gene such as RHAG, but we call this gene as Rh50 and the Rh blood group gene as Rh hereafter for simplicity. It has been shown that the Rh_{null} regulator and the Rh_{mod} phenotypes are suppressed by the Rh50 product (13), and a splicing mutant of this gene was shown to cause an Rh_{null} phenotype (14). These observations clearly indicate that the Rh50 gene is essential for expression of Rh antigens on erythrocytes.

These Rh gene and Rh related gene products seem to play an important role for erythrocytes. In this study, we determined nucleotide sequences for Rh genes and Rh50 genes of two mouse subspecies and rat as well as the Rh50 gene for crab-eating macaque, and compared these evolutionary relationships.

MATERIALS AND METHODS

PCR-direct sequencing of cDNA. Total RNAs were extracted from bone marrow of two mouse (*Mus musculus*) subspecies (*M. m. domesticus* and *M. m. brevirostris*), rat (*Rattus norvegicus*), and crab-eating macaque (*Macaca fascicularis*), using the AGPC (Acid Guanidinium-Phenol-Chloroform) method. After DNase reactions, reverse transcription was performed by using AMV (Avian Myeloblastosis Virus) reverse transcriptase and oligo dT-adaptor primer of RNA PCR Kit AMV Ver. 2.1 (TaKaRa). Degenerate PCR was performed and a partial product was obtained. We then performed

[1] Present address: Department of Biology, Yale University, New Haven, CT 06520-8103.

0006-291X/98 $25.00

5' RACE (rapid amplification of the 5' cDNA ends) using 5'RACE System for Rapid Amplification of cDNA Ends version 2.0 (Gibco-BRL). PCR was performed in a 20 μl reaction containing 0.5-1 μl of the first-strand cDNA, 1 × Gene Taq Universal Buffer (Mg^{2+} free) (Nippon Gene), 1.5 mM $MgCl_2$, 0.2 mM dNTP, 10 pmol of each primer (designed on sites of 5' and 3' ends), and 1 units of AmpliTaq Gold (Perkin-Elmer). Amplification was carried out in DNA GeneAmp PCR System 2400 (Perkin-Elmer) with the following temperature parameters: 10 min at 95°C followed by 40 cycles of 95°C for 30 sec, 65°C for 15 sec, and 72°C for 1 min. PCR products were purified using MicroSpin Columns S-300 HR (Pharmacia Biotech) and cloned in the TA cloning vectors pCRII (Invitrogen). DNA sequencing was performed on double-stranded plasmid DNA and PCR products using Dye Terminator Cycle Sequencing Kit and ABI prism 377 DNA sequencer (Perkin-Elmer). A progressive sequencing strategy was carried out with design of further primers to complete the sequence for coding region of both strands of the cDNA.

Sequence analyses. CLUSTAL W version 1.6 (15) was used for multiple alignments and tree analyses. The neighbor-joining method (16) was used for constructing phylogenetic trees. Numbers of amino acid substitutions were estimated by using Kimura's method (17). The PredictProtein server (EMBL) was used for analyses of transmembrane helix location of these proteins. ODEN package (18) was used to estimate numbers of synonymous and nonsynonymous substitutions (19), and the WINA program (20) was used for window analyses.

RESULTS AND DISCUSSION

PCR-cloning and sequencing of Rh and Rh50 cDNA coding regions. We sequenced Rh and Rh50 gene cDNAs for two mouse subspecies and rat. Those newly determined rodent sequences (DDBJ/EMBL/GenBank international nucleotide sequence database accession numbers are AB015189 - AB015194) were compared with the human Rh gene (1, 2) and the human Rh50 gene (12). Figure 1 shows the multiple alignment of nucleotide sequences of Rh genes. Nucleotide sequence lengths of human, crab-eating macaque, mouse, and rat are 1254, 1254, 1257, and 1269 bp, respectively. Four gaps (3, 15, 3, and 6 nucleotide long) were observed between primate and rodent sequences, and the rat Rh gene had extra 12 nucleotides (positions 337-348). Lengths of all gaps were multiplication of 3 and there is no frame shift. We also obtained an incomplete sequence for the rat Rh cDNA which lacks sites 149-661. These sites correspond to exons 2-4 of the human Rh gene, and this incomplete cDNA were probably produced by a splicing error.

Figure 2 shows the multiple alignment of nucleotide sequences of Rh50 genes. We obtained a Rh50 gene cDNA for crab-eating macaque (DDBJ/EMBL/GenBank accession number is AB015467), and it was also compared. Nucleotide sequence lengths of human, crab-eating macaque, mouse, and rat are 1230, 1287, 1317, and 1353 bp, respectively. The location of the stop codon of the human Rh50 gene is different from that of others, and its protein is 19 amino acids shorter corresponding to this region. There are repeats of 15 nucleotides around positions 100-150 (see Fig. 2), and its consensus sequence is AATGCTTCCCAGCAG. Rat and mouse have 5 and 3 repeats, respectively, while the two primate species have single repeat. Because all gaps were multiple of 3, they did not alter codon frames.

Sequence similarities (both for nucleotide and amino acid) are shown in Table 1. Nucleotide and amino acid sequence similarities between Rh genes and Rh50 genes are 47-49 and 34-38%, respectively. The GC contents of Rh and Rh50 genes were 53-55 and 45-47%, respectively (shown on the diagonal of Table 1). These values were similar to those previously reported (21), and may be related to gene locations on gnomes; the Rh gene is located on chromosome 1p34-36 (22-23), while the Rh50 gene is on chromosome 6p21-qter (12). We would like to note that similar sequencing results for mouse and macaque were recently obtained independently by G. Matassi *et al.* (personal communication).

Evolutionary rates and the phylogenetic tree. We estimated numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions for Rh and Rh50 genes (Table 2). d_S and d_N values between primates and rodents were estimated by averaging pairwise values. Numbers of synonymous substitutions (d_S) were similar between Rh and Rh50 genes, and they are more or less similar to those for other genes (24). Branching pattern of the Rh and Rh50 genes are also compatible with the established mammalian phylogeny. This indicates that we did orthologous comparison both for Rh and Rh50 genes.

Numbers of nonsynonymous substitutions (d_N) are about two times higher for the Rh gene than for the Rh50 gene; the ratios of Rh-d_N and Rh50-d_N are 2.0, 1.7, and 2.0 for human-macaque, mouse-rat, and primates-rodents comparisons, respectively (we neglected the comparison of the two mouse subspecies, for standard errors are so large). This evolutionary conservation of the Rh50 gene suggests that it may have more important function than the Rh gene. A relatively uniform ratio of Rh-d_N and Rh50-d_N for three different levels of divergence also suggests that a molecular clock (constancy of evolutionary rate) exists both for Rh and Rh50 genes.

Majority of genes are known to undergo neutral evolution, and number (d_S) of synonymous substitutions are expected to be higher than those (d_N) for nonsynonymous substitutions under this situation (17). We compared d_S and d_N values to see if there is any unusual pattern deviated from neutrality in Rh and Rh50 genes. d_N of both Rh and Rh50 genes were higher than d_S when human and macaque sequences were compared, while the situation is reversed for other comparisons (Table 2). We previously found that many branches of a phylogenetic tree of primate

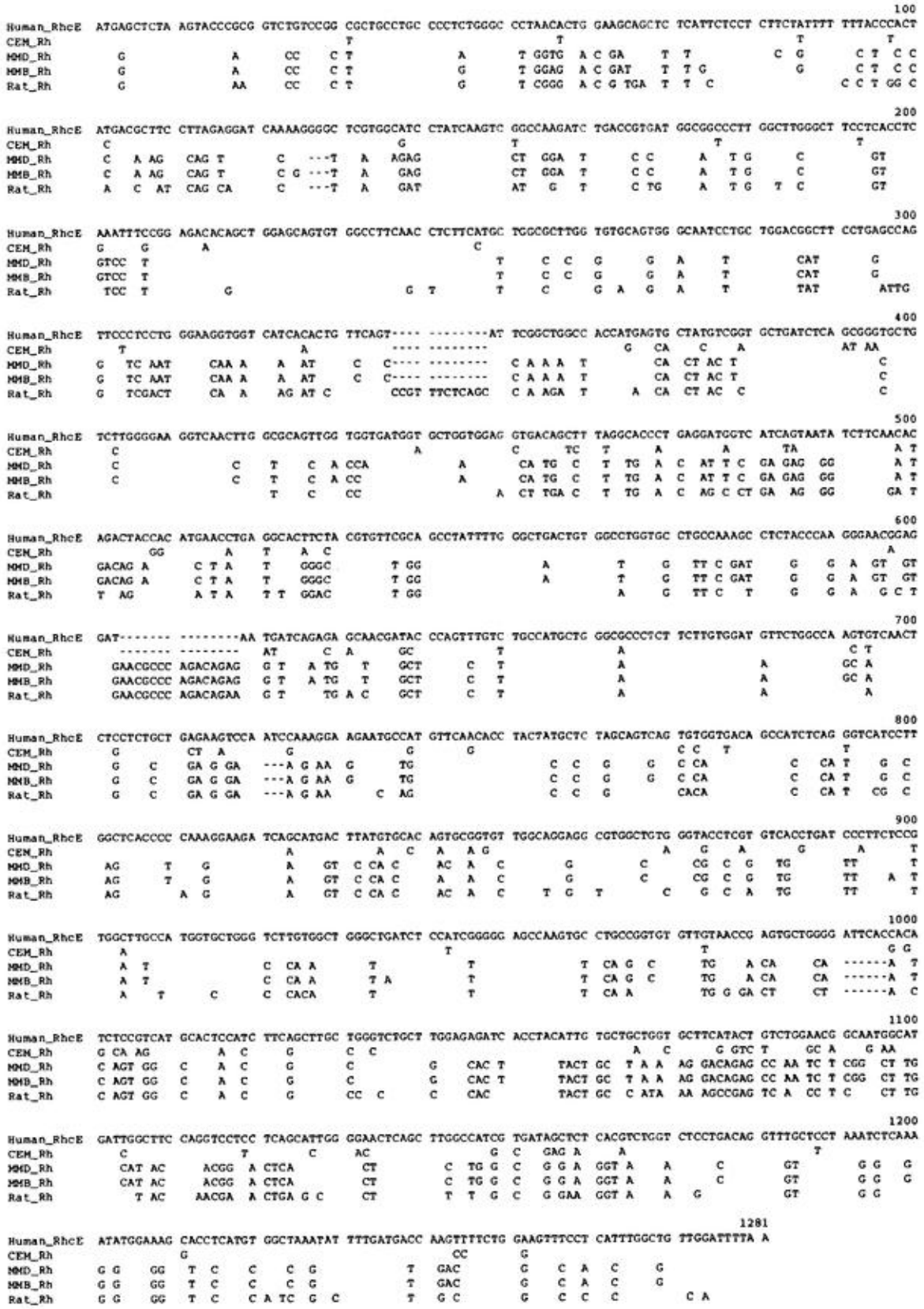


FIG. 1. The multiple alignment of nucleotide sequences of Rh genes. Nucleotide sequences for human RhcE (M34015 or X54534; they are identical) and for crab-eating macaque (L37054) were also included for comparison. Gaps are denoted by hyphens, and only nucleotides different from those of the human sequence are shown. MMD, MMB, and CEM denote *M. m. domesticus* and *M. m. brevirostris*, and crab-eating macaque, respectively.

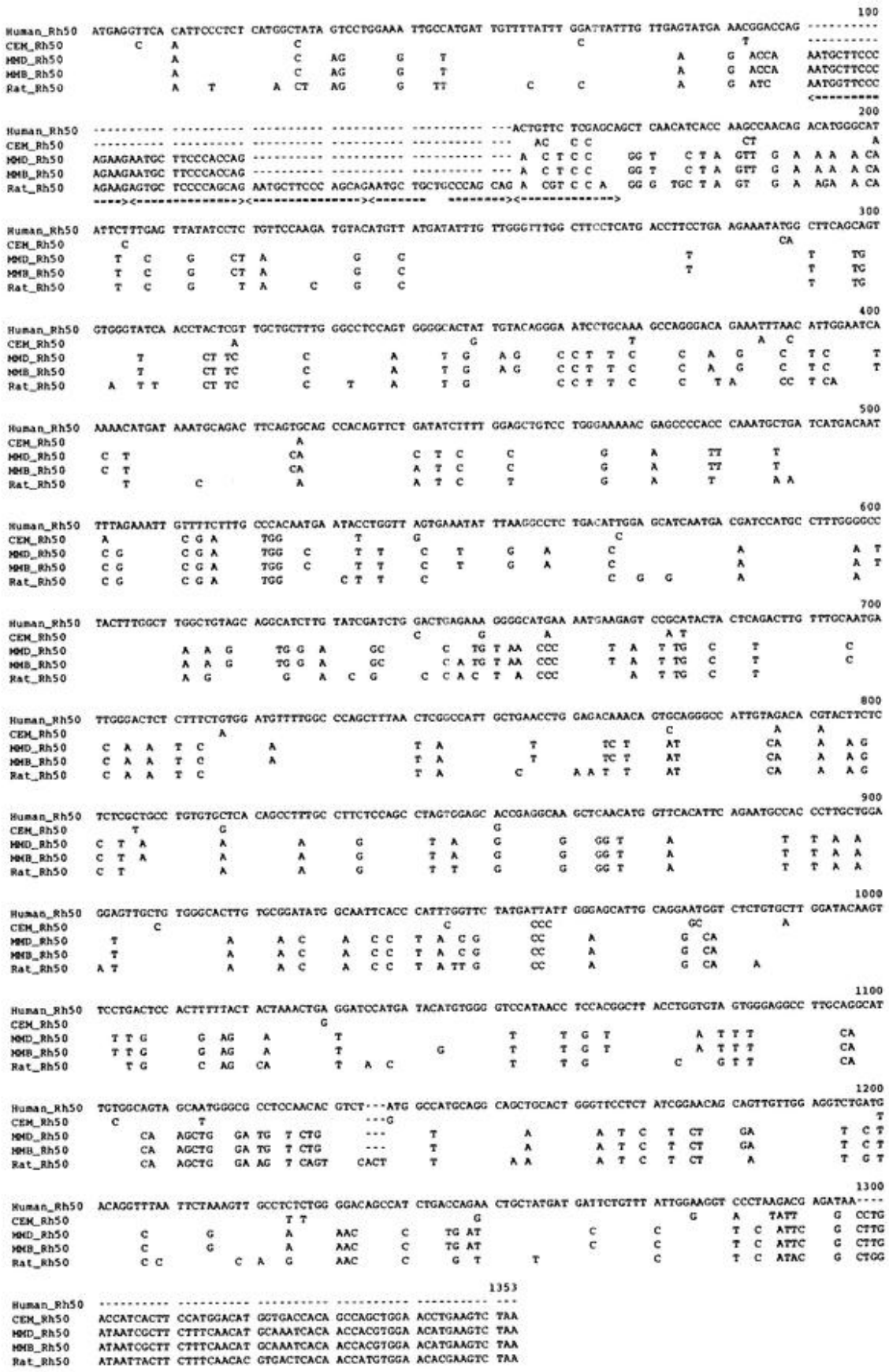


FIG. 2. The multiple alignment of nucleotide sequences of Rh50 genes. Human Rh50 (X64594) was also included. Gaps are denoted by hyphens, and only nucleotides different from those of the human sequence are shown. Equal signs surrounded with angled brackets designate the repeat unit of 15 nucleotides. Abbreviations of species are the same as Fig. 1.

TABLE 1

Similarities (%) of Rh and Rh50 Nucleotide Sequences (above Diagonal) and Amino Acid Sequences (below Diagonal) and GC Content (%) of Each Gene (on the Diagonal in Parentheses)

	1	2	3	4	5	6	7	8	9	10
1. Human RhcE	(53.7)	90.4	71.4	71.3	70.8	48.6	48.5	47.3	47.4	47.3
2. CEM Rh	79.1	(52.5)	71.9	71.8	70.9	48.8	48.9	48.4	48.5	48.2
3. MMD Rh	57.9	59.1	(55.2)	99.1	88.3	48.6	48.9	47.2	47.1	46.9
4. MMB Rh	58.1	59.3	98.3	(55.3)	88.3	48.5	48.8	47.1	47.0	46.8
5. Rat Rh	56.7	58.6	81.6	81.8	(54.0)	48.8	48.5	47.7	47.6	47.4
6. Human Rh50	35.2	35.7	37.1	36.8	35.3	(46.7)	94.6	80.0	79.8	79.4
7. CEM Rh50	35.4	37.7	37.8	37.5	36.3	88.8	(47.4)	79.6	79.4	78.6
8. MMD Rh50	34.4	35.4	35.0	35.0	35.5	77.0	74.3	(45.0)	99.8	91.6
9. MMB Rh50	34.7	35.7	34.8	34.8	35.3	76.8	74.1	99.8	(45.0)	91.7
10. Rat Rh50	33.8	35.1	34.7	34.7	34.4	75.8	73.1	88.8	89.0	(45.4)

Note. MMD, MMB, and CEM denote *M. m. domesticus, M. m. brevirostris,* and crab-eating macaque, respectively.

Rh genes showed higher d_N than d_S (Kitano and Saitou, unpublished), and this is compatible with a higher d_N for human and macaque Rh gene shown in Table 2. It is interesting that the Rh50 gene also showed a similar evolutionary pattern for primates, but not for rodents. If the heterotetramer structure of the Rh and Rh50 gene products is correct, it is possible that this erythrocyte membrane protein complex is under some kind of positive selection in primates but not in rodents.

We constructed a multiple alignment of amino acid sequences of Rh, Rh50, and their related genes (alignment not shown). Two genes of *C. elegans* (25) and an Rh-like gene of sponge (26) found by database searches were also included. The 12 predicted hydrophobic membrane-spanning regions did not include gaps and are relatively conserved. We thus used only the 216 amino acid sites for membrane-spanning regions for construction of the neighbor-joining tree (Fig. 3). The root was located by assuming the Rh-like protein of sponge as an outgroup. There are three clusters in this tree; Rh50 genes of mammals, Rh genes of mammals, and two genes of *C. elegans*. The branch lengths of Rh50 genes is much shorter than those of Rh genes, indicating a lower evolutionary rate in the Rh50 gene than in the Rh gene. This pattern is consistent with the result of d_N in Table 2 where all the coding regions were compared. It is interesting that after the gene duplication (node D in Fig. 3) which produced Rh and Rh50 genes, the Rh gene lineage started to evolve more rapidly than the Rh50 lineage.

Numbers of amino acid substitutions (d_A) are also estimated (Table 3). The phylogenetic tree of Fig. 3 was used and single-lineage d_A values were obtained applying Ishida *et al.*'s (27) method. Because d_A values for Rh were consistently two - three times higher than those for Rh50, a rough molecular clock exists for both genes. This is consistent with the result of Table 2. Therefore, we estimated evolutionary rates of Rh and Rh50 genes by using the regression through origin. Divergence times between human and macaque, between mouse and rat, and between primates and rodents were assumed to be 23.3, 40.7, and 112 million years (28), and they were used for calibration of the molecular clock. The rate of amino acid substitutions (per site per year) were thus obtained as 2.12×10^{-9} and 0.94×10^{-9} for Rh and Rh50 genes, respectively. If we use these rates, the time of gene duplication (node D in Fig. 3) producing Rh and Rh50 genes was estimated to be about 240 or 310 million years ago from the data for Rh or Rh50, respectively. This period roughly corresponds to the late Paleozoic where the mammalian lin-

TABLE 2

Numbers of Synonymous (d_S) and Nonsynonymous (d_N) Substitutions

	MMD vs MMB[a]	Human vs macaque[a]	Mouse vs rat[a]	Primates vs rodents[b]
d_S of Rh	0.013 ± 0.007	0.071 ± 0.016	0.226 ± 0.031	0.595
d_S of Rh50	0.007 ± 0.005	0.049 ± 0.013	0.200 ± 0.028	0.620
d_N of Rh	0.007 ± 0.003	0.115 ± 0.011	0.098 ± 0.011	0.302
d_N of Rh50	0.001 ± 0.001	0.057 ± 0.008	0.058 ± 0.008	0.153

Note. MMD and MMB designate *M. m. domesticus* and *M. m. brevirostris,* respectively.
[a] Pairwise values with standard errors.
[b] Averages of pairwise values.

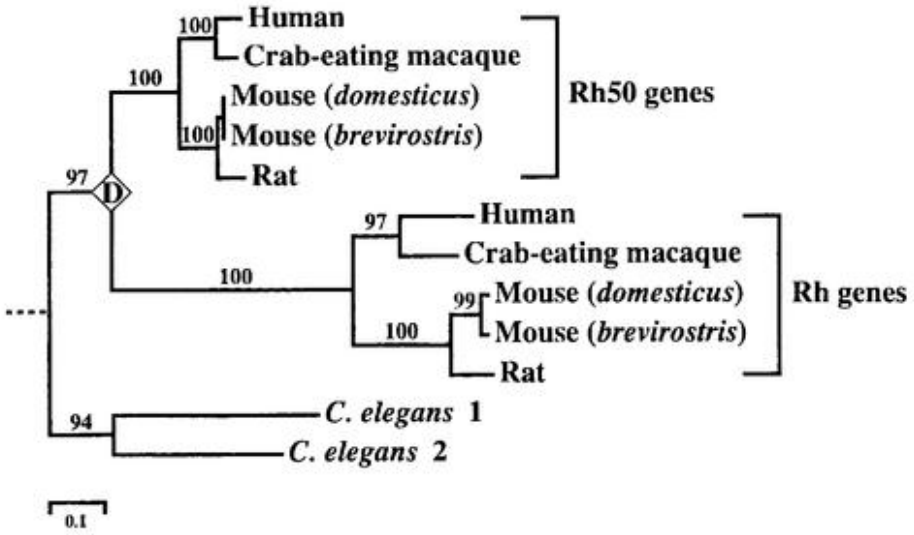


FIG. 3. The phylogenetic tree of mammalian Rh blood group genes and their homologous genes. Amino acid sequences of only membrane-spanning regions (216 sites) are used. Bootstrap probabilities (%) are shown on interior branches. The root was located by assuming the Rh-like protein of sponge as an outgroup. Accession numbers and gene IDs for *C. elegans* 1 and 2 are Z74026-B0240.1 and U64847-F08F3.3, respectively.

eage started to diverge from an ancestral reptilian lineage.

The Rh and its related genes have probably been existing as essential membrane proteins in many animal phyla. Because its evolutionary rate is lower than that for the Rh protein gene, the Rh50 protein may be closer to the ancestral form before gene duplication of Rh and Rh50 genes. In this context, it may be interesting to note a similarity between the Rh50 protein and NH_4^+ transporters (29). However, two coding regions of *C. elegans* (C05E11.4 in the DDBJ/EMBL/GenBank accession number U53338 and M195.3 in Z66498) are more similar to NH_4^+ transporters. Therefore, the actual function of Rh and Rh50 gene products remain to be found.

Window analyses of synonymous and nonsynonymous substitutions. We performed window analyses for synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions to investigate their possible correlation with the protein structure (Fig. 4). The twelve predicted hydrophobic membrane-spanning regions are shown by black boxes with numbers. There are several peaks (depicted by arrows) where nonsynonymous substitutions are higher than synonymous ones on putative outer membrane regions on primate Rh genes (Fig. 4A). One peak (designated as long arrows) is observed at the cell surface region between membrane-spanning regions 3 and 4 in all four comparisons, and four and two peaks were also observed for other cell surface regions in primate Rh and Rh50 comparisons, respectively (indicated by short arrows with asterisks). Whether an intermembrane region is on the cell surface or in the cytoplasm is based on the currently accepted model of the Rh and Rh50 protein structure (2, 30).

We examined Rh blood group gene and its homologous Rh50 genes in the present study. Blood group antigens may play a key role in pathogenesis of diseases (31), and there is a possibility of positive Darwinian selection caused by interaction between organisms (host mammals and parasites such as bacteria). In fact, Saitou and Yamamoto (32) and Kitano and Saitou (unpublished) found evidences of positive selection in the ABO and Rh blood group genes of primates, respectively. Comparison of synonymous and nonsynonymous substitutions for the Rh50 gene also revealed a possibility of existence of positive selection for this gene in primates. Because primates showed more clear sign of positive selection than rodents both for Rh and Rh50 genes, it is possible that the pattern of host-parasite interaction is different between primates and rodents.

TABLE 3

Numbers (d_A) of Amino Acid Substitutions and Divergence Times

Diverging node	Human/ macaque	Mouse/ rat	Primates/ rodents	Rh/Rh50
Single lineage d_A of Rh[a]	0.120	0.072	0.228	0.663
Single lineage d_A of Rh50[a]	0.042	0.029	0.104	0.225
Divergence time (MYA)	23.3[b]	40.7[b]	112[b]	240–310[c]

[a] Based on phylogenetic tree of Fig. 3.
[b] Taken from Ref. 28.
[c] Estimated from d_A values.

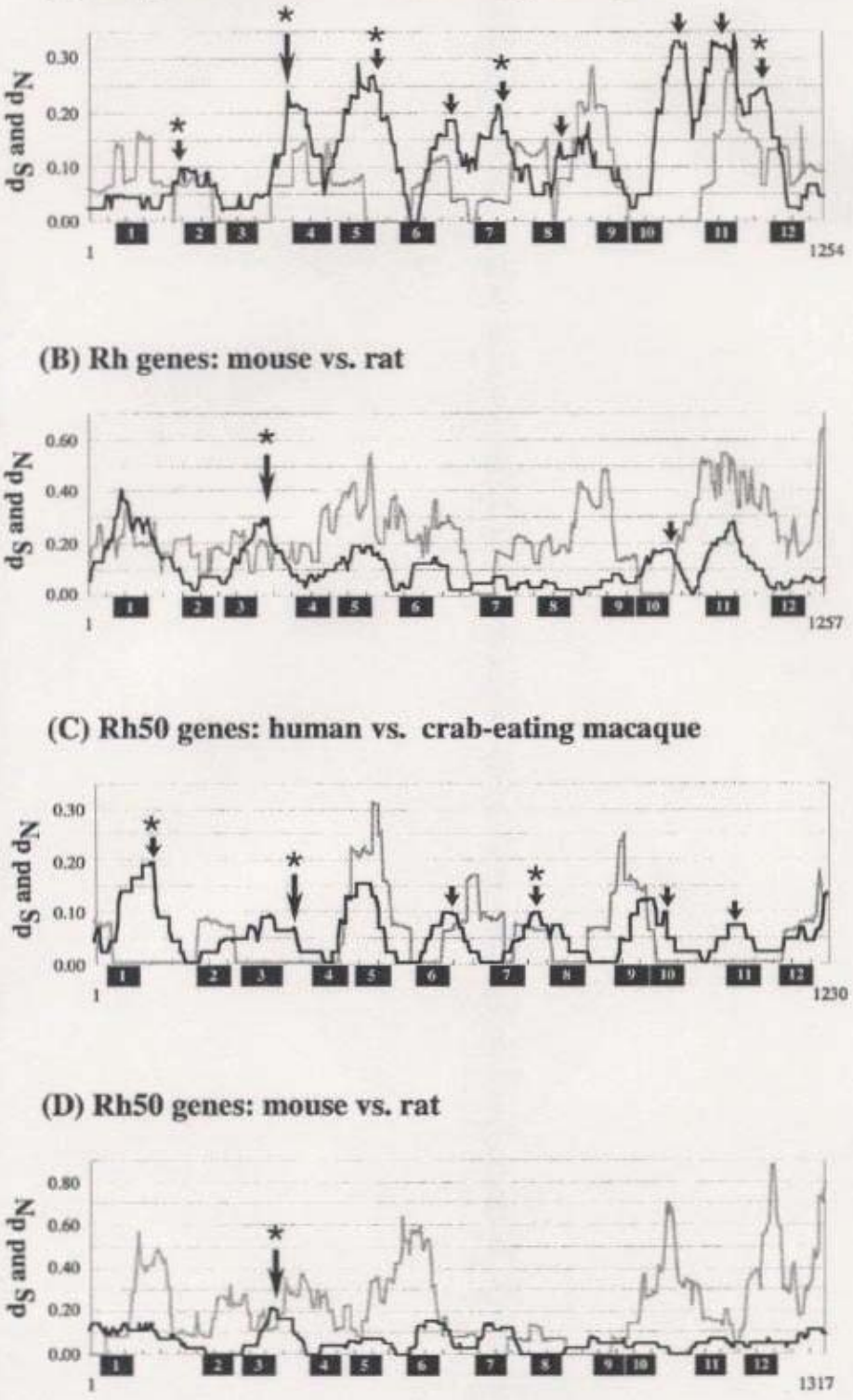


FIG. 4. Window analyses for synonymous (d_S: gray lines) and nonsynonymous (d_N: black lines) nucleotide substitutions for Rh genes between human and crab-eating macaque (A), for Rh genes between *M. m. domesticus* and rat (B), for Rh50 genes between human and crab-eating macaque (C), and for Rh50 genes between *M. m. domesticus* and rat (D). The 12 predicted hydrophobic membrane-spanning regions are shown by black boxes. Horizontal axes indicate numbers of nucleotide sites. Arrows indicate peaks of d_N higher than d_S, and arrows with asterisks are in the putative cell surface regions. Arrows in the same cell surface region for all the A-D comparisons are drawn longer than others.

ACKNOWLEDGMENTS

We thank Dr. Y. Ikehara for preparation of the crab-eating macaque tissues and Dr. T. Sagai for preparation of the mouse tissues. This study was partially supported by grants-in-aid for scientific studies from the Ministry of Education, Science, Sport, and Culture, Japan, to N. S.

REFERENCES

1. Cherif-Zahar, B., Bloy, C., Le Van Kim, C., Blanchard, D., Bailly, P., Hermand, P., Salmon, C., Cartron, J.-P., and Colin, Y. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 6243–6247.
2. Avent, N. D., Ridgwell, K., Tanner, M. J., and Anstee, D. J. (1990) *Biochem. J.* **271,** 821–825.
3. Le van Kim, C., Mouro, I., Cherif-Zahar, B., Raynal, V., Cherrier, C., Cartron, J.-P., and Colin, Y. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10925–10929.
4. Kajii, E., Umenishi, F., Iwamoto, S., and Ikemoto, S. (1993) *Hum. Genet.* **91,** 157–162.
5. Arce, M., Thompson, E. S., Wagner, S., Coyne, K. E., Ferdman, B. A., and Lublin, D. M. (1993) *Blood* **82,** 651–655.
6. Avent, N. D., Butcher, S. K., Liu, W., Mawby, W. J., Mallinson, G., Parsons, S. F., Anstee, D. J., and Tanner, M. J. (1992) *J. Biol. Chem.* **267,** 15134–15139.
7. Salvignol, I., Blancher, A., Calvas, P., Clayton, J., Socha, W. W., Colin, Y., and Ruffie, J. (1994) *Biochem. Genet.* **32,** 201–221.
8. Salvignol, I., Calvas, P., Socha, W. W., Colin, Y., Le Van Kim, C., Bailly, P., Ruffie, J., Cartron, J.-P., and Blancher, A. (1995) *Immunogenetics* **41,** 271–281.
9. Mouro, I., Le Van Kim, C., Cherif-Zahar, B., Salvignol, I., Blancher, A., Cartron, J.-P., and Colin, Y. (1994) *J. Mol. Evol.* **38,** 169–176.
10. Moore, S., and Green, C. (1987) *Biochem. J.* **244,** 735–741.
11. Eyers, S. A. C., Ridgwell, K., Mawby, W. J., and Tanner, M. J. A. (1994) *J. Biol. Chem.* **269,** 6417–6423.
12. Ridgwell, K., Spurr, N. K., Laguda, B., MacGeoch, C., Avent, N. D., and Tanner, M. J. (1992) *Biochem. J.* **287,** 223–228.
13. Cherif-Zahar, B., Raynal, V., Gane, P., Mattei, M. G., Bailly, P., Gibbs, B., Colin, Y., and Cartron, J.-P. (1996) *Nature Genet.* **12,** 168–173.
14. Kawano, M., Iwamoto, S., Okuda, H., Fukuda, S., Hasegawa, N., and Kajii, E. (1998) *Ann. Hum. Genet.* **62,** 107–113.
15. Thompson, J. D., Gibson, T. J., and Higgins, D. G. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
16. Saitou, N., and Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
17. Kimura, M. (1983) The Neutral Theory of Molecular Evolution, Cambridge Univ. Press, Cambridge.
18. Ina, Y. (1994) *Comput. Appl. Biosci.* **10,** 11–12.
19. Nei, M., and Gojobori, T. (1986) *Mol. Biol. Evol.* **3,** 418–426.
20. Endo, T., Ikeo, K., and Gojobori, T. (1996) *Mol. Biol. Evol.* **13,** 685–690.
21. Matassi, G., Cherif-Zahar, B., Raynal, V., Rouger, P., and Cartron, J.-P. (1998) *Genomics* **47,** 286–293.
22. Ruddle, F., Ricciuti, F., McMorris, F. A., Tischfield, J., Creagan, R., Darlington, G., and Chen, T. (1972) *Science* **176,** 1429–1431.
23. Cherif-Zahar, B., Mattei, M. G., Le Van Kim, C., Bailly, P., Cartron, J.-P., and Colin, Y. (1991) *Hum. Genet.* **86,** 398–400.
24. Li, W.-H., and Tanimura, M. (1987) *Nature* **326,** 93–96.
25. Wilson, R., Ainscough, R., Anderson, K., *et al.* (53 co-authors) (1994) *Nature* **368,** 32–38.
26. Seack, J., Pancer, Z., Mueller, I. M., and Mueller, W. E. G. (1997) *Immunogenetics* **47,** 493–498.
27. Ishida, N., Oyunsuren, T., Mashima, S., Mukoyama, H., and Saitou, N. (1995) *J. Mol. Evol.* **41,** 180–188.
28 Kumar, S., and Hedges, S. B. (1998) *Nature* **392,** 917–920.
29. Marini, A. M., Urrestarazu, A., Beauwens, R., and Andre, B. (1997) *Trends Biochem. Sci.* **22,** 460–461.
30. Avent, N. D., Butcher, S. K., Lie, W., Mawby, W. J., Mallinson, G., Parsons, S. F., Anstee, D. J., and Tanner, M. J. A. (1992) *J. Biol. Chem.* **267,** 15134–15139.
31. Eder, A. F., and Spitalnik, S. L. (1997) *in* Molecular Biology and Evolution of Blood Group and MHC Antigens in Primates (Blancher, A., Klein, J., and Socha, W. W., Eds.), pp. 268–304, Springer.
32. Saitou, N., and Yamamoto, F. (1997) *Mol. Biol. Evol.* **14,** 399–411.

APPENDIX II

These are multiple alignments of non-converted and ancestral sequences for tree A (A) and for tree B (B) of figure 2.5 in Chapter II. Cem and Rhm mean crab-eating macaque and rhesus macaque, respectively. N means undetermined site.

(A) Non-converted and ancestral sequences for tree A

```
Human_D         ATGAGCTCTAAGTACCCGCGGTCTGTCCGGCGCTGCCTGCCCCTCTGGGCCCTAACACTG
Human_CE        ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
Chimpanzee_1    ATGAGCTCTAAGTACCCGCGGTCTGTCCGGCGCTGCCTGCCCCTCTGCGCCCTAACACTG
Chimpanzee_2    ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
Gorilla_1       ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
Gorilla_2       ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
Cem_1           ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACTCTG
Cem_2           ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACTCTG
Rhm_1           ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACACTG
node_a          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
node_b          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
node_c          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGCGCTGCCTGCCCCTCTGCGCCCTAACACTG
node_d          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
node_e          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
node_f          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACACTG
node_g          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACTCTG
                ******************************** ***************** ******** ***

Human_D         GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTATGACGCTTCCTTAGAGGAT
Human_CE        GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
Chimpanzee_1    GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCAGTACGATGCTTCCTTAGAGGAT
Chimpanzee_2    GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
Gorilla_1       GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
Gorilla_2       GAAGCAGCTCTCACTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
Cem_1           GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
Cem_2           GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
Rhm_1           GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
node_a          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_b          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_c          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_d          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_e          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_f          GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
node_g          GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
                ************* *********** ********** * ** ** ***************

Human_D         CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Human_CE        CAAAAGGGGCTCGTGGCATCCTATCAAGTCGGCCAAGATCTGACCGTGATGGCGGCCCTT
Chimpanzee_1    CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Chimpanzee_2    CAAAAGGGGCTCGTGGCGTCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Gorilla_1       CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Gorilla_2       CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Cem_1           CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
Cem_2           CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
Rhm_1           CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
node_a          CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
node_b          CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
node_c          CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
node_d          CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
node_e          CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
node_f          CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
node_g          CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
                ***************** ***********  ************************* * **
```

```
Human_D         GGCTTGGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
Human_CE        GGCTTGGGCTTCCTCACCTCAAATTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
Chimpanzee_1    GGCTTTGGCTTCCTTACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAGC
Chimpanzee_2    GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
Gorilla_1       GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
Gorilla_2       GGCTTTGGCTTCCTCACCTCGAGTTTCCGGGGACACAGCTGGAGCAGTGTGGCCTTCAAC
Cem_1           GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
Cem_2           GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
Rhm_1           GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
node_a          GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_b          GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_c          GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_d          GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_e          GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_f          GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
node_g          GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
                ***** ****** * ***** * *** *** ** ************************ *

Human_D         CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Human_CE        CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Chimpanzee_1    CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATTCTGCTGGACGGCTTCCTGAGCCAG
Chimpanzee_2    CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Gorilla_1       CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Gorilla_2       CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Cem_1           CTCTTCCTGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Cem_2           CTCTTCCTGCTCGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Rhm_1           CTCTTCCTGCTGGCCCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_a          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_b          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_c          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_d          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_e          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_f          CTCTTCCTGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_g          CTCTTCCTGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
                ****** **** ** ********************** ************************

Human_D         TTCCCTTCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
Human_CE        TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
Chimpanzee_1    TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCACGAGTGCT
Chimpanzee_2    TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCACGAGTGCT
Gorilla_1       TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
Gorilla_2       TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
Cem_1           TTCTCTCCTGGGAAGGTGGTCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
Cem_2           TTCTCTCCTGGGAAGGTGGTCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
Rhm_1           TTCTCTCCTGGGAAGGTGGCCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
node_a          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_b          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_c          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_d          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_e          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_f          TTCTCTCCTGGGAAGGTGGTCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
node_g          TTCTCTCCTGGGAAGGTGGTCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
                *** ** ************ ***** *************************** ***  **

Human_D         TTGTCGGTGCTGATCTCAGTGGATGCTGTCTTGGGGAAGGTCAACTTGGCGCAGTTGGTG
Human_CE        ATGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGTACGTCAACTTGGTGCAGTTGGTG
Chimpanzee_1    TTGTCAGTGCTGATCTCAGTGGATGCTGTCTTGGGGAAGGTCAACTTGGTGCAGTTGGTG
Chimpanzee_2    TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGTACGTCAACTTGGTGCAGTTGGTG
Gorilla_1       TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGTACGTCAACTTGGTGCAGTTGGTG
Gorilla_2       TTGTCGGTGCTGATCTCAGTGGATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
Cem_1           ACGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGCGCAGTTGGTG
Cem_2           ACGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
Rhm_1           ATGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_a          TTGTCGGTGCTGATCTCAGTGGATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_b          TTGTCGGTGCTGATCTCAGTGGATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_c          TTGTCGGTGCTGATCTCAGTGGATGCTGTCTTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_d          TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGTACGTCAACTTGGTGCAGTTGGTG
node_e          TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGTACGTCAACTTGGTGCAGTTGGTG
node_f          ATGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_g          ACGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
                  ***  ***********  *  ******* ***** * ********** **********
```

```
Human_D         GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCAACCTGAGGATGGTCATCAGTAATATC
Human_CE        GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
Chimpanzee_1    GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCGTGAGGATGGTCATCAGTAATATC
Chimpanzee_2    GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
Gorilla_1       GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
Gorilla_2       GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
Cem_1           GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCTATAATATC
Cem_2           GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCTATAATATC
Rhm_1           GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCAATAATATC
node_a          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_b          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_c          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_d          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_e          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_f          GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCAATAATATC
node_g          GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCTATAATATC
                ******* ********** ******  ** **** * ******* ******  *******

Human_D         TTCAACACAGACTACCACATGAACATGATGCACATCTACGTGTTCGCAGCCTATTTTGGG
Human_CE        TTCAACACAGACTACCACATGAACCTGAGGCACTTCTACGTGTTCGCAGCCTATTTTGGG
Chimpanzee_1    TTCAATACAGACTACCACATGAACCTGATGCACATCTACGTGTTTGCAGCCTATTTTGGG
Chimpanzee_2    TTCAACACAGACTACCACATGAACATGATGCACTTCTACCTGTTCACAGCCTATTTTGGG
Gorilla_1       TTCAACACAGACTACCACATGAACATGACGCACTTCTACGTGTTCGCAGCCTATTTTGGG
Gorilla_2       TTCAACACAGACTACCACATGAACATGATGCACATCTACGTGTTCGCAGCCTGTTTTGGG
Cem_1           TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
Cem_2           TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
Rhm_1           TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
node_a          TTCAACACAGACTACCACATGAACATGATGCACATCTACGTGTTCGCAGCCTATTTTGGG
node_b          TTCAACACAGACTACCACATGAACATGATGCACATCTACGTGTTCGCAGCCTATTTTGGG
node_c          TTCAACACAGACTACCACATGAACATGATGCACATCTACGTGTTCGCAGCCTATTTTGGG
node_d          TTCAACACAGACTACCACATGAACATGATGCACTTCTACGTGTTCGCAGCCTATTTTGGG
node_e          TTCAACACAGACTACCACATGAACATGATGCACTTCTACGTGTTCGCAGCCTATTTTGGG
node_f          TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
node_g          TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
                ***** * *******  ******* *** **** ** ** ****  ****** *******

Human_D         CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCGAGGGAACGGAGGATAAAGATCAG
Human_CE        CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAACGTAAAGGAGGATAATGATCAG
Chimpanzee_1    CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAAAGATCAG
Chimpanzee_2    GTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCGACGTAAAGGAGGATAAAGATCAG
Gorilla_1       GTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCGACATAAAGGAGGATAAAGATCAG
Gorilla_2       CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTAGCCAAGGGAACGGAGGATAAAGATCAG
Cem_1           CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
Cem_2           CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
Rhm_1           CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
node_a          CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAAAGATCAG
node_b          CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAAAGATCAG
node_c          CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAAAGATCAG
node_d          CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAACGTAAAGGAGGATAAAGATCAG
node_e          CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAACGTAAAGGAGGATAAAGATCAG
node_f          CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
node_g          CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
                 ** ****************************** ** *   **  ********  *****

Human_D         ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Human_CE        AGAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Chimpanzee_1    ATAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Chimpanzee_2    ATAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Gorilla_1       ATAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
Gorilla_2       ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Cem_1           ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
Cem_2           ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
Rhm_1           ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
node_a          ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_b          ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_c          ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_d          ATAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_e          ATAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_f          ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
node_g          ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
                * * ******  ********** ************** ************************
```

```
Human_D         CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Human_CE        CCAAGTGTCAACTCTCCTCTGCTGAGAAGTCCAATCCAAAGGAAGAATGCCATGTTCAAC
Chimpanzee_1    CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Chimpanzee_2    CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Gorilla_1       CCAAGTTTCAATTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Gorilla_2       CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Cem_1           CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAATGCCGTGTTCAGC
Cem_2           CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAATGCCGTGTTCAGC
Rhm_1           CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAACGCCGTGTTCAGC
node_a          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_b          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_c          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_d          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_e          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_f          CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAATGCCGTGTTCAGC
node_g          CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAATGCCGTGTTCAGC
                **** * **** *** ********  ** ******* ********** *** ****** *

Human_D         ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGNNNNNNNNNNNNNNNNN
Human_CE        ACCTACTATGCTCTAGCAGTCAGTGTGGTGACAGCCATCTCAGNNNNNNNNNNNNNNNNN
Chimpanzee_1    ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGNNNNNNNNNNNNNNNNN
Chimpanzee_2    ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGNNNNNNNNNNNNNNNNN
Gorilla_1       ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGTGTCATCCTTGGCTCAC
Gorilla_2       ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGTGTCATCCTTGGCTCAC
Cem_1           ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
Cem_2           ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
Rhm_1           ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
node_a          ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGTGTCATCCTTGGCTCAC
node_b          ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGTGTCATCCTTGGCTCAC
node_c          ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGNNNNNNNNNNNNNNNNN
node_d          ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGTGTCATCCTTGGCTCAC
node_e          ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGNNNNNNNNNNNNNNNNN
node_f          ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
node_g          ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
                ************ ********** * *** **************

Human_D         NNNNNNNNNNNNNNNNCAAGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
Human_CE        NNNNNNNNNNNNNNNNCATGACTTATGTGCACAGTGCGGTGTTGGCAGGAGGCGTGGCT
Chimpanzee_1    NNNNNNNNNNNNNNNNCATGAGTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
Chimpanzee_2    NNNNNNNNNNNNNNNNCATGACTTATGTGCACAGTGCGGTGTTGGCAGGAGGCGTGGCT
Gorilla_1       CCCCAAGGGAAGATCAACATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
Gorilla_2       CCCCAAGGGAAGATCAACATGACTTATATGCACAATGCAGTGTTGGCAGGAGGTGTGGCT
Cem_1           CCCCAAAGGAAGATCAACATGACTTATATGCCCAATGCAGGGTTGGCAGGAGGCGTGGCT
Cem_2           CCCCGAAGGAAGATCAACATGACTTATATGCACAATGCAGCGTTGGCAGGAGGCGTGGTT
Rhm_1           CCCGGAGGGAAGATCAACATGACTTATATGCACAATGCAGCGTTGGCAGGAGGTGTGGCT
node_a          CCCCAAGGGAAGATCAACATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
node_b          CCCCAAGGGAAGATCAACATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
node_c          NNNNNNNNNNNNNNNNNCATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
node_d          CCCCAAGGGAAGATCAACATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
node_e          NNNNNNNNNNNNNNNNNCATGACTTATGTGCACAGTGCGGTGTTGGCAGGAGGCGTGGCT
node_f          CCCCGAGGGAAGATCAACATGACTTATATGCACAATGCAGCGTTGGCAGGAGGCGTGGCT
node_g          CCCCGAAGGAAGATCAACATGACTTATATGCACAATGCAGCGTTGGCAGGAGGCGTGGCT
                               ** ** **** *** ** *** * ************ **** *

Human_D         GTGGGTACCTCATGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Human_CE        GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Chimpanzee_1    GTGGGTACCTCATGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Chimpanzee_2    GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Gorilla_1       GTGGGTACCTCATGTCACCTGATTACTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Gorilla_2       GTGGCTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTCGGTCTTGTG
Cem_1           GTAGGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTGGGTCTTGTG
Cem_2           GTGAGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTGGGTCTTGTG
Rhm_1           CTGAGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTAGGTCTTGTG
node_a          GTGGGTACCTCATGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_b          GTGGGTACCTCATGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_c          GTGGGTACCTCATGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_d          GTGGGTACCTCATGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_e          GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_f          GTGAGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTGGGTCTTGTG
node_g          GTGAGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTGGGTCTTGTG
                 *   * **** ****** ****   ****** *** ************** *********
```

```
Human_D         GCTGGGCTGATCTCCGTCGGGGGAGCCAAGTACCTGCCGGGGTGTTGTAACCGAGTGCTG
Human_CE        GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGTGTGTTGTAACCGAGTGCTG
Chimpanzee_1    GCTGGGCTGATCTCCGTCGGGGGAGCCAAGTACTTGCCGGGGTGTTGTAACCGAGTGCTG
Chimpanzee_2    GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCGGGCCGGGGTGTTGTAACCGAGTGCTG
Gorilla_1       GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
Gorilla_2       GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
Cem_1           GCTGGGCTGATCTCCTTCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
Cem_2           GCTGGGCTGATCTCCTTCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
Rhm_1           GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
node_a          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
node_b          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
node_c          GCTGGGCTGATCTCCGTCGGGGGAGCCAAGTACCTGCCGGGGTGTTGTAACCGAGTGCTG
node_d          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
node_e          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
node_f          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
node_g          GCTGGGCTGATCTCCTTCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
                *************** *************** *  ***** ***** ************

Human_D         GGGATTCCCCACAGCTCCATCATGGGCTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Human_CE        GGGATTCACCACATCTCCGTCATGCACTCCATCTTCAGCTTGCTGGGTCTGCTTGGAGAG
Chimpanzee_1    GGGATTCCCCACAGCTCCGTCATGGGCTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Chimpanzee_2    GGGATTCCCGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Gorilla_1       GGGATTCATGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Gorilla_2       GGGATTCATGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Cem_1           GGGATTCACGAGAGCCACAGCATGCACTACACCTTCGGCTTGCCGGCTCTGCTTGGAGAG
Cem_2           GGGATTCACGAGAGCCACAGCATGCACTACACCTTCGGCTTGCCGGGTCTGCTTGGAGAG
Rhm_1           GGGATTCACGAGAGCCACAGCGTGCACTACACCTTCGGCTTGCCGGCTCTGCTTGGAGAG
node_a          GGGATTCACGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_b          GGGATTCACGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_c          GGGATTCCCCACAGCTCCGTCATGGGCTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_d          GGGATTCACGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_e          GGGATTCACGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_f          GGGATTCACGAGAGCCACAGCATGCACTACACCTTCGGCTTGCCGGCTCTGCTTGGAGAG
node_g          GGGATTCACGAGAGCCACAGCATGCACTACACCTTCGGCTTGCCGGCTCTGCTTGGAGAG
                *******   * * *  *  * **  ** ** **** ****** ** ************

Human_D         ATCATCTACATTGTGCTGCTGGTGCTTGATACCGTCGGAGCCGGCAATGGCATGATTGGC
Human_CE        ATCACCTACATTGTGCTGCTGGTGCTTCATACTGTCTGGAACGGCAATGGCATGATTGGC
Chimpanzee_1    ATCATCTACATTGTGCTGCTGGTGCTTGATACCGTCGGAGCCGGCAATGGCATGATTGGC
Chimpanzee_2    ATCATCTACATTGTGCTGGTGGTGCGTCATACCGTCTGGAACGGCAATGGCATGATTGGC
Gorilla_1       ATCACCTACATTGTGCTGATGGTGCTTCATACCGTCTGGGCTGGCAATGGCATGNNNNNN
Gorilla_2       ATCATCTACATTGTACTGCTGGTGCTTGATACCGTCGGAGCCGGCAATGGCATGNNNNNN
Cem_1           ATCACCTACATTGTGCTGATGGCGCTTCGTGTCTTCTGGGCCAGCAGTAACATGATCGGC
Cem_2           ATCACCTACATTGTGCTGATGGGGCTTCGTGTCTTCTGGGCCAGCAGTAACATGATCGGC
Rhm_1           ATCACCTACATTGTGCTGATGGCGCTTCGTGTCGTCTGGGCCAGCAGTAACATGATCGGC
node_a          ATCACCTACATTGTGCTGCTGGTGCTTCATACCGTCTGGGCCGGCAATGGCATGATTGGC
node_b          ATCATCTACATTGTGCTGCTGGTGCTTGATACCGTCGGAGCCGGCAATGGCATGATTGGC
node_c          ATCATCTACATTGTGCTGCTGGTGCTTGATACCGTCGGAGCCGGCAATGGCATGATTGGC
node_d          ATCACCTACATTGTGCTGCTGGTGCTTCATACCGTCTGGGCCGGCAATGGCATGATTGGC
node_e          ATCACCTACATTGTGCTGCTGGTGCTTCATACCGTCTGGAACGGCAATGGCATGATTGGC
node_f          ATCACCTACATTGTGCTGATGGCGCTTCGTGTCGTCTGGGCCAGCAGTAACATGATCGGC
node_g          ATCACCTACATTGTGCTGATGGCGCTTCGTGTCTTCTGGGCCAGCAGTAACATGATCGGC
                **** ********* *** *** ** *  *    ** *     *** *  ****

Human_D         TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATGGCGATAGCTCTCACGTCT
Human_CE        TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCT
Chimpanzee_1    TTCCAGGTCCTCCTCCGCATTGGGGAATTCAGCTTGGCCACGACGATAGCTCTCACGTCT
Chimpanzee_2    TTCCAGGTCCTCCTCAGCATGGGGGAACTCAGCTTGGCCATCGCGATAGCTCTCACGTCT
Gorilla_1       NNNNNNNNNNNNNNTCAGCACTGGGGAACTCAGCTTGGCCTTGGCGATAGCTGTCACGTCT
Gorilla_2       NNNNNNNNNNNNNNTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCT
Cem_1           TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
Cem_2           TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
Rhm_1           TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
node_a          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATGGCGATAGCTCTCACGTCT
node_b          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATGGCGATAGCTCTCACGTCT
node_c          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATGGCGATAGCTCTCACGTCT
node_d          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATGGCGATAGCTCTCACGTCT
node_e          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGCGATAGCTCTCACGTCT
node_f          TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
node_g          TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
                          ** ***  ***  * ***********     ***   * **** ***
```

```
Human_D          GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGAGGCTAAA
Human_CE         GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
Chimpanzee_1     GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGAGGCTAAA
Chimpanzee_2     GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
Gorilla_1        GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
Gorilla_2        GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGCGGCTAAA
Cem_1            GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
Cem_2            GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
Rhm_1            GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
node_a           GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
node_b           GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGTGGCTAAA
node_c           GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGAGGCTAAA
node_d           GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
node_e           GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
node_f           GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
node_g           GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
                 ******************** ******** ************ ******** *******

Human_D          TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Human_CE         TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Chimpanzee_1     TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Chimpanzee_2     TATTCTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Gorilla_1        TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Gorilla_2        TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Cem_1            TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
Cem_2            TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
Rhm_1            TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
node_a           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_b           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_c           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_d           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_e           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_f           TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
node_g           TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
                 **** ***********  ****** **************************
```

(B) Non-converted and ancestral sequences for tree B

```
Human_D         ATGAGCTCTAAGTACCCGCGGTCTGTCCGGCGCTGCCTGCCCCTCTGCGCCCTAACACTG
Human_CE        ATGAGCTCTAAGTACCCGCGGTCTGTCCGGCGCTGCCTGCCCCTCTGGGCCCTAACACTG
Chimpanzee_1    ATGAGCTCTAAGTACCCGCGGTCTGTCCGGCGCTGCCTGCCCCTCTGCGCCCTAACACTG
Chimpanzee_2    ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
Gorilla_1       ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
Gorilla_2       ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
Cem_1           ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACTCTG
Cem_2           ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACTCTG
Rhm_1           ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACACTG
node_a          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACACTG
node_b          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
node_c          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
node_d          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGCGCTGCCTGCCCCTCTGCGCCCTAACACTG
node_e          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGCGCCCTAACACTG
node_f          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACACTG
node_g          ATGAGCTCTAAGTACCCGCGGTCTGTCCGGTGCTGCCTGCCCCTCTGGGCCCTAACTCTG
                ******************************** ********** ******** ***

Human_D         GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
Human_CE        GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTATGACGCTTCCTTAGAGGAT
Chimpanzee_1    GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCAGTACGATGCTTCCTTAGAGGAT
Chimpanzee_2    GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
Gorilla_1       GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
Gorilla_2       GAAGCAGCTCTCACTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
Cem_1           GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
Cem_2           GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
Rhm_1           GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
node_a          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_b          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_c          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_d          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_e          GAAGCAGCTCTCATTCTCCTCTTCTATTTTTTTACCCACTACGACGCTTCCTTAGAGGAT
node_f          GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
node_g          GAAGCAGCTCTCATTCTCCTCTTCTTTTTTTTTACCTACTACGACGCTTCCTTAGAGGAT
                ************* *********** ********** * ** ** ***************

Human_D         CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Human_CE        CAAAAGGGGCTCGTGGCATCCTATCAAGTCGGCCAAGATCTGACCGTGATGGCGGCCCTT
Chimpanzee_1    CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Chimpanzee_2    CAAAAGGGGCTCGTGGCGTCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Gorilla_1       CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Gorilla_2       CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
Cem_1           CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
Cem_2           CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
Rhm_1           CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
node_a          CAAAAGGGGCTCGTGGCATCCTATCAAGTCGGCCAAGATCTGACCGTGATGGCGGCCCTT
node_b          CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
node_c          CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
node_d          CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
node_e          CAAAAGGGGCTCGTGGCATCCTATCAAGTTGGCCAAGATCTGACCGTGATGGCGGCCATT
node_f          CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
node_g          CAAAAGGGGCTCGTGGCGTCCTATCAAGTCTGCCAAGATCTGACCGTGATGGCGGTCCTT
                ***************** ***********  ************************* * **
```

```
Human_D         GGCTTGGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
Human_CE        GGCTTGGGCTTCCTCACCTCAAATTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
Chimpanzee_1    GGCTTTGGCTTCCTTACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAGC
Chimpanzee_2    GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
Gorilla_1       GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
Gorilla_2       GGCTTTGGCTTCCTCACCTCGAGTTTCCGGGGACACAGCTGGAGCAGTGTGGCCTTCAAC
Cem_1           GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
Cem_2           GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
Rhm_1           GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
node_a          GGCTTGGGCTTCCTCACCTCGAATTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_b          GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_c          GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_d          GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_e          GGCTTTGGCTTCCTCACCTCGAGTTTCCGGAGACACAGCTGGAGCAGTGTGGCCTTCAAC
node_f          GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
node_g          GGCTTGGGCTTCTTCACCTCGAATTTGCGGAGAAACAGCTGGAGCAGTGTGGCCTTCAAC
                ***** ****** * ***** * *** *** ** ************************ *

Human_D         CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Human_CE        CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Chimpanzee_1    CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATTCTGCTGGACGGCTTCCTGAGCCAG
Chimpanzee_2    CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Gorilla_1       CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Gorilla_2       CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Cem_1           CTCTTCCTGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Cem_2           CTCTTCCTGCTCGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
Rhm_1           CTCTTCCTGCTGGCCCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_a          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_b          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_c          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_d          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_e          CTCTTCATGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_f          CTCTTCCTGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
node_g          CTCTTCCTGCTGGCGCTTGGTGTGCAGTGGGCAATCCTGCTGGACGGCTTCCTGAGCCAG
                ****** **** ** ********************** ************************

Human_D         TTCCCTTCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
Human_CE        TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
Chimpanzee_1    TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCACGAGTGCT
Chimpanzee_2    TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCACGAGTGCT
Gorilla_1       TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
Gorilla_2       TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
Cem_1           TTCTCTCCTGGGAAGGTGGTCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
Cem_2           TTCTCTCCTGGGAAGGTGGTCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
Rhm_1           TTCTCTCCTGGGAAGGTGGCCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
node_a          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_b          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_c          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_d          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_e          TTCCCTCCTGGGAAGGTGGTCATCACACTGTTCAGTATTCGGCTGGCCACCATGAGTGCT
node_f          TTCTCTCCTGGGAAGGTGGTCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
node_g          TTCTCTCCTGGGAAGGTGGTCATCAAACTGTTCAGTATTCGGCTGGCCACCAGGAGCACT
                *** ** ************ ***** ************************** ***  **

Human_D         TTGTCGGTGCTGATCTCAGTGGATGCTGTCTTGGGGAAGGTCAACTTGGCGCAGTTGGTG
Human_CE        ATGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
Chimpanzee_1    TTGTCAGTGCTGATCTCAGTGGATGCTGTCTTGGGGAAGGTCAACTTGGTGCAGTTGGTG
Chimpanzee_2    TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGTACGTCAACTTGGTGCAGTTGGTG
Gorilla_1       TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGTACGTCAACTTGGTGCAGTTGGTG
Gorilla_2       TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
Cem_1           ACGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGCGCAGTTGGTG
Cem_2           ACGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
Rhm_1           ATGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_a          ATGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_b          TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_c          TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_d          TTGTCGGTGCTGATCTCAGTGGATGCTGTCTTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_e          TTGTCGGTGCTGATCTCAGCGGGTGCTGTCCTGGGGTACGTCAACTTGGTGCAGTTGGTG
node_f          ATGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
node_g          ACGTCGATGCTGATCTCAATGAATGCTGTCCTGGGGAAGGTCAACTTGGTGCAGTTGGTG
                  ***  ***********  *  ******* ***** * ********** **********
```

```
Human_D         GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCAACCTGAGGATGGTCATCAGTAATATC
Human_CE        GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
Chimpanzee_1    GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCGTGAGGATGGTCATCAGTAATATC
Chimpanzee_2    GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
Gorilla_1       GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
Gorilla_2       GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
Cem_1           GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCTATAATATC
Cem_2           GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCTATAATATC
Rhm_1           GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCAATAATATC
node_a          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_b          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_c          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_d          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_e          GTGATGGTGCTGGTGGAGGTGACAGCTTTAGGCACCATGAGGATGGTCATCAGTAATATC
node_f          GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCAATAATATC
node_g          GTGATGGAGCTGGTGGAGCTGACAGTCTTTGGCACCATGAGGATAGTCATCTATAATATC
                ******* ********** ******  ** **** * ******* ******  *******

Human_D         TTCAACACAGACTACCACATGAACATGATGCACATCTACGTGTTCGCAGCCTATTTTGGG
Human_CE        TTCAACACAGACTACCACATGAACCTGAGGCACTTCTACGTGTTCGCAGCCTATTTTGGG
Chimpanzee_1    TTCAATACAGACTACCACATGAACCTGATGCACATCTACGTGTTTGCAGCCTATTTTGGG
Chimpanzee_2    TTCAACACAGACTACCACATGAACATGATGCACTTCTACCTGTTCACAGCCTATTTTGGG
Gorilla_1       TTCAACACAGACTACCACATGAACATGACGCACTTCTACGTGTTCGCAGCCTATTTTGGG
Gorilla_2       TTCAACACAGACTACCACATGAACATGATGCACATCTACGTGTTCGCAGCCTGTTTTGGG
Cem_1           TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
Cem_2           TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
Rhm_1           TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
node_a          TTCAACACAGACTACCACATGAACATGATGCACTTCTACGTGTTCGCAGCCTATTTTGGG
node_b          TTCAACACAGACTACCACATGAACATGATGCACTTCTACGTGTTCGCAGCCTATTTTGGG
node_c          TTCAACACAGACTACCACATGAACATGATGCACATCTACGTGTTCGCAGCCTATTTTGGG
node_d          TTCAACACAGACTACCACATGAACATGATGCACATCTACGTGTTCGCAGCCTATTTTGGG
node_e          TTCAACACAGACTACCACATGAACATGATGCACTTCTACGTGTTCGCAGCCTATTTTGGG
node_f          TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
node_g          TTCAAAATAGACTACGGCATGAACATGATGCACATCCACGTGTTCGCAGCCTATTTTGGG
                ***** * *******  ******* *** **** ** ** ****  ****** *******

Human_D         CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCGAGGGAACGGAGGATAAAGATCAG
Human_CE        CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAATGATCAG
Chimpanzee_1    CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAAAGATCAG
Chimpanzee_2    GTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCGACGTAAAGGAGGATAAAGATCAG
Gorilla_1       GTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCGACATAAAGGAGGATAAAGATCAG
Gorilla_2       CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTAGCCAAGGGAACGGAGGATAAAGATCAG
Cem_1           CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
Cem_2           CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
Rhm_1           CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
node_a          CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAAAGATCAG
node_b          CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAAAGATCAG
node_c          CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAAAGATCAG
node_d          CTGTCTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACGGAGGATAAAGATCAG
node_e          GTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCGACGTAAAGGAGGATAAAGATCAG
node_f          CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
node_g          CTGACTGTGGCCTGGTGCCTGCCAAAGCCTCTACCCAAGGGAACAGAGGATAAATATCAG
                 ** ******************************** ** *   **  ********  *****

Human_D         ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Human_CE        AGAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Chimpanzee_1    ATAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Chimpanzee_2    ATAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Gorilla_1       ATAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
Gorilla_2       ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
Cem_1           ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
Cem_2           ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
Rhm_1           ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
node_a          ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_b          ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_c          ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_d          ACAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_e          ATAGCAACGATACCCAGTTTGTCTGCCATGCTGGGCGCCCTCTTCTTGTGGATGTTCTGG
node_f          ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
node_g          ACAACAACGAGCCCCAGTTTGTTTGCCATGCTGGGCACCCTCTTCTTGTGGATGTTCTGG
                * * ******  ********** ************** ************************
```

```
Human_D         CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Human_CE        CCAAGTGTCAACTCTCCTCTGCTGAGAAGTCCAATCCAAAGGAAGAATGCCATGTTCAAC
Chimpanzee_1    CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Chimpanzee_2    CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Gorilla_1       CCAAGTTTCAATTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Gorilla_2       CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
Cem_1           CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAATGCCGTGTTCAGC
Cem_2           CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAATGCCGTGTTCAGC
Rhm_1           CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAACGCCGTGTTCAGC
node_a          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_b          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_c          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_d          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_e          CCAAGTTTCAACTCTGCTCTGCTGAGAAGTCCAATCGAAAGGAAGAATGCCGTGTTCAAC
node_f          CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAATGCCGTGTTCAGC
node_g          CCAACTTTCAACTCTGCTCTGCTGCTAAATCCAATCGAAAGGAAGAATGCCGTGTTCAGC
                **** * **** *** ********  ** ******* ********** *** ****** *

Human_D         ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGGGTCATCCTTGGCTCAC
Human_CE        ACCTACTATGCTCTAGCAGTCAGTGTGGTGACAGCCATCTCAGGGTCATCCTTGGCTCAC
Chimpanzee_1    ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGGGTCATCCTTGGCTCAC
Chimpanzee_2    ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGGGTCATCCTTGGCTCAC
Gorilla_1       ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGNNNNNNNNNNNNNNNNN
Gorilla_2       ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGNNNNNNNNNNNNNNNNN
Cem_1           ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
Cem_2           ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
Rhm_1           ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
node_a          ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGGGTCATCCTTGGCTCAC
node_b          ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGGGTCATCCTTGGCTCAC
node_c          ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGGGTCATCCTTGGCTCAC
node_d          ACCTACTATGCTGTAGCAGTCAGCGTGGTGACAGCCATCTCAGGGTCATCCTTGGCTCAC
node_e          ACCTACTATGCTCTAGCAGTCAGCGTGGTGACAGCCATCTCAGGGTCATCCTTGGCTCAC
node_f          ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
node_g          ACCTACTATGCTCTAGCAGTCAGCGCGGTTACAGCCATCTCAGTGTCATCCTTGGCTCAC
                ************ ********** * *** ************

Human_D         CCCCAAGGGAAGATCAGCAAGACTTATNNNNNNNNTGCGGTGTTGGCAGGAGGCGTGGCT
Human_CE        CCCCAAAGGAAGATCAGCATGACTTATNNNNNNNNTGCGGTGTTGGCAGGAGGCGTGGCT
Chimpanzee_1    CCCCAAGGGAAGATCAGCATGAGTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
Chimpanzee_2    CCCCAAGGGAAGATCAGCATGACTTATGTGCACAGTGCGGTGTTGGCAGGAGGCGTGGCT
Gorilla_1       NNNNNNNNNNNNNNNNNCATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
Gorilla_2       NNNNNNNNNNNNNNNNNCATGACTTATATGCACAATGCAGTGTTGGCAGGAGGTGTGGCT
Cem_1           CCCCAAAGGAAGATCAACATGACTTATATGCCCAATGCAGGGTTGGCAGGAGGCGTGGCT
Cem_2           CCCCGAAGGAAGATCAACATGACTTATATGCACAATGCAGCGTTGGCAGGAGGCGTGGTT
Rhm_1           CCCGGAGGGAAGATCAACATGACTTATATGCACAATGCAGCGTTGGCAGGAGGTGTGGCT
node_a          CCCCAAGGGAAGATCAGCATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
node_b          CCCCAAGGGAAGATCAGCATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
node_c          CCCCAAGGGAAGATCAGCATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
node_d          CCCCAAGGGAAGATCAGCATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
node_e          CCCCAAGGGAAGATCAGCATGACTTATATGCACAATGCGGTGTTGGCAGGAGGCGTGGCT
node_f          CCCCGAGGGAAGATCAACATGACTTATATGCACAATGCAGCGTTGGCAGGAGGCGTGGCT
node_g          CCCCGAAGGAAGATCAACATGACTTATATGCACAATGCAGCGTTGGCAGGAGGCGTGGCT
                          ** ** ****        *** * ************ **** *

Human_D         GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Human_CE        GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Chimpanzee_1    GTGGGTACCTCATGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Chimpanzee_2    GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Gorilla_1       GTGGGTACCTCATGTCACCTGATTACTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
Gorilla_2       GTGGCTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTCGGTCTTGTG
Cem_1           GTAGGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTGGGTCTTGTG
Cem_2           GTGAGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTGGGTCTTGTG
Rhm_1           CTGAGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTAGGTCTTGTG
node_a          GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_b          GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_c          GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_d          GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_e          GTGGGTACCTCGTGTCACCTGATCCCTTCTCCGTGGCTTGCCATGGTGCTGGGTCTTGTG
node_f          GTGAGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTGGGTCTTGTG
node_g          GTGAGTGCCTCATGTCACGTGATCCATTCTCCTTGGATTGCCATGGTGCTGGGTCTTGTG
                 *   * **** ****** ****   ****** *** ************* *********
```

```
Human_D         GCTGGGCTGATCTCCGTCGGGGGAGCCAAGTACCTGCCGGGGTGTTGTAACCGAGTGCTG
Human_CE        GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGTGTGTTGTAACCGAGTGCTG
Chimpanzee_1    GCTGGGCTGATCTCCGTCGGGGGAGCCAAGTACTTGCCGGGGTGTTGTAACCGAGTGCTG
Chimpanzee_2    GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCGGGCCGGGGTGTTGTAACCGAGTGCTG
Gorilla_1       GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
Gorilla_2       GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
Cem_1           GCTGGGCTGATCTCCTTCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
Cem_2           GCTGGGCTGATCTCCTTCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
Rhm_1           GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
node_a          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGTGTGTTGTAACCGAGTGCTG
node_b          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
node_c          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
node_d          GCTGGGCTGATCTCCGTCGGGGGAGCCAAGTACCTGCCGGGGTGTTGTAACCGAGTGCTG
node_e          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGGGTGTTGTAACCGAGTGCTG
node_f          GCTGGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
node_g          GCTGGGCTGATCTCCTTCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTG
                *************** *************** *  ***** ***** *************

Human_D         GGGATTCCCCACAGCTCCATCATGGGCTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Human_CE        GGGATTCACCACATCTCCGTCATGCACTCCATCTTCAGCTTGCTGGGTCTGCTTGGAGAG
Chimpanzee_1    GGGATTCCCCACAGCTCCGTCATGGGCTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Chimpanzee_2    GGGATTCCCGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Gorilla_1       GGGATTCATGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Gorilla_2       GGGATTCATGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
Cem_1           GGGATTCACGAGAGCCACAGCATGCACTACACCTTCGGCTTGCCGGCTCTGCTTGGAGAG
Cem_2           GGGATTCACGAGAGCCACAGCATGCACTACACCTTCGGCTTGCCGGGTCTGCTTGGAGAG
Rhm_1           GGGATTCACGAGAGCCACAGCGTGCACTACACCTTCGGCTTGCCGGCTCTGCTTGGAGAG
node_a          GGGATTCACGACAGCTCCGTCATGCACTACATCTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_b          GGGATTCACGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_c          GGGATTCACGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_d          GGGATTCCCCACAGCTCCGTCATGGGCTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_e          GGGATTCACGACAGCTCCGTCATGCACTACAACTTCAGCTTGCTGGGTCTGCTTGGAGAG
node_f          GGGATTCACGAGAGCCACAGCATGCACTACACCTTCGGCTTGCCGGCTCTGCTTGGAGAG
node_g          GGGATTCACGAGAGCCACAGCATGCACTACACCTTCGGCTTGCCGGCTCTGCTTGGAGAG
                *******   * * *  *  * **  ** ** **** ****** ** *************

Human_D         ATCATCTACATTGTGCTGCTGGTGCTTGATACCGTCGGAGCCGGCAATGGCATGATTGGC
Human_CE        ATCACCTACATTGTGCTGCTGGTGCTTCATACTGTCTGGAACGGCAATGGCATGATTGGC
Chimpanzee_1    ATCATCTACATTGTGCTGCTGGTGCTTCATACCGTCGGAGCCGGCAATGGCATGATTGGC
Chimpanzee_2    ATCATCTACATTGTGCTGGTGGTGCGTCATACCGTCTGGAACGGCAATGGCATGATTGGC
Gorilla_1       ATCACCTACATTGTGCTGATGGTGCTTCATACCGTCTGGGCTGGCAATGGCATGNNNNNN
Gorilla_2       ATCATCTACATTGTACTGCTGGTGCTTGATACCGTCGGAGCCGGCAATGGCATGNNNNNN
Cem_1           ATCACCTACATTGTGCTGATGGCGCTTCGTGTCTTCTGGGCCAGCAGTAACATGATCGGC
Cem_2           ATCACCTACATTGTGCTGATGGGGCTTCGTGTCTTCTGGGCCAGCAGTAACATGATCGGC
Rhm_1           ATCACCTACATTGTGCTGATGGCGCTTCGTGTCGTCTGGGCCAGCAGTAACATGATCGGC
node_a          ATCACCTACATTGTGCTGCTGGTGCTTCATACCGTCTGGGCCGGCAATGGCATGATTGGC
node_b          ATCATCTACATTGTGCTGCTGGTGCTTCATACCGTCTGGGCCGGCAATGGCATGATTGGC
node_c          ATCATCTACATTGTGCTGCTGGTGCTTCATACCGTCGGAGCCGGCAATGGCATGATTGGC
node_d          ATCATCTACATTGTGCTGCTGGTGCTTCATACCGTCGGAGCCGGCAATGGCATGATTGGC
node_e          ATCATCTACATTGTGCTGGTGGTGCTTCATACCGTCTGGGCCGGCAATGGCATGATTGGC
node_f          ATCACCTACATTGTGCTGATGGCGCTTCGTGTCGTCTGGGCCAGCAGTAACATGATCGGC
node_g          ATCACCTACATTGTGCTGATGGCGCTTCGTGTCTTCTGGGCCAGCAGTAACATGATCGGC
                **** ********* *** *** ** *  *    ** *     *** *  ****

Human_D         TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCT
Human_CE        TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCT
Chimpanzee_1    TTCCAGGTCCTCCTCCGCATTGGGGAATTCAGCTTGGCCACGACGATAGCTCTCACGTCT
Chimpanzee_2    TTCCAGGTCCTCCTCAGCATGGGGGAACTCAGCTTGGCCATCGCGATAGCTCTCACGTCT
Gorilla_1       NNNNNNNNNNNNNNTCAGCACTGGGGAACTCAGCTTGGCCTTGGCGATAGCTGTCACGTCT
Gorilla_2       NNNNNNNNNNNNNTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCT
Cem_1           TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
Cem_2           TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
Rhm_1           TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
node_a          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCT
node_b          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCT
node_c          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCT
node_d          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCT
node_e          TTCCAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGCGATAGCTCTCACGTCT
node_f          TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
node_g          TTCCAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCT
                          ** ***  ***  * ***********     ***   * **** ***
```

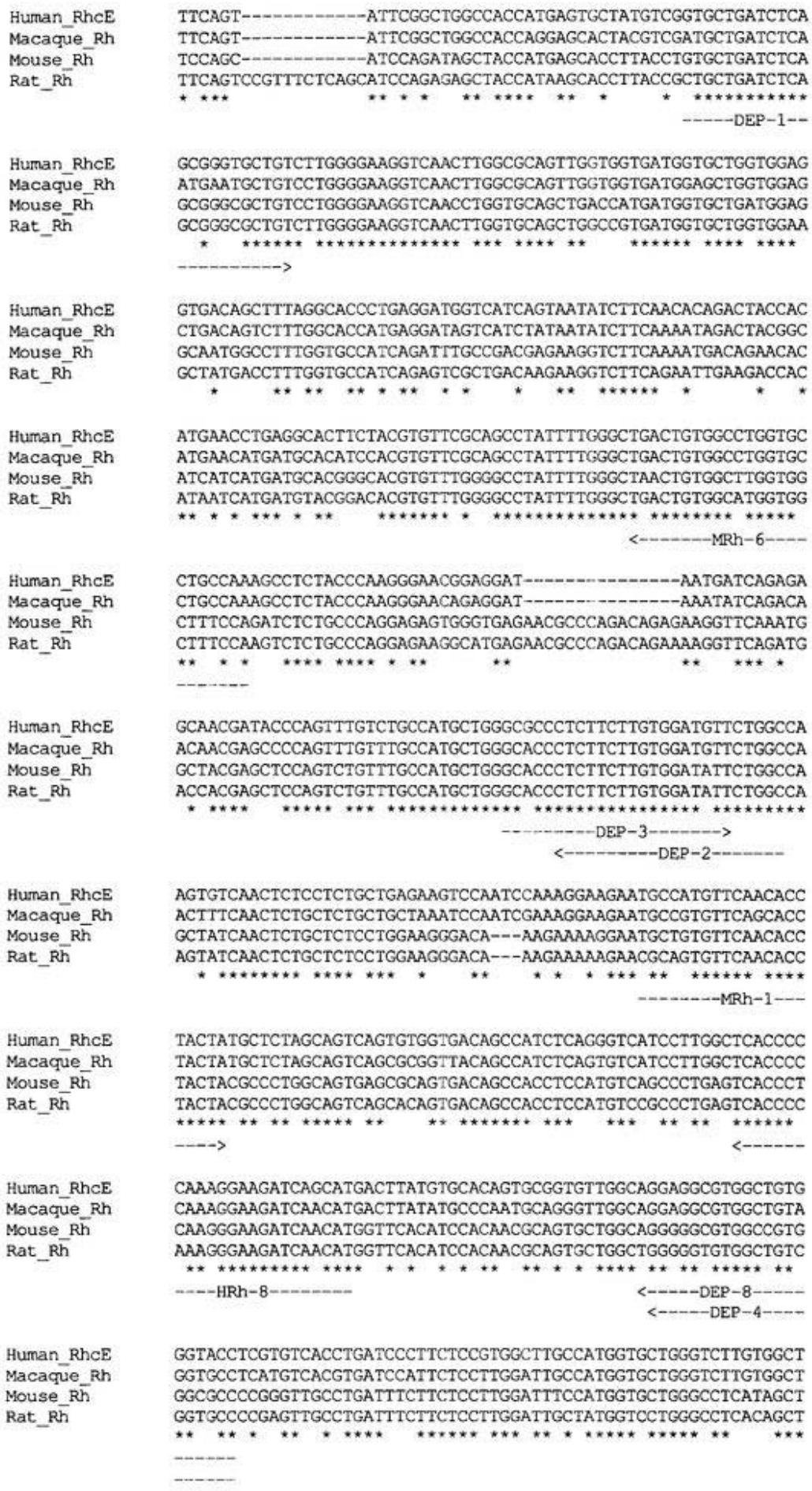
```
Human_D          GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGAGGCTAAA
Human_CE         GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
Chimpanzee_1     GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGAGGCTAAA
Chimpanzee_2     GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
Gorilla_1        GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
Gorilla_2        GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGCGGCTAAA
Cem_1            GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
Cem_2            GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
Rhm_1            GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
node_a           GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
node_b           GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
node_c           GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGTGGCTAAA
node_d           GGTCTCCTGACAGGTTTGCTCCTAAATCTTAAAATATGGAAAGCACCTCATGAGGCTAAA
node_e           GGTCTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAA
node_f           GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
node_g           GGTCTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAA
                 ******************** ******** ************* ******** *******

Human_D          TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Human_CE         TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Chimpanzee_1     TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Chimpanzee_2     TATTCTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Gorilla_1        TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Gorilla_2        TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
Cem_1            TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
Cem_2            TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
Rhm_1            TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
node_a           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_b           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_c           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_d           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_e           TATTTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTT
node_f           TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
node_g           TATTTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTT
                 **** ***********  ****** **************************
```

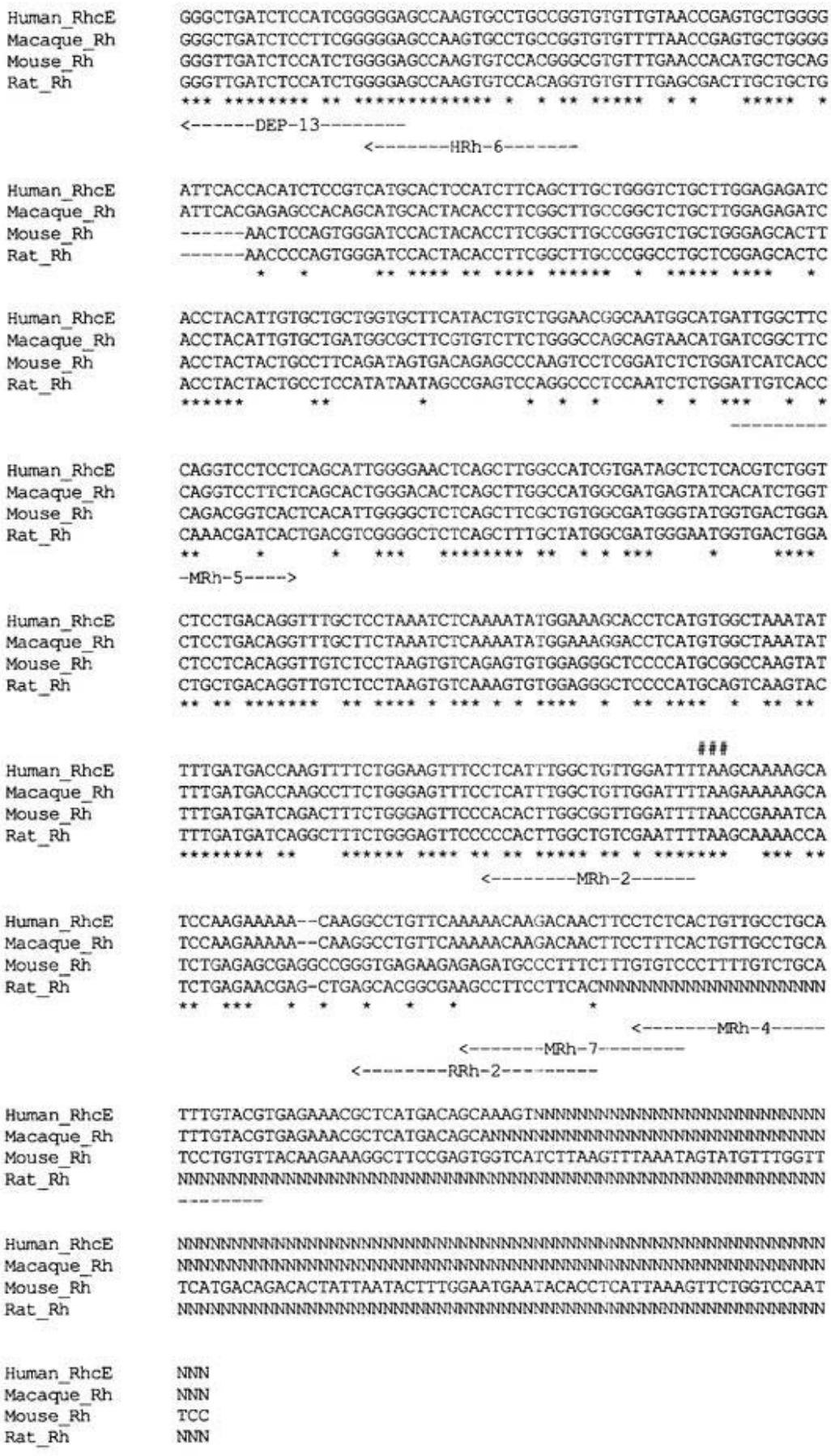
APPENDIX III

These are multiple alignments of Rh (A) and Rh50 (B) gene sequences, and degenerate primers (C). Primer sites are shown below alignments by dashes with an angled bracket. ### means start or stop codon. N means undetermined sites. HRh series: human Rh gene, MRh series: mouse Rh gene, RRh series: rat Rh gene, M50 series: mouse Rh50 gene, R50 series: rat Rh50 gene, X50 series: *xenopus* Rh50-like gene, O50 series: medaka Rh50-like gene, DEP series: Degenerate primers for Erythrocyte membrane Proteins.

(A) The multiple alignment of Rh genes

```
                                                    ###
Human_RhcE      AATCCCGGCCTGCACAGAGACGGACACAGGATGAGCTCTAAGTACCCGCGGTCTGTCCGG
Macaque_Rh      NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATGAGCTCTAAGTACCCGCGGTCTGTCCGG
Mouse_Rh        NNNNTACCCGGGCACAGCAACAGACACAAGATGGGCTCTAAGTACCCACGGTCCCTCCGC
Rat_Rh          NNNNNNNNCGGGCACAGCAACAGACACAAGATGGGCTCTAAGTACCCAAGGTCCCTCCGC
                                                *** ************** **** ****
                     --------MRh-3-------->                        ----------
                        --------RRh-3-------->                        --------

Human_RhcE      CGCTGCCTGCCCCTCTGGGCCCTAACACTGGAAGCAGCTCTCATTCTCCTCTTCTATTTT
Macaque_Rh      TGCTGCCTGCCCCTCTGGGCCCTAACTCTGGAAGCAGCTCTCATTCTCCTCTTCTTTTTT
Mouse_Rh        TGCTGCCTGCCCCTATGGGCCTTGGTGCTACAGACAGCTTTTATTCTCCTCTCCTGTTTT
Rat_Rh          TGCTGCCTGCCCCTGTGGGCCTTCGGGCTACAGGTGACTTTTATCCTCCTCTTCTATTTT
                 ************ ****** *    **  *     ** * ** ******* ** ****
                --DEP-12--->
                ---RRh-1----->

Human_RhcE      TTTACCCACTATGACGCTTCCTTAGAGGATCAAAAGGGGCTCGTGGCATCCTATCAAGTC
Macaque_Rh      TTTACCTACTACGACGCTTCCTTAGAGGATCAAAAGGGGCTCGTGGCGTCCTATCAAGTC
Mouse_Rh        TTCATCCCCCACGACACAGCCCAGGTGGATCACAAG---TTCATGGAGAGCTATCAAGTC
Rat_Rh          CTCATCGGCCAAGACCCTATCCAGGCAGATCACAAG---TTCATGGCGATCTATCAAGTC
                 * * *  * * *** *   *   *  ***** ***    ** ***    **********

Human_RhcE      GGCCAAGATCTGACCGTGATGGCGGCCCTTGGCTTGGGCTTCCTCACCTCAAATTTCCGG
Macaque_Rh      TGCCAAGATCTGACCGTGATGGCGGTCCTTGGCTTGGGCTTCTTCACCTCGAATTTGCGG
Mouse_Rh        CTCCGGAATTTGACCCTCATGGCAGCCTTGGGCTTCGGCTTCCTGTCCTCGTCCTTTCGG
Rat_Rh          ATCCAGGATTTGACCCTTGTGGCAGCCTTGGGTTTCGGCTTCCTGTCCTCATCCTTTCGG
                  **   ** ***** *  **** * * * ** ** ****** *  ****    ** ***
                              ----------DEP-9-------->
                                      -------DEP-5--->

Human_RhcE      AGACACAGCTGGAGCAGTGTGGCCTTCAACCTCTTCATGCTGGCGCTTGGTGTGCAGTGG
Macaque_Rh      AGAAACAGCTGGAGCAGTGTGGCCTTCAACCTCTTCCTGCTGGCGCTTGGTGTGCAGTGG
Mouse_Rh        AGACACAGCTGGAGCAGTGTGGCCTTCAACCTCTTCATGTTGGCCCTCGGGGTGCAGGGG
Rat_Rh          AGACACGGCTGGAGCAGTGTGGCCTTCAGCTTCTTCATGTTGGCCCTTGGGGTACAGGGG
                *** ** ********************** * ***** ** **** ** ** ** *** **
                 <-------HRh-11-------                                ----------
                       --------DEP-11------>

Human_RhcE      GCAATCCTGCTGGACGGCTTCCTGAGCCAGTTCCCTCCTGGGAAGGTGGTCATCACACTG
Macaque_Rh      GCAATCCTGCTGGACGGCTTCCTGAGCCAGTTCTCTCCTGGGAAGGTGGTCATCAAACTG
Mouse_Rh        ACAATCTTGCTGGACCATTTCCTGGGCCAGGTCCTCCAATGGAACAAGATCAACAATCTG
Rat_Rh          ACAATCTTGCTGGACTATTTCCTGAATTGGGTCCTCGACTGGAACATGATCAAGAATCCG
                 ***** ********   ******     * **       ****   * ***  *  * *
                --DEP-14----->
```

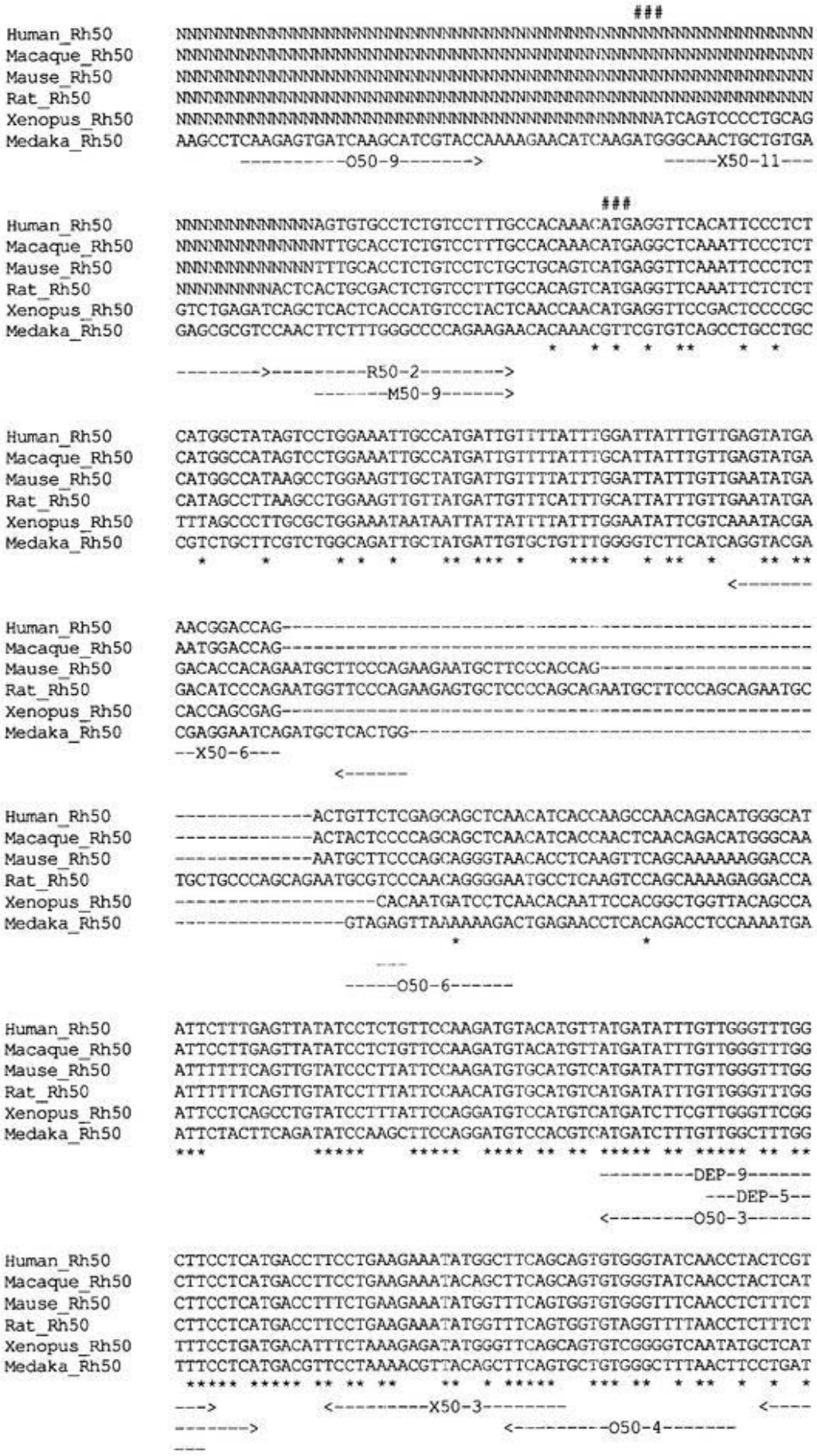

Human_RhcE
Macaque_Rh
Mouse_Rh
Rat_Rh
-----DEP-1--
---------->
<-------MRh-6----
---------DEP-3------->
<---------DEP-2-------
--------MRh-1---
---->
<------
----HRh-8--------
<-----DEP-8-----
<-----DEP-4----

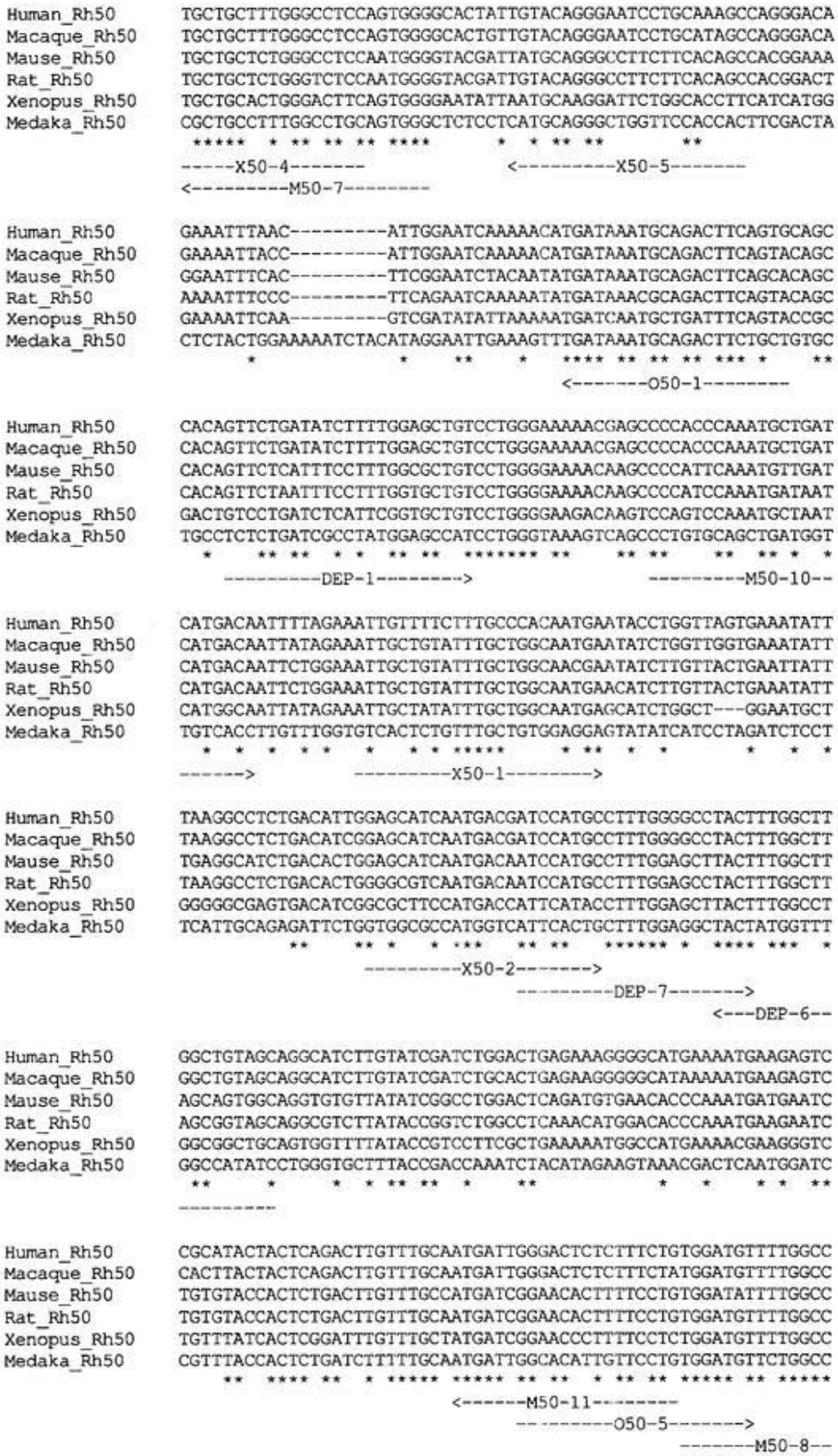

Human_RhcE GGGCTGATCTCCATCGGGGGAGCCAAGTGCCTGCCGGTGTGTTGTAACCGAGTGCTGGGG
Macaque_Rh GGGCTGATCTCCTTCGGGGGAGCCAAGTGCCTGCCGGTGTGTTTTAACCGAGTGCTGGGG
Mouse_Rh GGGTTGATCTCCATCTGGGGAGCCAAGTGTCCACGGGCGTGTTTGAACCACATGCTGCAG
Rat_Rh GGGTTGATCTCCATCTGGGGAGCCAAGTGTCCACAGGTGTGTTTGAGCGACTTGCTGCTG
<------DEP-13--------
<-------HRh-6-------
Human_RhcE ATTCACCACATCTCCGTCATGCACTCCATCTTCAGCTTGCTGGGTCTGCTTGGAGAGATC
Macaque_Rh ATTCACGAGAGCCACAGCATGCACTACACCTTCGGCTTGCCGGCTCTGCTTGGAGAGATC
Mouse_Rh ------AACTCCAGTGGGATCCACTACACCTTCGGCTTGCCGGGTCTGCTGGGAGCACTT
Rat_Rh ------AACCCCAGTGGGATCCACTACACCTTCGGCTTGCCCGGCCTGCTCGGAGCACTC
Human_RhcE ACCTACATTGTGCTGCTGGTGCTTCATACTGTCTGGAACGGCAATGGCATGATTGGCTTC
Macaque_Rh ACCTACATTGTGCTGATGGCGCTTCGTGTCTTCTGGGCCAGCAGTAACATGATCGGCTTC
Mouse_Rh ACCTACTACTGCCTTCAGATAGTGACAGAGCCCAAGTCCTCGGATCTCTGGATCATCACC
Rat_Rh ACCTACTACTGCCTCCATATAATAGCCGAGTCCAGGCCCTCCAATCTCTGGATTGTCACC

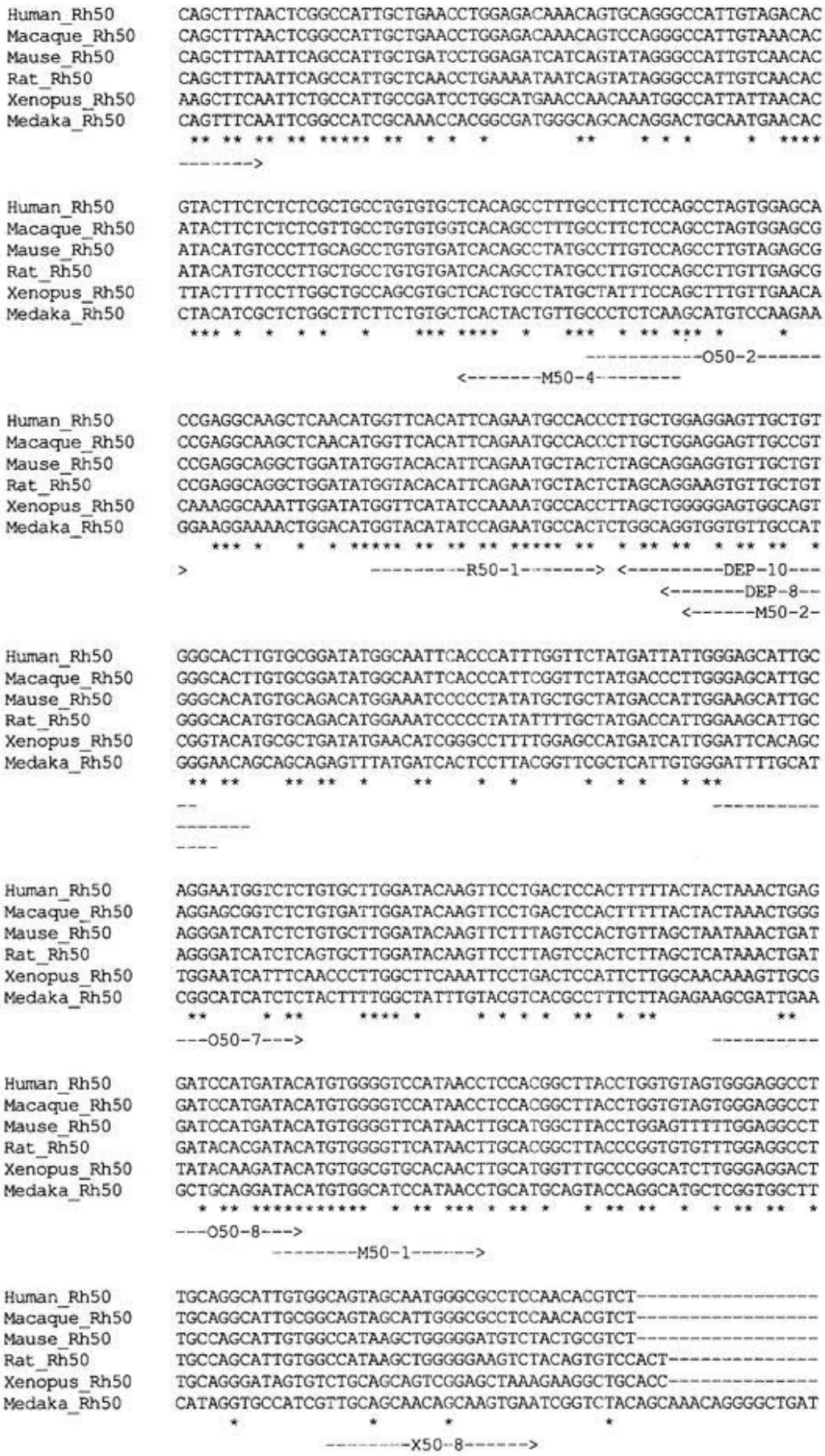
Human_RhcE CAGGTCCTCCTCAGCATTGGGGAACTCAGCTTGGCCATCGTGATAGCTCTCACGTCTGGT
Macaque_Rh CAGGTCCTTCTCAGCACTGGGACACTCAGCTTGGCCATGGCGATGAGTATCACATCTGGT
Mouse_Rh CAGACGGTCACTCACATTGGGGCTCTCAGCTTCGCTGTGGCGATGGGTATGGTGACTGGA
Rat_Rh CAAACGATCACTGACGTCGGGGCTCTCAGCTTTGCTATGGCGATGGGAATGGTGACTGGA
-MRh-5---->
Human_RhcE CTCCTGACAGGTTTGCTCCTAAATCTCAAAATATGGAAAGCACCTCATGTGGCTAAATAT
Macaque_Rh CTCCTGACAGGTTTGCTTCTAAATCTCAAAATATGGAAAGGACCTCATGTGGCTAAATAT
Mouse_Rh CTCCTCACAGGTTGTCTCCTAAGTGTCAGAGTGTGGAGGGCTCCCCATGCGGCCAAGTAT
Rat_Rh CTGCTGACAGGTTGTCTCCTAAGTGTCAAAGTGTGGAGGGCTCCCCATGCAGTCAAGTAC
###
Human_RhcE TTTGATGACCAAGTTTTCTGGAAGTTTCCTCATTTGGCTGTTGGATTTTAAGCAAAAGCA
Macaque_Rh TTTGATGACCAAGCCTTCTGGGAGTTTCCTCATTTGGCTGTTGGATTTTAAGAAAAAGCA
Mouse_Rh TTTGATGATCAGACTTTCTGGGAGTTCCCACACTTGGCGGTTGGATTTTAACCGAAATCA
Rat_Rh TTTGATGATCAGGCTTTCTGGGAGTTCCCCCACTTGGCTGTCGAATTTTAAGCAAAACCA
<--------MRh-2------
Human_RhcE TCCAAGAAAAA--CAAGGCCTGTTCAAAAACAAGACAACTTCCTCTCACTGTTGCCTGCA
Macaque_Rh TCCAAGAAAAA--CAAGGCCTGTTCAAAAACAAGACAACTTCCTTTCACTGTTGCCTGCA
Mouse_Rh TCTGAGAGCGAGGCCGGGTGAGAAGAGAGATGCCCTTTCTTTGTGTCCCTTTTGTCTGCA
Rat_Rh TCTGAGAACGAG-CTGAGCACGGCGAAGCCTTCCTTCACNNNNNNNNNNNNNNNNNNNN
<-------MRh-4-----
<-------MRh-7--------
<--------RRh-2---------
Human_RhcE TTTGTACGTGAGAAACGCTCATGACAGCAAAGTNNNNNNNNNNNNNNNNNNNNNNNNNNN
Macaque_Rh TTTGTACGTGAGAAACGCTCATGACAGCANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Mouse_Rh TCCTGTGTTACAAGAAAGGCTTCCGAGTGGTCATCTTAAGTTTAAATAGTATGTTTGGTT
Rat_Rh NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

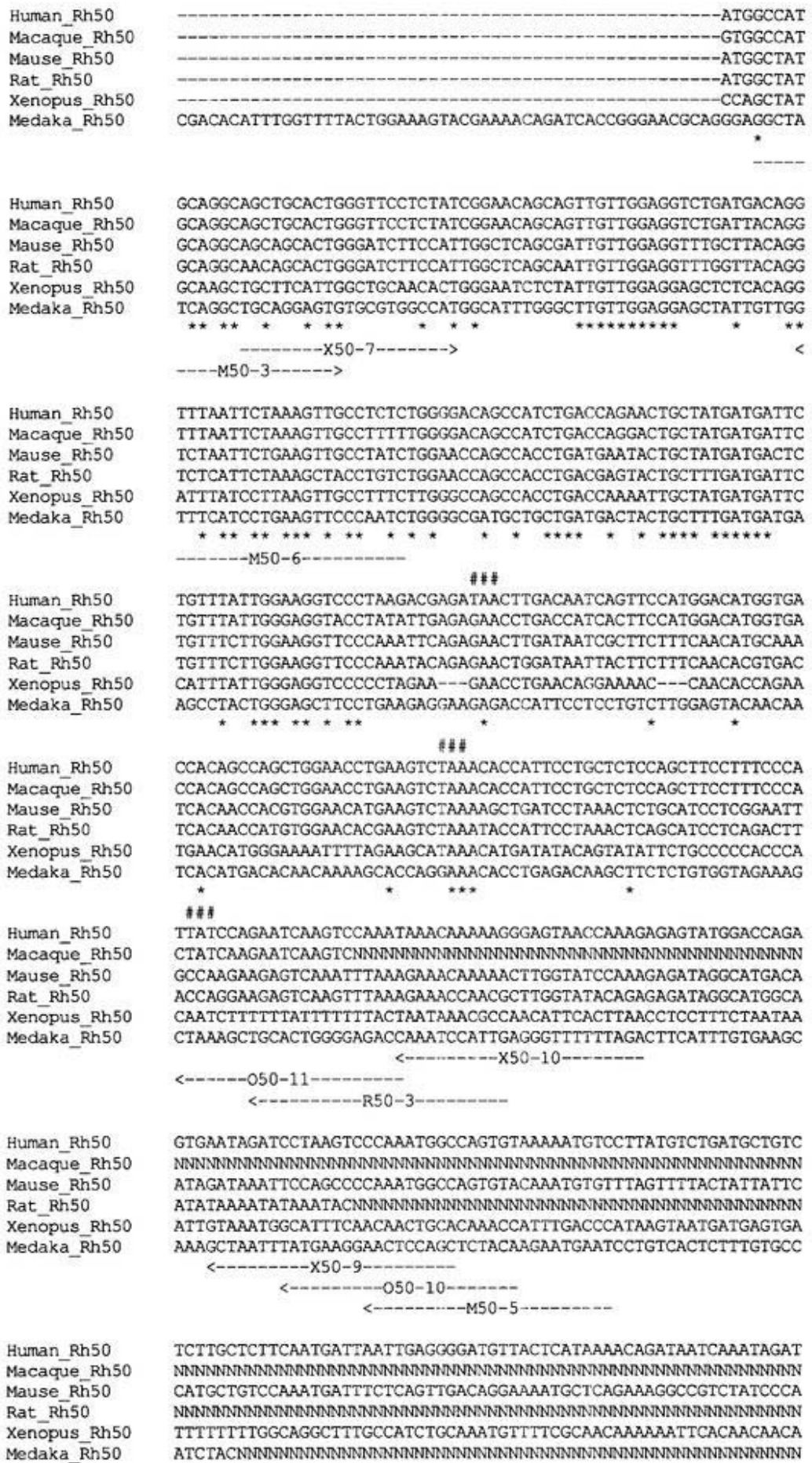
Human_RhcE NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Macaque_Rh NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Mouse_Rh TCATGACAGACACTATTAATACTTTGGAATGAATACACCTCATTAAAGTTCTGGTCCAAT
Rat_Rh NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Human_RhcE NNN
Macaque_Rh NNN
Mouse_Rh TCC
Rat_Rh NNN

(B) The multiple alignment of Rh50 genes




Human_Rh50 TGCTGCTTTGGGCCTCCAGTGGGGCACTATTGTACAGGGAATCCTGCAAAGCCAGGGACA
Macaque_Rh50 TGCTGCTTTGGGCCTCCAGTGGGGCACTGTTGTACAGGGAATCCTGCATAGCCAGGGACA
Mause_Rh50 TGCTGCTCTGGGCCTCCAATGGGGTACGATTATGCAGGGCCTTCTTCACAGCCACGGAAA
Rat_Rh50 TGCTGCTCTGGGTCTCCAATGGGGTACGATTGTACAGGGCCTTCTTCACAGCCACGGACT
Xenopus_Rh50 TGCTGCACTGGGACTTCAGTGGGGAATATTAATGCAAGGATTCTGGCACCTTCATCATGG
Medaka_Rh50 CGCTGCCTTTGGCCTGCAGTGGGCTCTCCTCATGCAGGGCTGGTTCCACCACTTCGACTA
X50-4
M50-7
X50-5
Human_Rh50 GAAATTTAAC---------ATTGGAATCAAAAACATGATAAATGCAGACTTCAGTGCAGC
Macaque_Rh50 GAAAATTACC---------ATTGGAATCAAAAACATGATAAATGCAGACTTCAGTACAGC
Mause_Rh50 GGAATTTCAC---------TTCGGAATCTACAATATGATAAATGCAGACTTCAGCACAGC
Rat_Rh50 AAAATTTCCC---------TTCAGAATCAAAAATATGATAAACGCAGACTTCAGTACAGC
Xenopus_Rh50 GAAAATTCAA---------GTCGATATATTAAAAATGATCAATGCTGATTTCAGTACCGC
Medaka_Rh50 CTCTACTGGAAAAATCTACATAGGAATTGAAAGTTTGATAAATGCAGACTTCTGCTGTGC
O50-1
Human_Rh50 CACAGTTCTGATATCTTTTGGAGCTGTCCTGGGAAAAACGAGCCCCACCCAAATGCTGAT
Macaque_Rh50 CACAGTTCTGATATCTTTTGGAGCTGTCCTGGGAAAAACGAGCCCCACCCAAATGCTGAT
Mause_Rh50 CACAGTTCTCATTTCCTTTGGCGCTGTCCTGGGGAAAACAAGCCCCATTCAAATGTTGAT
Rat_Rh50 CACAGTTCTAATTTCCTTTGGTGCTGTCCTGGGGAAAACAAGCCCCATCCAAATGATAAT
Xenopus_Rh50 GACTGTCCTGATCTCATTCGGTGCTGTCCTGGGGAAGACAAGTCCAGTCCAAATGCTAAT
Medaka_Rh50 TGCCTCTCTGATCGCCTATGGAGCCATCCTGGGTAAAGTCAGCCCTGTGCAGCTGATGGT
DEP-1
M50-10
Human_Rh50 CATGACAATTTTAGAAATTGTTTTCTTTGCCCACAATGAATACCTGGTTAGTGAAATATT
Macaque_Rh50 CATGACAATTATAGAAATTGCTGTATTTGCTGGCAATGAATATCTGGTTGGTGAAATATT
Mause_Rh50 CATGACAATTCTGGAAATTGCTGTATTTGCTGGCAACGAATATCTTGTTACTGAATTATT
Rat_Rh50 CATGACAATTCTGGAAATTGCTGTATTTGCTGGCAATGAACATCTTGTTACTGAAATATT
Xenopus_Rh50 CATGGCAATTATAGAAATTGCTATATTTGCTGGCAATGAGCATCTGGCT---GGAATGCT
Medaka_Rh50 TGTCACCTTGTTTGGTGTCACTCTGTTTGCTGTGGAGGAGTATATCATCCTAGATCTCCT
X50-1
Human_Rh50 TAAGGCCTCTGACATTGGAGCATCAATGACGATCCATGCCTTTGGGGCCTACTTTGGCTT
Macaque_Rh50 TAAGGCCTCTGACATCGGAGCATCAATGACGATCCATGCCTTTGGGGCCTACTTTGGCTT
Mause_Rh50 TGAGGCATCTGACACTGGAGCATCAATGACAATCCATGCCTTTGGAGCTTACTTTGGCTT
Rat_Rh50 TAAGGCCTCTGACACTGGGGCGTCAATGACAATCCATGCCTTTGGAGCCTACTTTGGCTT
Xenopus_Rh50 GGGGGCGAGTGACATCGGCGCTTCCATGACCATTCATACCTTTGGAGCTTACTTTGGCCT
Medaka_Rh50 TCATTGCAGAGATTCTGGTGGCGCCATGGTCATTCACTGCTTTGGAGGCTACTATGGTTT
X50-2
DEP-7
DEP-6
Human_Rh50 GGCTGTAGCAGGCATCTTGTATCGATCTGGACTGAGAAAGGGGCATGAAAATGAAGAGTC
Macaque_Rh50 GGCTGTAGCAGGCATCTTGTATCGATCTGCACTGAGAAGGGGGCATAAAAATGAAGAGTC
Mause_Rh50 AGCAGTGGCAGGTGTGTTATATCGGCCTGGACTCAGATGTGAACACCCAAATGATGAATC
Rat_Rh50 AGCGGTAGCAGGCGTCTTATACCGGTCTGGCCTCAAACATGGACACCCAAATGAAGAATC
Xenopus_Rh50 GGCGGCTGCAGTGGTTTTATACCGTCCTTCGCTGAAAAATGGCCATGAAAACGAAGGGTC
Medaka_Rh50 GGCCATATCCTGGGTGCTTTACCGACCAAATCTACATAGAAGTAAACGACTCAATGGATC
Human_Rh50 CGCATACTACTCAGACTTGTTTGCAATGATTGGGACTCTCTTTCTGTGGATGTTTTGGCC
Macaque_Rh50 CACTTACTACTCAGACTTGTTTGCAATGATTGGGACTCTCTTTCTATGGATGTTTTGGCC
Mause_Rh50 TGTGTACCACTCTGACTTGTTTGCCATGATCGGAACACTTTTCCTGTGGATATTTTGGCC
Rat_Rh50 TGTGTACCACTCTGACTTGTTTGCAATGATCGGAACACTTTTCCTGTGGATGTTTTGGCC
Xenopus_Rh50 TGTTTATCACTCGGATTTGTTTGCTATGATCGGAACCCTTTTCCTCTGGATGTTTTGGCC
Medaka_Rh50 CGTTTACCACTCTGATCTTTTTGCAATGATTGGCACATTGTTCCTGTGGATGTTCTGGCC
M50-11
O50-5
M50-8


Human_Rh50
Macaque_Rh50
Mause_Rh50
Rat_Rh50
Xenopus_Rh50
Medaka_Rh50
O50-2
M50-4
R50-1
DEP-10
DEP-8
M50-2
O50-7
O50-8
M50-1
X50-8


Human_Rh50
Macaque_Rh50
Mause_Rh50
Rat_Rh50
Xenopus_Rh50
Medaka_Rh50
X50-7
M50-3
M50-6
X50-10
O50-11
R50-3
X50-9
O50-10
M50-5

```
Human_Rh50      CTTCTCCAGGATTCCCAAAAAGCTTTTGGCAGTG
Macaque_Rh50    NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Mause_Rh50      NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Rat_Rh50        NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Xenopus_Rh50    AATAAGTTGCTGCAAATAAAAATGCCTATGACTT
Medaka_Rh50     NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

(C) Degenerate primers

```
DEP-1   GTICTGATITCIIIKGRWGCTGT
DEP-2   GCCARAACATCCACARRAAGAG
DEP-3   GCRCYCTCTTYTTGTGGATGTT
DEP-4   GGTRCCCACAGCMACKCCTCC
DEP-5   TIGGIYTIGGCTTCCT
DEP-6   GCYACIGCYARICCAAAGTA
DEP-7   ATCCATRCMTTYGGRGCYTACT
DEP-8   TGYKCCCACIGCAACWCCTCCT
DEP-9   ATGATATTTGTIGGIYTIGGCTTC
DEP-10  CCACIGCAACWCCTCCTGCIA
DEP-11  GCTGGAGCAGYGTNGCNTTCA
DEP-12  GGTCNCTNCGNTGYTGYYTNCC
DEP-13  ICCCCAGATGGAGATCARCCC
DEP-14  GTRCARGGRACAATCYTGCTNGA
```