J Mol Evol (1996) 42:459–468

JOURNAL OF MOLECULAR EVOLUTION

Model of Amino Acid Substitution in Proteins Encoded by Mitochondrial DNA

Jun Adachi,[1] **Masami Hasegawa**[2]

[1] Department of Statistical Science, The Graduate University for Advanced Studies, 4-6-7 Minami-Azabu, Minato-Ku, Tokyo 106, Japan
[2] The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan

Received: 3 May 1995 / Accepted: 31 October 1995

Abstract. Mitochondrial DNA (mtDNA) sequences are widely used for inferring the phylogenetic relationships among species. Clearly, the assumed model of nucleotide or amino acid substitution used should be as realistic as possible. Dependence among neighboring nucleotides in a codon complicates modeling of nucleotide substitutions in protein-encoding genes. It seems preferable to model amino acid substitution rather than nucleotide substitution. Therefore, we present a transition probability matrix of the general reversible Markov model of amino acid substitution for mtDNA-encoded proteins. The matrix is estimated by the maximum likelihood (ML) method from the complete sequence data of mtDNA from 20 vertebrate species. This matrix represents the substitution pattern of the mtDNA-encoded proteins and shows some differences from the matrix estimated from the nuclear-encoded proteins. The use of this matrix would be recommended in inferring trees from mtDNA-encoded protein sequences by the ML method.

Key words: General reversible Markov model — Amino acid substitution — Maximum likelihood method

Introduction

Any method for inferring molecular phylogeny assumes explicitly or implicitly a model for the fundamental process of evolution, that is, nucleotide or amino acid substitution. Clearly, the assumed model should be as realistic as possible. Dependence among neighboring nucleotides in a codon complicates the problem in modeling the nucleotide substitution in protein-encoding genes, and it seems preferable to model the amino acid substitution.

Since selective constraints are more likely to be operating at the codon level rather than at the individual nucleotide level, it would be more realistic to construct a model for amino acid (rather than for nucleotide) substitutions to perform phylogenetic analyses of protein-encoding genes. The transition matrices of amino acid substitutions have previously been estimated by the parsimony method for data sets which consist mainly of nuclear-encoded proteins (Dayhoff et al. 1978; Jones et al. 1992). However, the parsimony method sometimes gives a biased estimate of the transition matrix (Collins et al. 1994; Perna and Kocher 1995).

Collins et al. (1994) pointed out that, in the presence of compositional bias, the transition matrix estimated by parsimony might be systematically distorted. From the method, common-to-rare state changes tend to predominate over rare-to-common changes, and therefore in the common ancestral node the estimated compositional bias tends to be more extreme than those of the contemporary species. By using the cytochrome *b* gene sequences from the gastropods (their original data) and from the pecoran ruminants (Irwin et al. 1991), they demonstrated this trend for both of the data sets. It is clear that this is due to the bias of the parsimony in inferring the ancestral

Correspondence to: M. Hasegawa

state when compositional bias exists. Perna and Kocher (1995) also demonstrated the same characteristic of the parsimony. Furthermore, the parsimony method has no time structure (Goldman 1990), and it is not effective when the proteins being used as not closely enough related to detect all the replacement events with parsimony procedures. The maximum likelihood (ML) method can overcome these defects of the parsimony, and therefore it is desirable to estimate the matrix by using the ML (Yang 1994).

Recently, Naylor et al. (1995) have pointed out that, since the bias for T and C at second codon positions is directly correlated with hydrophobicity of an encoded amino acid and since mtDNA-encoded proteins contain a high proportion of hydrophobic amino acids, the second codon positions of mtDNA, hitherto regarded as perhaps the most reliable for inferring evolutionary histories of distantly related species, may actually carry less phylogenetic information than the more fast-evolving first positions whose compositional bias is less skewed. Thus, it seems difficult to take fully into account different constraints operating on different codon positions when the analysis is carried out at the nucleotide sequence level.

Mitochondrial DNA sequences encoding proteins have been widely used for inferring the phylogenetic relationships among species (e.g., Irwin et al. 1991; Horai et al. 1992; Adachi et al. 1993; Janke et al. 1994; Cao et al. 1994). However, since the mitochondrial code is different from the universal code and since most of the mtDNA-encoded proteins are membranous, the transition matrix of the mtDNA-encoded proteins might be different from that estimated from nuclear-encoded proteins. Thus, it seems desirable to model the amino acid substitution of mtDNA-encoded proteins, and therefore we estimated the 20 × 20 transition probability matrix of the general reversible Markov model for mtDNA-encoded proteins by the ML method. This model is an extension to amino acids of the general reversible Markov model of nucleotide substitution proposed by Yang (1994).

Markov Models of Amino Acid Substitution

Transition Probability Matrix

We assume that each site evolves independently of the other sites according to a reversible Markov process. A probability of an amino acid *i* being replaced by an amino acid *j* in an infinitesimally short time interval, *dt*, is represented by $P_{ij}(dt)$. We would like to derive a transition probability matrix for a finite time *t*, $\mathbf{P}(t)$, where $\sum_{j=1}^{20} P_{ij}(t) = 1$ $(i = 1, \ldots, 20)$. A time interval during which one amino acid substitution occurs per 100 sites is taken as a unit of time, and we consider a transition probability matrix **M** for a unit time interval; $\mathbf{P}(1) = \mathbf{M}$. Adoption of a shorter time interval as a unit does not make any significant difference of the transition matrix estimated below (data not shown). Kishino et al. (1990) presented a method for deriving a transition probability matrix $\mathbf{P}(t)$ from **M.** We will follow their procedure.

If the unit time interval is sufficiently short, the transition probability matrix $\mathbf{P}(t)$ for time interval *t* is well approximated by

$$\mathbf{P}(t) = \exp(t\mathbf{W}) \tag{1}$$

where **W** is a function of eigen-values λ_i and eigenvectors $\mathbf{u}_i$ of **M,** and is represented by

$$\boldsymbol{W} = \boldsymbol{U} \begin{pmatrix} \lambda_1 & & 0 \\ & \cdots & \\ 0 & & \lambda_{20} \end{pmatrix} \boldsymbol{U}^{-1} \tag{2}$$

and

$$\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{20}) \tag{3}$$

Therefore,

$$P_{ij}(t) = \sum_{k=1}^{20} U_{ik} U_{kj}^{-1} \exp(t\lambda_k) \tag{4}$$

Thus, if the transition probability matrix for a unit time is given, the matrix for time *t* can be calculated.

Poisson Model

The simplest model for amino acid substitution is the Poisson model, in which an amino acid is replaced by any other amino acids with an equal probability. Let δ be the number of amino acid substitutions per site per unit time interval, and we take $\delta = 0.01$. The transition probability for a unit time of the Poisson model is,

$$M_{ij} = \begin{cases} \delta/19 & (i \neq j) \\ 1 - \delta & (i = j) \end{cases} \tag{5}$$

Although the representation of **M** is simple for the Poisson model, it becomes complicated for models in which the transition rate differs among different pairs of amino acids. In order to derive **M** in these models, we define the relative substitution rate **R** as follows:

$$R_{ii} = 0 \qquad (i = 1, \ldots, 20)$$

$$R_{ij} = R_{ji} \geqslant 0 \quad (i,j = 1, \ldots, 20)$$

R is related to the accepted mutation matrix **A** in Fig. 80 of Dayhoff et al. (1978) by the following formula:

$$R_{ij} = A_{ij}/(\pi_i^A \pi_j^A) \tag{6}$$

where π_i^A is the frequency of amino acid i in the data set used in constructing **A** (given in Table 22 of Dayhoff et al.). The matrix **R** represents relative rate of substitutions, and its absolute value has no special meaning. Differing from the transition probability matrix, a summation of a row need not be 1. Because of this freedom from the constraint, we can give the matrix easily.

The relative substitution rate for the Poisson model is

$$R_{ij} = \begin{cases} \alpha \ (i \neq j) \\ 0 \ (i = j) \end{cases} \tag{7}$$

Usually we take $\alpha = 1$.

From ***R,*** we can derive ***M*** as follows:

$$M_{ij} = \begin{cases} 20\delta R_{ij}/s & (i \neq j) \\ 1 - 20\delta \sum_{k=1}^{20} R_{ik}/s & (i = j) \end{cases} \tag{8}$$

where

$$s = \sum_{i=1}^{20} \sum_{j=1}^{20} R_{ij} \tag{9}$$

Proportional Model

In the proportional model, which is an extension to amino acids of the model for nucleotides proposed by Felsenstein (1981), P_{ij} is proportional to the frequency of amino acid j, π_j (where $\sum_{j=1}^{20} \pi_j = 1$), and the relative substitution rate is identical with that of the Poisson model (Eq. 7). If the amino acid frequency of the data under analysis is taken as π, this means that the frequency of the data is at the stationary state of the Markov process. A higher abundance of an amino acid than others is interpreted to be due to higher substitution probability to the amino acid than to the others. The transition probability matrix **M** for the proportional model is given by

$$M_{ij} = \begin{cases} \delta \pi_j R_{ij}/s & (i \neq j) \\ 1 - \delta \sum_{k=1}^{20} \pi_k R_{ik}/s & (i = j) \end{cases} \tag{10}$$

where

$$s = \sum_{i=1}^{20} \sum_{j=1}^{20} \pi_i \pi_j R_{ij} \tag{11}$$

By using this transformation, we can easily construct a model dependent on π.

General Reversible Markov Model

By increasing the number of parameters in **R,** we can construct various Markov models for amino acid substitutions. Yang (1994) estimated 4 × 4 transition matrices of the most general reversible Markov model (REV model) of nucleotide substitution for primate ψη-globin pseudogenes and for primate mtDNA sequences.

The relative substitution rate of the REV model of amino acid substitution has 20 ×19/2 minus 1 degree of freedom, and is given by

$$R_{ij} = \begin{cases} r_{ij} \ (i \neq j) \\ 0 \quad (i = j) \end{cases} \tag{12}$$

where $r_{ij} = r_{ji}$.

By using the transformation of Eq. 10, we can obtain the transition probability matrix **M** of the REV model for a unit time interval. Provided the tree topology which generated the nucleotide sequence data is known, we can estimate the relative substitution rate **R** by the ML, and the procedure is given by Adachi (1995).

Sequence Data

The transition probability matrix of the REV model for mtDNA-encoded proteins (the mtREV model) was estimated through ML by using the complete mtDNA sequences from 20 vertebrate species (3 individuals from human) listed in Table 1. Only the 12 proteins encoded in the same strand of mtDNA were used and NADH dehydrogenase subunit 6 (ND6) was omitted, because it is coded on the complementary strand and thus has different nucleotide and accordingly different amino acid compositions (Hasegawa and Kishino 1989). Positions with gaps and regions where the alignment was ambiguous were excluded. Overlapping regions between ATPase subunits 6 and 8 and between ND4 and ND4L were also excluded. The following protein-encoding regions were used in this work: ND1 (3322–4050, 4054–4251 in the numbering of Anderson et al., 1981), ND2 (4473–5180, 5184–5423, 5430–5447, 5451–5456, 5460–5471, 5475–5483), COI (5907–6350, 6354–7421), COII (7589–7735, 7739–8245), ATPase8 (8369–8446, 8474–8497, 8501–8503, 8507–8524), ATPase6 (8575–8607, 8644–8703, 8707–8880, 8884–8985, 8989–9030, 9040–9081, 9088–9204). COIII (9210–9272, 9276–9914, 9918–9920, 9924–9989). ND3 (10,092–10,109, 10,116–10,154, 10164–10,400), ND4L (10,476–10,496, 10,503–10,646, 10,659–10,757), ND4 (10,769–11,035, 11,039–11,677, 11,690–12,007, 12,011–12,127), ND5 (12,355–12,372, 12,469–12,933, 12,973–13,299, 13,303–13,680, 13,684–13,827, 13,900–13,992, 13,996–14,028, 14,074–14,109), and Cyt-b (14,750–15,598, 15,602–15,880). The total number of deduced amino acid sites was 3,357.

Table 1. List of data used in estimating the mtREV matrix

Species name		Reference	Database
Bos taurus	Cow	Anderson et al. 1982	V00654
Balaenoptera physalus	Fin whale	Árnason et al. 1991	X61145
Balaenoptera musculus	Blue whale	Árnason and Gullberg 1993	X72204
Phoca vitulina	Harbor seal	Árnason and Johnsson 1992	X63726
Halichoerus grypus	Grey seal	Árnason et al. 1993	X72004
Equus caballus	Horse	Xu and Árnason 1994	X79547
Mus musculus	Mouse	Bibb et al. 1981	V00711
Rattus norvegicus	Rat	Gadaleta et al. 1989	X14848
Homo sapiens	European	Anderson et al. 1981	J01415[a]
Homo sapiens	Japanese (DCM1)	Ozawa et al. 1991	
Homo sapiens	African (SB17F)	Horai et al. 1995	D38112
Pan troglodytes	Chimpanzee	Horai et al. 1995	D38113
Pan paniscus	Bonobo	Horai et al. 1995	D38116
Gorilla gorilla	Gorilla	Horai et al. 1995	D38114
Pongo pygmaeus	Orangutan	Horai et al. 1995	D38115
Didelphis virginiana	Opossum	Janke et al. 1994	Z29573
Gallus gallus	Chicken	Desjardins and Morais 1990	X52392
Xenopus laevis	Clawed frog	Roe et al. 1985	X02890
Cyprinus carpio	Carp	Chang et al. 1994	X61010
Crossostoma lacustre	Loach	Tzeng et al. 1992	M91245
Oncorhynchus mykiss	Trout	Zardaya et al. 1995	L29771
Petromyzon marinus	Sea laprey	Lee and Kocher 1995	U11880

[a] Revised according to Horai et al. (1995)

Transition Probability Matrix of the mtREV Model

Figure 1 shows the unrooted tree (Cao et al. 1994; Horai et al. 1995), among species from which complete mtDNA sequences are available, assumed in the estimation of the transition probability matrix. The placing of lamprey in this figure is not the ML tree but the second highest likelihood tree, and ((Birds, Mammals), (Xenopus, Fishes), Lamprey) is the ML tree. However, since the difference of log-likelihood of this tree from that of the ML tree is minor (9.6 ± 15.6 where ± 1 SE estimated by the formula in Kishino and Hasegawa, 1989), we used this biological tree. Since the branching order among Carnivora, Perissodactyla, and the Cetacea/Artiodactyla clade cannot be resolved by the mtDNA data, it was left as a trifurcation.

Starting from initial values of **R** and of branch lengths **t,** we continued to iterate ML estimation of **R** by the Brent method and of *t* by the Newton-Raphson method alternately until convergence was attained. An SE of **R** was estimated from the second derivative of the likelihood function by using the same procedure in the quasi-Newton method. Table 2 gives the estimated relative substitution rate matrix **R** for the mtREV model with its SE. We carried out the estimation starting from three different initial matrices for Poisson, Dayhoff, and JTT and obtained the same estimate as shown in this table. Therefore, we think that we found the global maximum, not a local one, of likelihood. Table 3 gives the estimated transition probability matrix **M** of the mtREV model for a unit time interval. The estimated transition matrix is not sensitive to the assumed tree (Adachi 1995; Yang 1994).

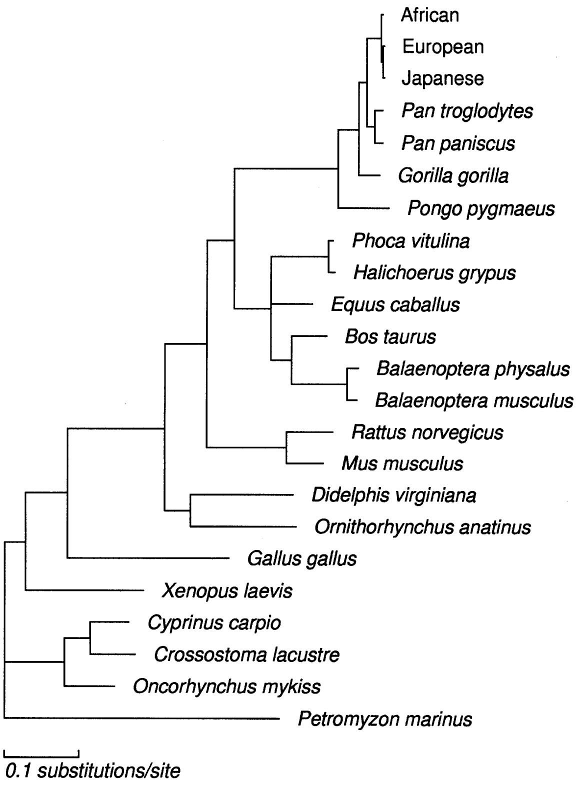

Fig. 1. The unrooted tree (Cao et al. 1994; Horai et al. 1995) used in estimating the transition probability matrix of Table 3. The horizontal length of each branch is proportional to the number of amino acid substitutions estimated by the ML method based on the mtREV model.

Table 2. Relative substitution rate matrix of mtREV model[a]

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala		41	70	48	192	42	41	86	52	79	30	—	115	28	61	146	150	—	35	131
Arg	57		69	—	426	316	—	71	256	—	33	268	—	34	81	42	27	96	—	48
Asn	142	58		496	264	223	167	78	322	54	34	354	86	52	106	233	140	51	184	55
Asp	60	1	4,338		—	191	567	117	258	40	14	138	—	40	51	129	81	70	81	—
Cys	297	636	212	1		248	—	153	440	199	115	—	—	219	115	410	315	210	539	—
Gln	18	1,245	767	334	212		352	40	389	55	53	451	131	81	161	117	128	—	130	81
Glu	59	1	434	3,394	1	1,833		60	148	—	—	325	—	—	51	101	61	—	89	74
Gly	696	142	273	348	192	42	115		—	25	9	53	25	—	—	88	29	33	—	20
His	87	826	2,468	738	883	3,223	265	1		51	29	199	—	93	89	107	90	48	429	—
Ile	590	1	152	39	299	70	1	41	93		83	48	189	69	34	57	125	—	60	305
Leu	129	83	110	4	219	209	1	6	61	1,696		30	141	76	35	48	56	40	49	70
Lys	1	771	2,684	66	1	2,566	1,545	83	391	54	45		143	51	112	152	143	112	152	64
Met	729	1	195	1	1	293	1	7	1	2,490	2,815	397		102	52	122	196	87	94	244
Phe	48	30	40	37	401	166	1	1	214	404	1,196	51	429		38	68	51	39	239	36
Pro	294	159	435	50	90	761	41	1	233	68	227	250	85	94		104	86	32	50	36
Ser	2,074	41	2,791	363	1,706	370	334	689	344	200	422	515	627	360	852		172	66	86	—
Thr	2,655	17	1,123	165	978	545	81	49	307	1,991	628	714	2,846	157	674	3,155		47	68	138
Trp	1	135	57	51	208	1	1	46	42	1	176	164	134	51	28	178	86		80	35
Tyr	35	1	818	58	1,390	201	110	1	3,672	161	228	274	164	2,298	82	292	153	135		55
Val	947	48	63	1	1	67	129	16	1	6,465	467	9	2,169	33	48	1	1,117	34	30	

[a] Relative substitution rate matrix (lower left) and its SE (upper right). Sum of lower left matrix elements is fixed to be 10^5. Elements smaller than 1 are fixed to be 1, and their SEs are not estimated

Table 3. Transition probability matrix P_{ij} ($\times 10^5$) of the amino acid i being replaced by the amino acid j during a time interval of one substitution per 100 amino acids (1PAM) for the mtREV model, and average amino acid frequencies π of the mtDNA-encoded proteins

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	99,007	2	9	2	3	1	2	64	4	84	35	0	63	5	26	243	380	0	2	67
Arg	7	99,794	4	0	7	51	0	13	38	0	23	29	0	3	14	5	2	6	0	3
Asn	17	2	98,905	134	2	31	17	25	114	22	30	102	17	4	39	327	161	3	44	4
Asp	7	0	279	99,410	0	14	135	32	34	6	1	3	0	4	4	42	24	2	3	0
Cys	35	20	14	0	99,290	9	0	18	41	42	60	0	0	40	8	200	140	10	75	0
Gln	2	38	49	10	2	99,261	73	4	148	10	57	98	25	16	68	43	78	0	11	5
Glu	7	0	28	105	0	75	99,634	11	12	0	0	59	0	0	4	39	12	0	6	9
Gly	82	4	18	11	2	2	5	99,774	0	6	2	3	1	0	0	81	7	2	0	1
His	10	25	159	23	9	132	10	10	99,260	13	17	15	0	21	21	40	44	2	198	0
Ile	70	0	10	1	3	3	0	4	4	98,398	465	2	216	40	6	23	285	0	9	461
Leu	15	3	7	0	2	9	0	1	3	241	99,142	2	244	118	20	49	90	8	12	33
Lys	0	24	173	2	0	105	61	8	18	8	12	99,342	34	5	22	60	102	8	15	1
Met	86	0	13	0	0	12	0	1	0	354	772	15	98,047	42	8	73	408	6	9	155
Phe	6	1	3	1	4	7	0	0	10	57	328	2	37	99,343	8	42	23	2	124	2
Pro	35	5	28	2	1	31	2	0	11	10	62	10	7	9	99,583	100	96	1	4	3
Ser	246	1	179	11	18	15	13	64	16	28	116	20	54	36	76	98,631	452	8	16	0
Thr	315	1	72	5	10	22	3	5	14	283	172	27	247	16	60	369	98,287	4	8	80
Trp	0	4	4	2	2	0	0	4	2	0	48	6	12	5	3	21	12	99,866	7	2
Tyr	4	0	53	2	14	8	4	0	169	23	62	10	14	227	7	34	22	6	99,336	2
Val	112	1	4	0	0	3	5	2	0	918	128	0	188	3	4	0	160	2	2	98,467
π	0.072	0.019	0.039	0.019	0.006	0.025	0.024	0.056	0.028	0.087	0.168	0.023	0.053	0.060	0.055	0.072	0.088	0.029	0.033	0.044

Comparison Between the mtREV and JTT-F Models

The mtREV model can be compared with Jones, Taylor, and Thornton's (1992) model of nuclear-encoded proteins adjusted with the amino acid frequencies of the mtDNA-encoded proteins as the equilibrium frequencies (JTT-F model; Cao et al. 1994; Adachi and Hasegawa 1996). The log-likelihood of the tree in Fig. 1 for the mtREV model is −46,240, while that for the JTT-F model is −47,039, showing much improved fitting of the mtREV model to the mtDNA-encoded protein data.

Table 4 shows the difference of the transition probability matrix of mtREV model from that of the JTT-F model. One of the most remarkable characteristics of the transition matrix for the mtREV model is that the transitions between Arg and Lys are very rare compared to those observed in nuclear-encoded proteins. The transition probability of Arg ↔ Lys for 1PAM in the mtREV model is only one-fifth of that in the JTT-F model. The SE of the **R** shown in Table 3 suggests that this difference is significant. This might be due to the difference between universal and mitochondrial genetic codes. In the universal code, Lys can be substituted by Arg with a

Table 4. Difference of the transition probability matrix of the mtREV model for 1PAM from that of the JTT-F model (×10^5)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	181	−7	−12	−13	−1	−12	−23	−30	−2	54	−14	−8	40	−3	−75	−18	2	−2	−2	−55
Arg	−30	555	−13	−3	1	−21	−7	−61	−48	−20	−37	−117	−21	−1	−23	−63	−52	−30	−7	−4
Asn	−21	−6	−66	37	0	13	4	−17	7	−19	8	46	1	0	33	−10	−31	2	22	−3
Asp	−49	−3	76	102	−1	2	−41	−36	7	−4	−9	−3	−9	2	−2	3	−11	0	−11	−13
Cys	−5	1	2	−2	−164	7	−1	−12	21	29	33	−1	−20	0	2	53	101	−22	7	−26
Gln	−36	−16	20	1	1	31	−6	−9	−4	3	−56	32	2	13	−15	6	35	−5	3	−3
Glu	−66	−6	7	−32	0	−6	177	−49	5	−9	−16	20	−9	−3	−6	18	−15	−3	4	−10
Gly	−39	−21	−11	−12	−1	−4	−21	245	−5	1	−9	−3	−5	−3	−11	−46	−19	−14	−2	−19
His	−5	−32	10	5	5	−3	4	−11	154	−2	−69	3	−16	−5	−38	−10	5	−2	13	−5
Ile	45	−4	−8	−1	2	1	−3	1	−1	−208	88	−3	−25	−6	1	−5	72	−4	−1	59
Leu	−6	−4	2	−1	1	−8	−2	−3	−11	46	−37	−1	54	−28	−35	9	67	−7	4	−41
Lys	−26	−94	79	−3	0	34	21	−7	4	−9	−12	−67	3	4	11	28	21	5	12	−4
Met	54	−8	1	−3	−2	1	−4	−6	−9	−40	171	1	−500	17	0	54	237	0	3	31
Phe	−4	0	0	0	0	6	−1	−3	−2	−9	−77	1	15	172	0	−22	11	−14	−52	−23
Pro	−99	−8	24	0	0	−7	−2	−11	−19	2	−106	5	−1	0	315	−86	0	−1	0	−6
Ser	−19	−17	−6	1	5	2	6	−37	−4	−5	22	10	40	−18	−66	46	61	−1	−4	−17
Thr	2	−11	−14	−2	7	10	−5	−12	1	71	129	6	144	8	0	50	−419	2	2	33
Trp	−6	−19	3	1	−5	−4	−3	−26	−2	−12	−39	4	2	−27	−1	0	5	154	−17	−8
Tyr	−6	−4	27	−6	1	2	3	−3	11	−3	23	8	5	−96	1	−9	5	−16	59	−5
Val	−90	−2	−2	−6	−4	−1	−6	−24	−3	117	−155	−3	37	−32	−7	−28	65	−5	−3	151

one-step change, while in the vertebrate mitochondrial code it requires a two-step change. Therefore, although Arg and Lys are chemically similar (both are basic amino acids) and hence are frequently substituted with each other in nuclear-encoded proteins, Arg ↔ Lys substitutions are much less frequent in vertebrate mitochondria. This probably explains why Arg is the second-most-conservative amino acid in the mtREV model, while it is only the ninth-most-conservative in the JTT-F model. These observations demonstrate the importance of the mutation-driven neutral evolution (Kimura 1983) under the constraint of the genetic code.

The substitutions between chemically similar amino acids with a one-step nucleotide change, such as Val ↔ Ile, Ala ↔ Thr, Met ↔ Leu, Ile ↔ Leu, Met ↔ Ile, Ser ↔ Thr, and Phe ↔ Leu, are very frequent both in the mtREV and the JTT-F models. In agreement with the neutral theory (Kimura 1983), this suggests that most of the amino acid substitutions in evolution are conservative rather than progressive (McLachlan 1971; Grantham 1974). Met ↔ Thr substitutions are more frequent in the mtREV model than in the JTT-F model by 2.4-fold. Again, this might be due to peculiarities of the mitochondrial code, in which there are two codons for Met, while there is only one in the universal code.

The transition probability of Pro (codons: CCX) ↔ Ala (GCX), in which transversion in a codon is needed, for the mtREV model is only 0.26 of that for the JTT-F model. Increased nucleotide transition rate of mtDNA relative to transversion rate (Brown et al. 1982) might be responsible to this difference. Lower rates of Val (GUX) ↔ Leu (CUX, UUR) and Tyr (UAY) ↔ Phe (UUY) and higher rates of Val (GUX) ↔ Ile (AUY) and Thr (AUX) ↔ Ile (AUY) (in spite of the decreased number of codons for Ile in mitochondria) in the mtREV model than in the JJT-F model might also be due to the difference of transition/transversion mutation ratio between mtDNA and nuclear DNA. However, not all the differences between the mtREV and JTT-F model conform to this expectation. For example, transition probabilities of Pro (CCX) ↔ Leu (CUX, UUR), Pro (CCX) ↔ Ser (UCX, AGY), Val (GUX) ↔ Ala (GCX), and Phe (UUY) ↔ Leu (CUX, UUR), which are achieved by a transition, for the mtREV model are 0.37, 0.54, 0.55, and 0.81 of those for the JTT-F model, respectively, and the probability of Lys (AAR) ↔ Asn (AAY), which requires a transversion, is 1.84 times higher. These differences are not interpretable.

Cys is the fourth-most-conservative amino acid in the JTT-F model, while it is only the tenth in the mtREV model. This might be due to the fact that, since most of the mtDNA-encoded proteins are membranous, cysteines in the mtDNA-encoded proteins are not involved in disulfide bonds so often as in the nuclear-encoded proteins in which globular proteins occupy a larger portion. All the differences of the transition matrix between the mtREV and the JTT-F models are not necessarily interpretable in straightforward ways. Some of the differences might be due to the biased estimate of the JTT-F matrix by the parsimony method (Collins et al. 1994; Perna and Kocher 1995; Goldman 1990; Yang 1994), and some of the others might be due to the small sample size of the data in estimating the mtREV matrix.

Table 5. Comparison of amino acid frequencies between mitochondrial and nuclear-encoded proteins[a]

	Mt code	Mitochondria	Nuclear	Mt/nuc
Trp	UGR	0.029	0.014	2.07
Tyr	UAY	0.033	0.032	1.03
Phe	UUY	0.060	0.040	1.50
Leu	UUR, CUX	0.168	0.091	1.85
Ile	AUY	0.087	0.053	1.64
Met	AUR	0.053	0.024	2.21
Val	GUX	0.044	0.066	0.67
Ala	GCX	0.072	0.077	0.94
Pro	CCX	0.055	0.051	1.08
Gly	GGX	0.056	0.074	0.76
Thr	ACX	0.088	0.059	1.49
Ser	UCX, AGY	0.072	0.069	1.04
Asn	AAY	0.039	0.043	0.91
Asp	GAY	0.019	0.052	0.37
Gln	CAR	0.025	0.041	0.61
Glu	GAR	0.024	0.062	0.39
His	CAY	0.028	0.023	1.22
Lys	AAR	0.023	0.059	0.39
Arg	CGX	0.019	0.051	0.37
Cys	UGY	0.006	0.020	0.30

[a] Average amino acid frequencies of the mtDNA-encoded proteins (the mtREV model) and of the nuclear-encoded proteins (the JTT model)

Table 5 gives amino acid frequencies of the mtDNA-encoded proteins used in the estimation of the mtREV matrix (12 proteins) and of the proteins used in the estimation of the JTT matrix which consist mainly of nuclear-encoded ones. Cys is scarce in the mtDNA-encoded proteins probably because this amino acid is not involved in disulfide bonds so often as in the nuclear-encoded proteins, as mentioned before. The mtDNA-encoded proteins are mostly membranous, and probably for this reason, hydrophobic amino acids, such as Met, Trp, Leu, Ile, and Phe, are more abundant, and hydrophilic amino acids, such as Arg, Lys, Glu, Asp, and Gln, are more scarce than in the nuclear-encoded proteins. Of course, that Met and Trp are more abundant in the mtDNA-encoded proteins than in the nuclear-encoded proteins might also be due to their having two codons in mitochondria and only one in the universal code. However, in disagreement with the above expectation, the frequencies of hydrophobic amino acids, such as Val (codon: GUX) and Gly (GGX), are less in the mtDNA-encoded proteins than in the nuclear-encoded proteins. This might be due to the fact that the codons of these amino acids contain G, which is scarce in the L-strand of mtDNA (the 12 proteins used in this analysis are encoded by the H-strand, and the mRNAs are complementary to the H-strand). In agreement with this consideration, Val and Gly are three times more abundant in ND6, which is encoded by the L-strand (G is abundant in its mRNA) than in the 12 mtDNA-encoded proteins. This suggests that amino acid frequencies of the mtDNA-encoded proteins are governed not only by the structural-functional requirements of the individual proteins but also by the bias and skewness of mtDNA caused by its asymmetric replication pattern (Tanaka and Ozawa 1994; W.K. Thomas, personal communication).

Discussion

Previously, the JTT model for nuclear-encoded proteins was used even in the ML analyses of mtDNA-encoded proteins (Adachi et al. 1993; Cao et al. 1994; Adachi and Hasegawa 1995), because no appropriate model for mtDNA-encoded proteins was available. The conclusions of these phylogenetic analyses hold when the mtREV model presented in this paper is used. This suggests that the ML method is robust to some extent against the violation of the assumed model (Hasegawa and Fujiwara (1993). Nevertheless, phylogenetic conclusions derived from a realistic model should be more reliable than that from a less realistic one, and therefore we must continue to improve the model. Once a probabilistic model as shown in Table 3, which is realistic to some extent, is obtained, the ML method would be the preferred method in inferring trees from mtDNA-encoded protein sequences (Felsenstein 1981; Kishino et al. 1990; Edwards 1995). Although the amino acid frequencies of the individual protein under analysis might be different from the average frequencies of the 12 proteins used in estimating the transition matrix, the ProtML program of our package MOLPHY (Adachi and Hasegawa 1996) can adjust the equilibrium frequencies of the model to the actual frequencies of the protein under study (F-option).

The mtREV-F model gives much higher likelihood and better approximates the evolution of the individual proteins encoded by mtDNA than the JTT-F and Dayhoff-F models as far as we have examined for cytochrome *b* and cytochrome oxidase subunit II from vertebrates (data not shown). It remains to be discovered whether the model presented in this paper is also applicable to proteins in invertebrate, fungus, and plant mitochondria whose codes differ from that of vertebrate mitochondria.

If we are to analyze closely related sequences, synonymous substitutions provide us with important information, and therefore a codon-based model of nucleotide substitution (Schöniger et al. 1990; Muse and Gaut 1994; Goldman and Yang 1994) might be preferable to the amino acid substitution model. However, in constructing the model of nucleotide substitution, it must be noted that the nucleotide frequencies of the third codon positions are significantly different even between closely related species in Hominoidea (T is significantly more scarce and C is more abundant in orangutan than in gorilla; Adachi 1995), and that the reversible Markov model no longer holds for these sites. One of the advantages of the ML method over the other existing methods in molecular phylogenetics is that, as is demonstrated in this work, we can incorporate complexity in the pattern of substitution and can improve the model as the relevant data accumulate, because the method is based on an explicit model (Thorne et al. 1992). The parsimony method is used widely (Stewart 1993), but it is not based on the explicit model, and therefore it suffers limitations in taking account directly of the complex pattern of the actual process of evolution (Sidow 1994).

Acknowledgments. This paper constitutes part of the research done by J.A. in partial fulfillment of a Ph.D. degree at The Graduate University for Advanced Studies. We thank A. Sidow for discussions, M. Milinkovitch for comments on an earlier version of the manuscript, and Y. Cao for providing us with the aligned sequence data. We also thank three anonymous reviewers for constructive suggestions to improve the manuscript. This work was supported by grants from the Ministry of Education, Science, and Culture of Japan.

References

Adachi J (1995) Modeling of molecular evolution and maximum likelihood inference of molecular phylogeny. PhD dissertation, The Graduate University for Advanced Studies, Tokyo, Japan

Adachi J, Cao Y, Hasegawa M (1993) Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid sequence level: rapid evolution in warm-blooded vertebrates. J Mol Evol 36:270–281

Adachi J, Hasegawa M (1995) Phylogeny of whales: dependence of the inference on species sampling. Mol Biol Evol 12:177–179

Adachi J, Hasegawa M (1996) MOLPHY: programs for molecular phylogenetics, ver 2.3. Institute of Statistical Mathematics, Tokyo

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith ALH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–464

Anderson S, de Bruijn MHL, Coulson AR, Eperon IC, Sanger F, Young IG (1982) The complete sequence of bovine mitochondrial DNA: conserved features of the mammalian mitochondrial genome. J Mol Biol 156:683–717

Árnason Ú, Gullberg A (1993) Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. J Mol Evol 37:312–322

Árnason Ú, Gullberg A, Johnsson E, Ledje C (1993) The nucleotide sequence of the mitochondrial DNA molecule of the grey seal, *Halichoerus grypus,* and a comparison with mitochondrial sequences of other true seals. J Mol Evol 37:323–330

Árnason Ú, Gullberg A, Widegren B (1991) The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenoptera physalus.* J Mol Evol 33:556–568

Árnason Ú, Johnsson E (1992) The complete mitochondrial DNA sequence of the harbor seal, *Phoca vitulina.* J Mol Evol 34:493–505

Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA (1981)

Sequence and gene organization of mouse mitochondrial DNA. Cell 26:167–180

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primtes: tempo and mode of evolution. J Mol Evol 18:225–239

Cao Y, Adachi J, Janke A, Pääbo S, Hasegawa M (1994) Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J Mol Evol 39:519–527

Chang YS, Huang FL, Lo TB (1994) The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. J Mol Evol 38:138–155

Collins TM, Wimberger PH, Naylor GJP (1994) Compositional bias, character-state bias, and character-state reconstruction using parsimony. Syst Biol 43:482–496

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington DC, pp 345–352

Desjardins P, Morais R (1990) Sequence and gene organization of the chicken mitochondrial genome: a novel gene order in higher vertebrates. J Mol Biol 212:599–634

Edwards AWF (1995) Assessing molecular phylogenies. Science 267: 253–253

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Gadaleta G, Pepe G, De Candia G, Quagliariello C, Sbisa E, Saccone C (1989) The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: cryptic signals revealed by comparative analysis between vertebrates. J Mol Evol 28:497–516

Goldman N (1990) Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. Syst Zool 39:345–361

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736

Grantham R (1974) Amino acid differences formula to help explain protein evolution. Science 185:862–864

Hasegawa M, Fujiwara M (1993) Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. Mol Phyl Evol 2:1–5

Hasegawa M, Kishino H (1989) Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. Jpn J Genet 64:243–258

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) The recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 92:532–536

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S, Takahata N (1992) Man's place in Hominoidea revealed by mitochondrial DNA genealogy. J Mol Evol 35:32–43

Irwin DM, Kocher TD, Wilson AC (1991) Evolution of the cytochrome *b* gene of mammals. J Mol Evol 32:128–144

Janke A, Feldmaier-Fuchs G, Thomas WK, von Haeseler A, Pääbo S (1994) The marsupial mitochondrial genome and the evolution of placental mammals. Genetics 137:243–256

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comp Appl Biosci 8:275–282

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

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179

Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. J Mol Evol 31:151–160

Lee WJ, Kocher TD (1995) Complete sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome: early establishment of the vertebrate genome organization. Genetics 139:873–887

McLachlan AD (1971) Tests for comparing related amino-acid sequences. Cytochrome *c* and cytochrome c_{551}. J Mol Biol 61:409–424

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

Naylor GJP, Collins TM, Brown WM (1995) Hydrophobicity and phylogeny. Nature 373:565–566

Ozawa T, Tanaka M, Ino H, Ohno K, Sano T, Wada Y, Yoneda M, Tanno Y, Miyatake T, Tanaka T, Itoyama S, Ikebe S, Hattori N, Mizuno Y (1991) Distinct clustering of point mutations in mitochondrial DNA among patients with mitochondrial encephalomyopathies and Parkinson's disease. Biochem Biophys Res Commun 176:938

Perna NT, Kocher TD (1995) Unequal base frequencies and the estimation of substitution rates. Mol Biol Evol 12:359–361

Roe BA, Ma DP, Wilson RK, Wong JFH (1985) The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J Biol Chem 260:9759–9774

Schöniger M, Hofacker GL, Borstnik B (1990) Stochastic traits of molecular evolution—acceptance of point mutations in native actin genes. J Theor Biol 143:287–306

Sidow A (1994) Parsimony of statistics? Nature 367:26–26

Stewart CB (1993) The powers and pitfalls of parsimony. Nature 361: 603–607

Tanaka M, Ozawa T (1994) Strand asymmetry in human mitochondrial DNA mutations. Genomics 22:327–335

Thorne JL, Kishino H, Felsenstein J (1992) Inching toward reality: an improved likelihood model of sequence evolution. J Mol Evol 34: 3–16

Tzeng CS, Hui CF, Shen Huang PC (1992) The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome: conservation and variations among vertebrates. Nucleic Acids Res 20: 4853–4858

Xu X, Árnason Ú (1994) The complete mitochondrial DNA sequence of the horse, *Equus caballus:* extensive heteroplasmy of the control region. Gene 148:357–362

Yang Z (1994) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

Zardoya R, Garrido-Pertierra A, Bautista JM (1995) The complete nucleotide sequence of mitochondrial DNA genome of the rainbow trout, *Oncorhynchus mykiss.* J Mol Evol 41:942–951