J Mol Evol (1984) 20:86–93

Journal of Molecular Evolution

A New Method for Calculating Evolutionary Substitution Rates

Cecilia Lanave,[1] Giuliano Preparata,[2] Cecilia Saccone,[1] and Gabriella Serio[3]

[1] Centro CNR Mitocondri e Metabolismo Energetico, Istituto di Chimica Biologica, Università, 70126 Bari, Italy
[2] Dipartimento di Fisica, Università, 70126 Bari, Italy [3] Dipartimento di Matematica, Università, 70121 Bari, Italy

Summary. In this paper we present a new method for analysing molecular evolution in homologous genes based on a general stationary Markov process. The elaborate statistical analysis necessary to apply the method effectively has been performed using Monte Carlo techniques. We have applied our method to the silent third position of the codon of the five mitochondrial genes coding for identified proteins of four mammalian species (rat, mouse, cow and man). We found that the method applies satisfactorily to the three former species, while the last appears to be outside the scope of the present approach. The method allows one to calculate the evolutionarily effective silent substitution rate (v_s) for mitochondrial genes, which in the species mentioned above is 1.4×10^{-8} nucleotide substitutions per site per year. We have also determined the divergence time ratios between the couples mouse–cow/rat–mouse and rat–cow/rat–mouse. In both cases this value is approximately 1.4.

Key words: Silent substitution — Molecular evolution — Evolution of mitochondrial DNA — Stochastic Markov chain — Monte Carlo simulation

Introduction

Since the development of techniques for sequencing amino acids in proteins there has been a considerable flourishing of studies on molecular evolution. A number of interesting properties have thus been revealed, all clearly demonstrating that proteins retain in their primary structures the evolutionary history of the organisms to which they belong. This in turn has allowed the construction of phylogenetic trees based on molecular data (see Wilson et al. 1977). With the discovery of the genetic code, these results could be extended to the genes themselves, even though as far as the silent third codon positions are concerned nothing could be established. New and powerful tools for the study of molecular evolution have become available recently via the direct sequencing of genes, which has uncovered such properties as the high rate of nucleotide substitution at the silent sites (Grunstein et al. 1976; Salser et al. 1976; Kimura 1977; Jukes 1980; Miyata et al. 1980; Perler et al. 1980).

There have been several attempts to determine from the number of nucleotide changes the rates of the silent substitutions in various genes. However, reliable estimates of the evolutionarily effective substitution rates of the four nucleotides strongly depend both on a trustworthy determination of the divergence time and a correct calculation of the nucleotide substitution number. Several methods have been suggested for the correct estimation of the evolutionary distance between homologous genes, and even though such methods have different bases, there is a wide consensus that the substitution process is stochastic in nature (Jukes and Cantor 1969; Kimura 1981; Takahata and Kimura 1981; Gojobori et al. 1982). It thus appears worthwhile to approach the problem in the most general terms, by carrying out an analysis of the process without any a priori assumptions about either the evolutionary time lengths involved or the structure of the substitution rate matrix. This is what we do in this paper.

To carry out our programme we need a statistical analysis whose results are more stringent the longer the nucleotide sequences of the genes considered are.

Abbreviations used: B, cow; H, human; M, mouse; MY, million years; R, rat

Offprint requests to: C. Saccone, Istituto di Chimica Biologica, Via Amendola 165/A, Università, 70126 Bari, Italy

A sufficient body of such data has begun to be available to permit a meaningful application of our approach. In the case of mammalian mitochondrial (mt) genes we are in possession of the nucleotide sequences of the entire or almost entire genomes for various species, so that our method can fruitfully be applied (Sekiya et al. 1980; Anderson et al. 1981; Bibb et al. 1981; Grosskopf and Feldman 1981; Kobayashi et al. 1981; Saccone et al. 1981; Anderson et al. 1982; Cantatore et al. 1982; Koike et al. 1982; Pepe et al. 1983). In this paper we analyse the nucleotide substitution patterns of five mammalian mt genes coding for identified products, namely the three cytochrome oxidase subunits (CoI, CoII, and CoIII), the 25-kilodalton ATPase subunit and the cytochrome b (Cyt b) subunit of four mammalian mt genomes [mouse (M), rat (R), cow (B) and human (H)].[1] We shall only consider the silent third codon positions, since all mutations in them should be selectively neutral.

To determine the evolutionary distance between mt genes a stationary Markov model is proposed, the prerequisite for its applicability being that the frequencies of each base at the silent positions (q_i) must be, within statistical error, constant both in the species considered and in their common ancestor. We found that this condition is satisfied for the mt genes of mouse, rat and cow. For the mt genome of human cells, however, the values q_i we obtained turn out to be significantly different from those of the other species, thus calling for a different mathematical model consistent with the observed divergence.

The Stochastic Model

Let us consider two codon sequences of length L on homologous genes (of the mt genome) of two different species originating from a common ancestor. We will call the ancestor A and the two species under study B and C, and let us suppose that B and C are at a time distance T from A. We denote the four nucleotide bases A, C, G, and T as 1, 2, 3, and 4, respectively, and concentrate our attention on the silent positions of each homologous gene of A, B and C.

The problem we want to solve is to determine the rate of evolutionarily effective substitutions from the rate of divergence between different coding sequences at the silent codon position. To achieve this we construct a Markov chain in which the four bases form the finite set of possible states.

For each member of the sequences we introduce the matrix $P_{ik}(T)$, which represents the probability that the base k of A has undergone the mutation into the base i of B or C. The argument T is the evolutionary divergence time, defined above.

By definition we have:

$$\sum_{i=1}^{4} P_{ik}(T) = 1 \text{ for each } k = 1, \ldots, 4 \qquad (1)$$

If we denote by $q_i(T)_{B,C}$, i = 1, . . . , 4, the frequencies of the nucleotide i in each of the sequences of B and C, and by $q_i(0)$ the same quantities for the ancestor A, we obviously have

$$q_i(T)_B = q_i(T)_C = \sum_k P_{ik}(T) q_k(0) \qquad (2)$$

Furthermore, if the Markov process is assumed, as we do in the following, to be stationary, we must have

$$q_i(0) = q_i(T)_B = q_i(T)_C \qquad (3)$$

This condition plays a fundamental role in our analysis, and constitutes the necessary prerequisite for the applicability of a stationary Markov model. Thus, within the statistical fluctuations to be analysed in the following section, we will require that the data obey condition (3).

Let us now indicate by R_{ik} the time-independent rate matrix which characterizes the stationary Markov process through the following Kolmogorov differential equations:

$$\frac{d}{dT} P_{ik}(T) = \sum_{r=1}^{4} P_{ir}(T) R_{rk} \qquad (4)$$

with the initial conditions

$$P_{ik}(0) = \delta_{ik} \qquad (5)$$

From Eqs. (4) and (5) we get:

$$R_{rk} = \frac{d}{dT} P_{rk}(T)|_T = 0 \qquad (6)$$

In matrix notation the solution of (4), with boundary conditions (5), can be written as

$$P(T) = e^{TR} \qquad (7)$$

To carry out our analysis it is necessary to determine explicitly the individual transition probabilities $P_{ik}(T)$ obeying the differential equations (4). These functions depend on the eigenvalues and eigenvectors of the rate matrix R in the manner we shall now describe.

In the case under study R possesses four eigenvalues λ_α $\{\alpha = 0, 1, 2, 3\}$, one of which is equal to zero and the other three of which are real and nonpositive. Thus, we may write

[1] The lengths in base pairs of the genes for R, M, B and H are, respectively: CoI: 1545, 1546, 1545, 1541; CoII: 684, 684, 684, 684; CoIII: 784, 784, 784, 784; ATPase: 683, 680, 681, 679; Cyt b: 1143, 1144, 1130, 1141

$$R_{ik} = \sum_{\alpha=0}^{3} \lambda_\alpha u_i^{(\alpha)} v_k^{(\alpha)}, \tag{8}$$

where $u_i^{(\alpha)}$, $v_k^{(\alpha)}$ are the left and right eigenvectors, respectively, and satisfy the orthogonality relation:

$$\sum_i u_i^{(\alpha)} v_i^{(\beta)} = \delta_{\alpha\beta} \tag{9}$$

It is easy to see that for $\lambda = 0$ the eigenvectors can be determined explicitly and are given by

$$\begin{aligned} v_i^{(0)} &\equiv (1, 1, 1, 1), \\ u_i^{(0)} &\equiv (q_1, q_2, q_3, q_4) \end{aligned} \tag{10}$$

For $\alpha = r = 1, 2, 3$, we introduce the vectors $w_i(r)$, according to the definitions

$$\begin{aligned} v_i^{(r)} &= \frac{1}{(q_i)^{1/2}} w_i^{(r)}, \\ u_i^{(r)} &= (q_i)^{1/2} w_i^{(r)} \end{aligned} \tag{11}$$

It is easily verified that the orthogonality conditions (9) are satisfied if

$$\sum_{i=1}^{4} (q_i)^{1/2} w_i^{(r)} = 0, \tag{12}$$

$$\sum_{i=1}^{4} w_i^{(r)} w_i^{(s)} = \delta_{rs} \tag{13}$$

By making use of Eqs. (10)–(13), we can rewrite (8) as

$$R_{ik} = \sum_{r=1}^{3} \lambda_r \left(\frac{q_i}{q_k}\right)^{1/2} w_i^{(r)} w_k^{(r)} \tag{14}$$

and through Eq. (7) we can give the matrix $P_{ik}(T)$ the representation

$$P_{ik}(T) = u_i^{(0)} v_i^{(0)} + \sum_{r=1}^{3} e^{\lambda_r T} \left(\frac{q_i}{q_k}\right)^{1/2} w_i^{(r)} w_k^{(r)} \tag{15}$$

Let us now consider the two sequences B and C evolved from the common ancestor A after time T. The probability that at a given site one finds the nucleotide i in the sequence of B and the nucleotide j in the sequence of C is given by

$$S_{ij}(T) = \sum_{k=1}^{4} P_{ik}(T) P_{jk}(T) q_k \tag{16}$$

where $S_{ij}(T)$ is obviously symmetric.

On substituting (15) in (16) we obtain

$$S_{ij}(T) = q_i q_j + (q_i q_j)^{1/2} \sum_{r=1}^{3} e^{2T\lambda_r} w_i^{(r)} w_j^{(r)} \tag{17}$$

This expression suggests the introduction of a new matrix

$$\begin{aligned} M_{ij}(T) &= \frac{S_{ij}(T)}{(q_i q_j)^{1/2}} \\ &= (q_i q_j)^{1/2} + \sum_{r=1}^{3} e^{2T\lambda_r} w_i^{(r)} w_j^{(r)} \end{aligned} \tag{18}$$

Equation (18) relates the experimental data M_{ij} (see the next section) to the evolution time T and to the structure of the rate matrix R through the eigenvalues λ_r and the vectors $w_i^{(r)}$.

Finally, defining the average substitution rate per site for unit time

$$v_s = \sum_{\substack{i,k \\ i \neq k}} R_{ik} q_k \tag{19}$$

we get easily

$$v_s = -\sum_{i,r} q_i w_i^{(r)} w_i^{(r)} \lambda_r \tag{20}$$

Data Analysis

The matrix $S_{ij}(T)$ defined in Eq. (16) is obtained from the data through the following procedure. (1) Consider in the codon sequences all silent positions relative to the compared species (B, C) and record the nucleotides (i, j) that occupy such positions. (2) Count the number of times that the nucleotide i in the sequence of B has become the nucleotide j in the sequence of C; this number shall be denoted by N_{ij}. (3) Calling the length of the sequences L, compute the matrix S_{ij} as

$$S_{ij} = \frac{N_{ij}}{L} \tag{21}$$

This procedure is repeated by interchanging C and B. Within the statistical fluctuations due to the finiteness of L, one finds that $N_{ij} \simeq N_{ji}$, thus implying that S_{ij} is (approximately) symmetric, consistent with the model discussed in the preceding section. This check, together with the one on the constancy of the frequencies q_i of the single bases, is thus fundamental to the meaningfulness of our analysis.

Having constructed, by the above procedure, S_{ij} and consequently M_{ij} [Eq. (18)], one proceeds to the diagonalization of the latter matrix and to the determination of its eigenvalues $\Lambda_r = e^{2T\lambda_r}$ and eigenvectors $w^{(r)}$. To determine the rate matrix R_{ik} according to Eq. (14), we must then extract the eigenvalues

$$\lambda_r = \frac{1}{2T} \log \Lambda_r \tag{22}$$

from the eigenvalues of M_{ij}. As remarked in the Introduction, our method needs as input time T for a pair of species, or if we have p pairs of species we must get

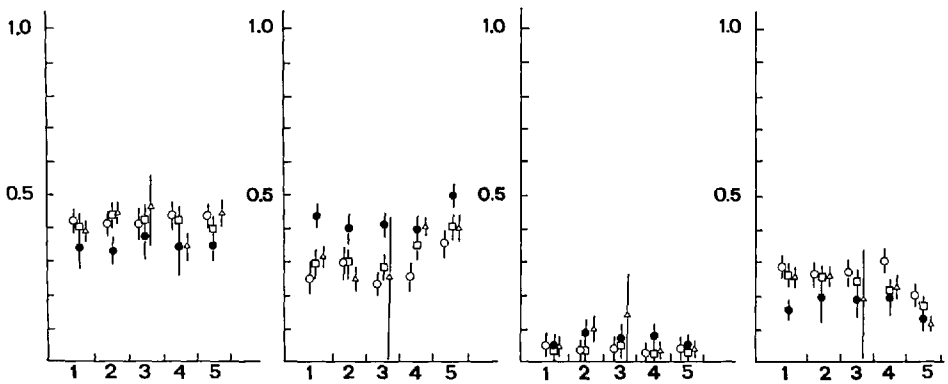

Fig. 1. The frequencies of the four bases (left to right: A, C, G, T) at the silent positions in the genes CoI (1), CoII (2), ATPase (3), CoIII (4) and Cyt b (5) for rat (□), mouse (○), cow (△) and human (●). The error bars give the standard deviation

$$\lambda_r = \frac{1}{2T_1} \log \Lambda_r^{(1)} = \frac{1}{2T_2} \log \Lambda_r^{(2)} = \ldots = \frac{1}{2T_p} \log \Lambda_r^{(p)} \tag{23}$$

for every pair. The $\Lambda_r^{(a)}$ $\{a = 1, \ldots, p\}$ being determined from experimental data, Eq. (23) fixes all other T_a's in terms of a single T, which can be taken from the most trustworthy set of paleontological data.

Excluding for the time being the presence of systematic errors, our analysis, as we have emphasized, is subject to statistical fluctuations in such a way that chain equalities like (23) must be understood within these fluctuations.

Thus, to proceed to a sensible interpretation of our experimental results it is necessary to evaluate quantitatively the size of the statistical fluctuations in the different sequence comparisons that we are going to carry out.

We analyse first the effects of fluctuations on the frequencies q_i. Given a sequence of length L and the number of times, N_i, that the base i appears in this sequence, we obviously have

$$q_i = \frac{N_i}{L}, \sum_i q_i = 1 \quad (i = 1, 2, 3, 4) \tag{24}$$

The statistical fluctuations are determined by generating, via a Monte Carlo method, a large number of sequences ($\sim$1000) of fixed length L by first extracting in a random way a value of i and then generating the populations N_j ($j \neq i$) according to a Poisson law with parameter $\bar{N}_j$ equal to the experimentally observed value. For the generic sequence we obviously have

$$N_i = L - \sum_{j \neq i} N_j \tag{25}$$

N_i's, as determined by Eq. (25) and by Poisson distributions for N_j ($j \neq i$), are not a priori guaranteed to be positive. We obviously reject those sequences for which N_i turns out to be negative. The average values $\bar{N}_i$ and the dispersions σ_i of the distributions of the N_i thus obtained determine q_i and its statistical error δq_i through the relations

$$q_i = \frac{\bar{N}_i}{L}, \tag{26}$$

and

$$\delta q_i = \frac{\sigma_i}{L} \tag{27}$$

Our results are reported in Fig. 1, where we have plotted the frequencies q_i of the four bases A, C, G, and T for the genes CoI, CoII, ATPase, CoIII and Cyt b of rat, mouse, cow and man. We observe that whereas for the triplet rat, mouse, cow the q_i's coincide, within the statistical fluctuations, roughly independently of the genes considered, for humans the q_i's are again gene independent but differ from those of the other three species (see also Table 1 and Table 2). This obviously means that our model cannot be applied to humans.

Although we find it rather surprising that our description works so well for the triplet of mammals considered and not for humans, it is conceivable that the considerable length of time separating humans from the other three mammals requires a more sophisticated model in which transient phenomena might play an important role. Thus, in the remaining part of this paper, we shall focus our attention on rat, mouse and cow only.

The observed gene independence of the q_i's suggests the usefulness of considering a "supergene" obtained by considering the genes as linked one after the other. In this way we can substantially reduce the statistical errors affecting our analysis, thus producing a determination of the various quantities relevant to the model, such as the q_i's, the divergence times T and the rate matrix R_{ij}, in a more precise fashion. From now on we shall consider only the supergene. In Table 2 we report the q_i's we ob-

Table 1. The frequencies (q_i) and standard deviations of the frequencies of the four bases at the silent position in the genes CoI, CoII, ATPase, CoIII and Cyt b[a]

Gene	Base			
	A	C	G	T
CoI	0.40 ± 0.03	0.29 ± 0.03	0.04 ± 0.02	0.26 ± 0.03
	0.42 ± 0.03	0.25 ± 0.03	0.04 ± 0.02	0.28 ± 0.03
	0.39 ± 0.03	0.31 ± 0.03	0.04 ± 0.02	0.25 ± 0.03
	0.35 ± 0.03	0.43 ± 0.03	0.05 ± 0.02	0.16 ± 0.03
CoII	0.42 ± 0.05	0.30 ± 0.04	0.03 ± 0.02	0.25 ± 0.04
	0.41 ± 0.04	0.29 ± 0.04	0.03 ± 0.02	0.26 ± 0.04
	0.43 ± 0.05	0.25 ± 0.04	0.10 ± 0.04	0.22 ± 0.04
	0.32 ± 0.05	0.39 ± 0.05	0.09 ± 0.04	0.19 ± 0.04
ATPase	0.43 ± 0.05	0.28 ± 0.04	0.05 ± 0.03	0.24 ± 0.03
	0.49 ± 0.05	0.23 ± 0.04	0.02 ± 0.03	0.26 ± 0.04
	0.44 ± 0.21	0.21 ± 0.20	0.15 ± 0.21	0.20 ± 0.20
	0.35 ± 0.05	0.40 ± 0.05	0.07 ± 0.04	0.18 ± 0.04
CoIII	0.42 ± 0.04	0.35 ± 0.04	0.02 ± 0.03	0.21 ± 0.04
	0.43 ± 0.04	0.25 ± 0.04	0.02 ± 0.03	0.30 ± 0.04
	0.33 ± 0.04	0.37 ± 0.04	0.03 ± 0.03	0.29 ± 0.04
	0.33 ± 0.04	0.40 ± 0.04	0.07 ± 0.04	0.19 ± 0.04
Cyt b	0.39 ± 0.03	0.40 ± 0.03	0.03 ± 0.03	0.16 ± 0.03
	0.43 ± 0.04	0.35 ± 0.04	0.02 ± 0.02	0.20 ± 0.03
	0.44 ± 0.04	0.39 ± 0.04	0.04 ± 0.03	0.12 ± 0.03
	0.33 ± 0.04	0.49 ± 0.04	0.05 ± 0.03	0.13 ± 0.03

[a] Values for rat, mouse, cow and human are listed from top to bottom within a gene

Table 2. The base frequencies and their standard deviations at the silent positions in the "supergene" for four mammalian species

Base	Species			
	Rat	Mouse	Cow	Human
A	0.41 ± 0.02	0.43 ± 0.02	0.41 ± 0.03	0.34 ± 0.02
C	0.33 ± 0.02	0.27 ± 0.01	0.32 ± 0.03	0.43 ± 0.02
G	0.03 ± 0.01	0.03 ± 0.01	0.06 ± 0.03	0.06 ± 0.01
T	0.22 ± 0.01	0.26 ± 0.01	0.21 ± 0.03	0.17 ± 0.02

tained by analysing the supergene in the four mammalian species.

We next determine the statistical errors on the matrix elements of the symmetric matrix S_{ij}. We start from the matrix N_{ij} experimentally determined in the way outlined above. Let us fix the jth column and extract randomly one value for the row i; we will call this value k. For each j and $i \neq k$ we Monte Carlo generate the nonnegative numbers N_{ij} according to a Poisson distribution with parameter $\bar{N}_{ij}$ coinciding with the experimentally observed value. N_{kj} is then determined by the relation

$$N_{kj} = N_j - \sum_{i \neq k} N_{ik} \tag{28}$$

We again reject those matrices for which one of their N_{kj} is negative. By repeating this procedure a large number of times (~1000) we can produce distributions for each matrix element N_{ij}, whose average value $\langle N_{ij} \rangle$ is the matrix we seek and whose dispersion σ_{ij} is the statistical error affecting that matrix element.

Having checked that within the statistical errors $\langle N_{ij} \rangle \cong \langle N_{ji} \rangle$ (see Table 3) we determine the matrix $\bar{S}_{ij}$ by the equation

$$\bar{S}_{ij} = \frac{1}{2L}[\langle N_{ij} \rangle + \langle N_{ji} \rangle] \tag{29}$$

attributing to it the statistical error $\delta \bar{S}_{ij} = \sigma_{ij}/L$.

A large number (~1000) of M_{ij} matrices is then generated by Monte Carlo methods according to a normal distribution of mean value $\bar{S}_{ij}$ and dispersion $\delta \bar{S}_{ij}$. Each matrix of this sample is then diagonalized, producing the eigenvalues Λ_r and the eigenvectors $w_i^{(r)}$. From the obtained distributions we extract the average eigenvalues $\bar{\Lambda}_r$ and eigenvectors $\bar{w}_i^{(r)}$ together with their dispersions.

In Table 4 we report, together with their statistical errors, the eigenvalues Λ_r for the comparisons rat–mouse, cow–mouse and cow–rat. Note that the errors reported are only a representation of the uncertainties affecting our determination, due to the non-normality of our distribution. We observe that, within the errors, the Λ_r's for the comparisons cow–rat and cow–mouse coincide. This fact corroborates the phylogenetic tree shown in Figure 2.

The high degree of consistency between our data and the predictions of our stochastic model allows us to calculate, given one divergence time T, the rate matrix R_{ij} and the statistical error affecting it. By

Table 3. The matrix N_{ij} defined in Eq. (25) for the couple rat–mouse[a]

	A	C	G	T
A	478.48 ± 18.8	51.23 ± 13.03	31.21 ± 11.12	50.11 ± 12.43
C	76.83 ± 12.79	227.24 ± 14.73	7.23 ± 8.16	177.28 ± 14.60
G	28.18 ± 4.69	6.42 ± 3.39	8.27 ± 3.63	7.98 ± 3.72
T	56.95 ± 10.86	121.76 ± 12.01	3.51 ± 4.95	149.63 ± 11.15

[a] Errors coincide with the standard deviations, through our Monte Carlo simulation

Table 4. Determination of the eigenvalues Λ_r $\{\Lambda_1 \leq \Lambda_2 \leq \Lambda_3\}$ from the comparisons of the "supergenes" for the couples rat–mouse (R–M), rat–cow (R–B) and mouse–cow (M–B)

	Λ_1	Λ_2	Λ_3
R–M	0.06 ± 0.04	0.15 ± 0.05	0.65 ± 0.03
R–B	0.01 ± 0.04	0.10 ± 0.04	0.50 ± 0.04
M–B	0.03 ± 0.04	0.12 ± 0.04	0.51 ± 0.03

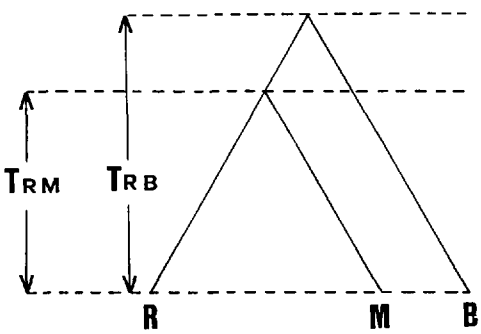

Fig. 2. The phylogenetic tree for rat, mouse and cow. T_{RM} is the divergence time between rat and mouse while T_{RB} denotes the divergence time between rodents and cow

Table 5. The rate matrix R_{ik} determined from the rat–mouse comparison by fixing $T_{RM} = 35$ MY[a]

	A	C	G	T
A	−5.09 ± 1.05	2.45 ± 1.24	22.17 ± 9.61	2.61 ± 1.55
C	1.76 ± 0.88	−17.25 ± 5.08	2.36 ± 10.49	18.34 ± 5.81
G	1.83 ± 0.92	0.29 ± 1.39	−28.04 ± 10.49	0.49 ± 1.76
T	1.47 ± 0.85	14.75 ± 4.72	3.49 ± 10.98	−21.42 ± 6.04

[a] All values should be multiplied by 10^{-9}

making use of the paleontological data on the divergence time T_{RM} between rat and mouse we obtain $T_{RM} \simeq 35$ million years (MY), which leads to the rate matrix reported in Table 5.

It is worth noticing that, due to the low content of G of the supergene ($q_G = 0.04 \pm 0.02$), the statistical errors on the transitions from G are necessarily quite large. However, from the data reported in Table 5, it emerges quite clearly that the rates of the transitions exceed those of transversions by at least an order of magnitude.

We have also determined the average substitution rate v_s, given by Eqs. (19) and (20), to be

$$v_s = (13.6 \pm 7.1) \cdot 10^{-9}/\text{site per year}. \quad (30)$$

Finally, we determine the divergence times for the couples rat–cow (T_{RB}) and mouse–cow (T_{MB}), which, if the phylogenetic tree reported above is correct, should coincide. In Figure 3 we report the determinations (A) of the ratio T_{MB}/T_{RM} and (B) of the ratio T_{RB}/T_{RM}, afforded by each of the three eigenvalues Λ_r. It is readily apparent that, within the statistical errors, our expectations are completely fulfilled. To obtain a more precise determination of T'/T_{RM} $\{T' = T_{RB}, T_{MB}\}$, we combined the different determinations $(T'/T)_r$ and their errors, σ_r, obtained from each eigenvalue, with weights p_r inversely proportional to σ_r^2. In this way we get

$$\frac{T_{RB}}{T_{RM}} = 1.45 \pm 0.19 \quad (31)$$

$$\frac{T_{MB}}{T_{RM}} = 1.35 \pm 0.16 \quad (32)$$

which, keeping T_{RM} fixed at 35 MY, yield

$$T_{RB} = 50.7 \pm 6.6 \text{ MY} \quad (31')$$

and

$$T_{MB} = 47.2 \pm 5.6 \text{ MY} \quad (32')$$

Introducing these values for T_{RB} and T_{MB} in Eq. (23), and making use of Eq. (14), we obtain the rate matrices R_{ik} for the rat–cow and the mouse–

cow comparisons, respectively, which we report in Table 6.

Discussion

The most relevant results of our work are contained in Figure 1 and Tables 1–6. They show that at least for some mammalian mt genes and in a time span shorter than 80 MY (the presumable divergence time between rodents and primates) the substitution process of the nucleotides at the third codon position follows a stationary Markov process. For divergence times larger than 50 MY the definite differences among q_i values calls for a substantial modification of the model, which would have to take into account a possible time dependence of the rate matrices. To this end we are studying a simple such model, which will be reported on in a future paper. Thus our discussion here concentrates on the triplet mouse–rat–cow.

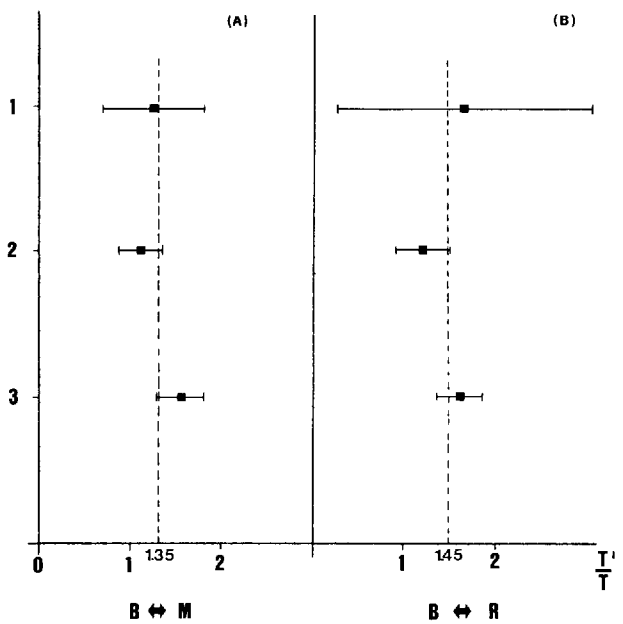

Fig. 3. The values for the divergence time ratios T'/T for A cow–mouse, and B cow–rat, rat–mouse obtained using the three different eigenvalues (1, 2, 3) taken in increasing order. The average value for R and M is equal to 1.40.

As emphasized in the previous sections of this paper, to obtain the substitution rate matrix for different couples of genes one needs as input the divergence time for one couple. Taking 35 MY as the divergence time between rat and mouse (Brown et al. 1979), we obtained the results shown in Eqs. (31′) and (32′). We emphasize that it is a nontrivial result that there exists a $T_{RB} = T_{MB} = 1.4\ T_{RM}$ whose determination is such that the rate matrices R_{RB} and R_{MB} turn out, within the statistical fluctuations, to be precisely the same (Table 6).

The high degree of consistency we find gives us confidence that, at least within the time span specified above, our method gives a reliable and general way to determine the evolutionary distance between different species. In particular, our result that $T_{RB} \cong 50$ MY turns out to be somewhat smaller than that generally accepted from paleontological data (see Brown et al. 1979). On the other hand, we should like to suggest that, in view of the considerable uncertainties and difficulties involved in dating fossil records, our method might be of definite help.

From Eq. (19) we calculated the average silent substitution rate in mammalian mt genes to be $v_s = 1.4 \times 10^{-8}$ substitutions per site per year. This value is about four times higher than the one previously calculated by some of us using the Kimura model (Saccone et al. 1983). This should not be too surprising, in view of the fact that the assumption upon which the Kimura model rests (Kimura 1981) is in fact not supported by our experimental determinations. This same critique appears to apply to the majority of models based on a priori assumptions about the rate matrices.

To compare the evolutionary rate of mt DNA with that of nuclear DNA it is necessary to perform the same analysis on the structure of nuclear genes. Even though the data on nuclear genes are unfor-

Table 6. The rate matrices R_{ik} obtained for the comparisons rat–cow (upper entry) and mouse–cow (lower entry)[a]

	A	C	G	T
A	−5.39 ± 1.41 −5.79 ± 1.42	2.73 ± 1.63 3.02 ± 1.33	16.31 ± 11.84 19.34 ± 12.19	3.24 ± 2.62 3.39 ± 1.66
C	2.10 ± 1.25 2.09 ± 0.92	−11.17 ± 4.58 −12.47 ± 3.28	1.88 ± 12.05 0.66 ± 8.45	12.35 ± 6.76 12.05 ± 4.36
G	1.64 ± 1.22 1.88 ± 1.23	0.22 ± 1.61 0.10 ± 1.24	−20.46 ± 11.72 −23.08 ± 12.03	0.41 ± 2.41 0.53 ± 1.71
T	1.52 ± 1.20 1.83 ± 0.90	7.94 ± 4.24 9.34 ± 3.30	1.89 ± 11.72 3.07 ± 9.07	−16.34 ± 7.52 −15.95 ± 4.69

[a] All values should be multiplied by 10^{-9}

tunately still too scanty, we have carried out a preliminary study on the growth hormone genes of rat and cow and we have found a v_s approximately one third the size of that for mt DNA genes.

We end this discussion by recalling the peculiar facts, made evident by our measurements and in a previous paper (Pepe et al. 1983) that the G content in the silent third codon position of mt DNA is very low and that the rate of replacement by G of bases other than G is not significantly different from zero. This property is shared not only by the mt genes of man but also by mt genes from lower eukaryotes (with the possible exception of plant mt genes). We believe that the low G content of the majority of the mt genes is a peculiar feature linked to the basic structure and expression of the mitochondrial genome.

Acknowledgments. This work was supported by the Progetto Finalizzato Ingegneria Genetica e Basi Molecolari delle Malattie Ereditarie of the CNR.

References

Anderson S, Bankier AT, Barrell GB, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, 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) Complete sequence of bovine mitochondrial DNA. J Mol Biol 156:683–717

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, George M Jr, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci USA 76: 1967–1971

Cantatore P, De Benedetto C, Gadaleta G, Gallerani R, Kroon AM, Holtrop M, Lanave C, Pepe G, Quagliariello C, Saccone C, Sbisà E (1982) The nucleotide sequences of several tRNA genes from rat mitochondria: common features and relatedness to homologous species. Nucleic Acids Res 10:3279–3289

Gojobori T, Ishii K, Nei M (1982) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J Mol Evol 18:414–423

Grosskopf R, Feldman H (1981) Analysis of a DNA segment from rat liver mitochondria containing the genes for the cytochrome oxidase subunits I, II and III, ATPase subunit 6 and several tRNA genes. Curr Genet 4:151–158

Grunstein M, Schede P, Kedes L (1976) Sequence analysis and evolution of sea urchin (*Lytechinus pictus* and *Strongylocentrotus purpuratus*) histone H4 messenger RNAs. J Mol Biol 104:351–369

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian Protein Metabolism, Vol III. Academic Press, New York, pp 21–132

Jukes TH (1980) Silent nucleotide substitutions and the molecular evolutionary clock. Science 210:973–978

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

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Kobayashi M, Seki T, Yaginuma K, Koike K (1981) Nucleotide sequences of small ribosomal RNA and adjacent transfer RNA genes in rat mitochondrial DNA. Gene 16:297–307

Koike K, Kobayashi M, Yaginuma K, Taira M, Yoshida E, Imai M (1982) Nucleotide sequence and evolution of the rat mitochondrial cytochrome b gene containing the *ochre* termination codon. Gene 20:177–185

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

Pepe G, Holtrop M, Gadaleta G, Kroon AM, Cantatore P, Gallerani R, De Benedetto C, Quagliariello C, Sbisà E, Saccone C (1983) Non-random patterns of nucleotide substitutions and codon strategy in the mammalian mitochondrial genes coding for identified and unidentified reading frames. Biochem Intern 6:553–563

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

Saccone C, Cantatore P, Gadaleta G, Gallerani R, Lanave C, Pepe G, Kroon AM (1981) The nucleotide sequence of the large ribosomal RNA gene and the adjacent tRNA genes from rat mitochondria. Nucleic Acids Res 9:4139–4148

Saccone C, De Benedetto C, Gadaleta G, Lanave C, Pepe G, Sbisà E, Cantatore P, Gallerani R, Quagliariello C, Holtrop M, Kroon AM (1983) Studies on the evolutionary history of the mammalian mitochondrial genome. In: Nagley P, Linnane AW, Peacock WJ, Pateman JA (eds) Manipulation and Expression of Genes in Eukaryotes. Academic Press, Sydney, pp 325–332

Salser W, Bowen S, Browne D, Eladli F, Fedoroff N, Fry K, Heindell H, Paddock G, Poon R, Wallace B, Whitcome R (1976) Investigation of the organization of mammalian chromosomes at the DNA sequence level. Fed Proc 35:23–35

Sekiya T, Kobayashi M, Seki T, Koike K (1980) Nucleotide sequence of a cloned fragment of rat mitochondrial DNA containing the replication origin. Gene 11:53–62

Takahata N, Kimura M (1981) A model of evolutionary base substitutions and its applications with special reference to rapid change of pseudogenes. Genetics 98:641–657

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Annu Rev Biochem 46:573–639

Received May 15, 1983/Revised September 21, 1983