J Mol Evol (1991) 32:6–15

Journal of Molecular Evolution

Codon Usage Divergence of Homologous Vertebrate Genes and Codon Usage Clock

Manyuan Long and John H. Gillespie

Department of Genetics, University of California, Davis, CA 95616, USA

Summary. This paper is concerned with the divergence of synonymous codon usage and its bias in three homologous genes within vertebrate species. Genetic distances among species are described in terms of synonymous codon usage divergence and the correlation is found between the genetic distances and taxonomic distances among species under study. A codon usage clock is reported in alpha-globin and beta-globin. A method is developed to define the synonymous codon preference bias and it is observed that the bias changes considerably among species.

Key words: Codon usage — molecular clock — Genetic distance — Vertebrate

Introduction

It is well known that synonymous codons are not used with equal frequencies (Watson 1976; Grantham et al. 1980, 1981; Ikemura 1985; Maruyama et al. 1986). Based on a multivariate analysis of codon usage data from unicellular organisms, Grantham et al. (1980) proposed the genome hypothesis implying some relationship between codon usage and taxonomic distance. Ikemura (1985) and Maruyama et al. (1986) noticed a correlation between taxonomic divergence and the similarity of the codon dialect. In the last two years, there has been an increasing number of publications that explore codon usage evolution in vertebrates (e.g., Bernardi et al. 1988; Filipski 1988; Sueoka 1988; Wolfe et al. 1989). These studies increase our understanding of how codon usage evolves and what mechanisms are responsible. However, as more and more DNA sequence data become available, it is necessary to extend these results.

At first, the conclusions about correlations came from qualitative observations and the studies seemed not to isolate exactly the information of synonymous codon usage from amino acid usage. This cannot lead us to a precise knowledge of the problem. On the other hand, the comparison of nonhomologous gene data sometimes leads to wrong conclusions when examining the relationship between codon usage and taxonomic distance. We find this is so while checking Maruyama et al.'s tabulation. Mouchiroud and Gautier (1988) observed high codon usage changes in mammalian genes by examining the differences in G+C percentage at the third position of codons among species and the asymmetry of differences between the sequences. This suggests the possibility of describing interspecific divergence in terms of synonymous codon usages within vertebrate species and to confirm if such a difference corresponds to taxonomic distance. Finally, Bernardi et al. (1988) reported two interesting evolutionary modes of compositional patterns in vertebrate genomes, namely, the conservative mode and the shifting mode. They found that, in the first mode, the similar compositional patterns could be maintained over a relatively long period (many million years) even though a lot of substitutions accumulated; in the shifting mode, on the contrary, the compositional patterns were observed to change quickly. From their findings, furthermore, we can ask a question relevant to our study: how codon

Table 1. Homologous DNA sequences

Species	Accession no.	Species	Accession no.
Alpha-globin			
Human	J00153	Mouse	J00410
Chimpanzee	X00226	Chicken	J00852
Orangutan	M12157	Duck	X02008
Horse	X01086	*Xenopus*	X02796
Goat	J00043	Salamander	M13365
Rabbit	J00658		
Beta-globin			
Human	J00179	Mouse	J00413
Lemur	M15734	Chicken	J00860
Rabbit	J00659	Duck	†
Lepus	Y00347	*Xenopus*	J00978
Bovine	X00376	*Rana*	M19548
Goat	*		
Insulin			
Human	J00265	Rat	J00747
Monkey	J00336	Guinea pig	K02233
Dog	J00042	Carp	K00036

* Data from Lingrel et al. (1983)
† Data from Hampe et al. (1981)

usage undergoes evolution within a long period of compositional stability.

The present paper aims at describing codon usage divergence among vertebrate species by appropriate statistical techniques with homologous genes and revealing the possible relationship between codon usage difference and taxonomic distance. We will show a new kind of molecular clock that might be called the codon usage clock. Meanwhile, we will analyze the relation of the genetic distances—our description of codon usage divergences among species with the isochore proposed by Bernardi et al. (Bernardi and Bernardi 1985, 1986; Bernardi et al. 1985), defined by GC contents. Moreover, we will develop a method to describe the extent of the bias of synonymous codon choice in a gene, which is more simple than other methods (e.g., McLachlan et al. 1984; Sharp and Li 1987). By this method, we will investigate the difference of codon preference bias among species.

Methods

Data Preparation. All of the DNA sequences in this study were obtained from the GenBank Genetic Sequence Data Bank (Bilofsky et al. 1986) except for goat and duck beta-globin. Three homologous sequences that have been sequenced from many vertebrate species were chosen: alpha-globin with 11 species, beta-globin with 11 species, and insulin with 6 species. The sequences are listed in Table 1.

The codon usage for each sequence was calculated from a computer program in the database (item 9 of GenBank Menu). These numbers were then converted to frequencies in units of individual amino acids. For example, the 13 glycine residues in alpha-globin are partitioned into four synonymous codons:

	Codon	Count	Frequency
Gly:	GGA	0	0
	GGC	8	0.6154
	GGG	1	0.0769
	GGT	4	0.3077

Obviously, different codon frequencies in a given amino acid show different codon preference or codon usage. Once we find there are different percentages for a particular codon in an amino acid between homologous sequences of two species under study, we know that the difference is due to the different synonymous codon choices, not because of different amino acid usages. In this way, we isolate the codon usage information completely from the amino acid usage information.

Interspecific Divergence Measurement. Nei's (1972) genetic distance is employed to measure interspecific divergence of synonymous codon usage. Assume species 1 and species 2 have a pair of homologous DNA sequences consisting of 20 amino acids, each amino acid with t synonymous codons. Suppose that the frequency of synonymous codon C_i at an amino acid A_j is x_{ij} and y_{ij} ($i = 1, 2, \ldots t$) in the two species, respectively. Then, calculate

$$J_{j11} = \sum_{i=1}^{t} x_{ij}^2,\ J_{j22} = \sum_{i=1}^{t} y_{ij}^2,\ \text{and}\ J_{j12} = \sum_{i=1}^{t} x_{ij}y_{ij}$$

where 1 and 2 in the subscripts of J_{j11}, J_{j22}, and J_{j12} mean the species 1 and 2, respectively. Take the average over all s amino acids that have synonymous codons (usually s = 18, except for mitochondrial genes and certain genes encoding proteins without some amino acids).

$$J_{11} = \frac{1}{s}\sum_{j=1}^{s} J_{j11},\ J_{22} = \frac{1}{s}\sum_{j=1}^{s} J_{j22},\ \text{and}\ J_{12} = \frac{1}{s}\sum_{j=1}^{s} J_{j12}$$

Thus, a statistic, which may be called the similarity coefficient between species 1 and species 2, can be defined as

$$I = \frac{J_{12}}{\sqrt{J_{11}J_{22}}}$$

The genetic distance is given by the formula

$$D = -\ln(I) = -\ln\left(\frac{J_{12}}{\sqrt{J_{11}J_{22}}}\right)$$

which describes the divergence between two species in synonymous codon usage.

It needs to be indicated that an elegant population genetic model underlies the original genetic distance of Nei. In this paper, however, that model seems no longer to hold. Hence, we are unable to use some of the original arguments such as D = 2tu (t = divergent time between two species and u = substitution rate) to explain D values obtained in our research. Fortunately, we notice that Nei's distance has a mathematic property such that it can describe the divergence between two sets of frequency tables. The more different the two sets of frequency tables, the bigger the distance value. Thus, it is reasonable to use Nei's distance in an empirical description of codon usage divergence among species.

Bias Measurement of Synonymous Codon Preference. In population genetics, the notation J_{ii} in the above formula stands for the average genetic homozygosity of population i, i.e., the fraction of loci that are homozygous. In other words, from the magnitude of J_{ii}, we can know whether the existing alleles at a locus (or

average over a group of loci) are evenly distributed or not. Similarly, in our study J_{ii} reflects the synonymous codon choice bias. The larger the J_{ii}, the greater the bias. However, we cannot apply it to comparing different genes directly because of its changing minimum even though the maxima are always unity. The J_{ii} have a minimum when synonymous codons are evenly distributed. The minimum is changing because different genes have different amino acid compositions. Different amino acids have different minimum value $J_{min.j}$ (here subscript j stands for the amino acid A_j). For example, sextet amino acids (e.g., arginine) have $J_{min.j} = 6x(1/6)^2 = 1/6$, whereas quartet amino acids (e.g., alanine) have $J_{min.j} = 4x(1/4)^2 = 1/4$ and duet amino acids (e.g., lysine) have $J_{min.j} = 2x(1/2)^2 = 1/2$. Now, let us define

$$J_{min.} = \frac{1}{s}\sum_{j=1}^{s} J_{min.j} = \frac{1}{s}\left(s_2\cdot\frac{1}{2} + s_3\cdot\frac{1}{3} + s_4\cdot\frac{1}{4} + s_6\cdot\frac{1}{6}\right)$$

where s_2, s_4, and s_6 are the number of amino acid types in duet, quartet, and sextet amino acid (for their definitions, see Grantham et al. 1981) groups in a particular gene, respectively. The s_3 value depends on whether isoleucine is available ($s_3 = 1$) or not ($s_3 = 0$) in this gene. In alpha-globin and beta-globin, because many species lack isoleucine, this amino acid is not included in our analyses. Note $s_2 + s_3 + s_4 + s_6 = s$. We can standardize J_{ii} by $J_{min.}$. Thus, we are able to develop a formula to measure the bias of synonymous codon choice as follows

$$b = \frac{J}{J_{min.}} = \frac{\sum_{j=1}^{s} J_j}{\sum_{j=1}^{s} J_{min.j}}$$

where J has same meaning of J_{ii}. By this formula, we can compare the bias difference of codon preference in various genes. b has a minimum of unity when synonymous codons are evenly distributed. When b is larger than one, that implies existence of synonymous codon preference. b increases while the bias increases. The maximum of b is equal to $J_{min.}^{-1}$, i.e., it depends on amino acid composition. Thus, b will take a value in the range 1 to $J_{min.}^{-1}$. Furthermore, we can transform the statistic b into B

$$B = \frac{1}{J_{min.}^{-1} - 1}(b - 1) = \frac{1}{1 - J_{min.}}(J - J_{min.})$$

such that the measurement of bias can take its value in the range 0 to 1. Obviously, B = 0 means no bias whereas B = 1 implies maximum bias. We are going to use B, which we will call standardized synonymous codon bias, to describe the degree of bias in the sequences under study.

Clustering Analysis. After genetic distances were calculated in terms of synonymous codon usage divergence among species, a phylogenetic classification was made using a clustering algorithm for the purpose of comparison with other well known molecular phylogenies (e.g., Goodman et al. 1985). The hierarchical clustering method is used for this purpose. Suppose there are n species with $\frac{1}{2}n(n-1)$ distance values. At the beginning, let n species be n clusters. Then, group the two species with the shortest distance as a new cluster. After this is done, define distances between the new cluster and old clusters by the formula

$$D_{ir} = \frac{n_p}{n_r} D_{ip} + \frac{n_q}{n_r} D_{iq}$$

where D_{ir} is the defined distance between new cluster r consisted of cluster p and q and the old cluster i, n_p and n_q are numbers of the distances included in the clusters p and q ($n_r = n_p + n_q$). This means D_{ir} is weighted as an average distance of D_{ip} and D_{iq}. Group the two clusters with the shortest distances again and repeat the above procedures until all species are grouped together. The whole clustering process can be expressed as a phylogenetic graph. From this graph, we can easily understand the relationship among species in terms of similarity (see, e.g., Fitch and Margoliash 1967).

Relationship between Genetic Distance and Divergence Time. The divergence time data from Romer (1966), Hickman et al. (1979), Goodman et al. (1982, 1985), Romero-Herrera et al. (1982) and Britten (1986) were used to analyze the possible relation.

Relative Rate Test. In order to test the clocklike divergent pattern, this test (Wilson et al. 1977) was used. The basic appoach in this test is to check if genetic distances between two within-group species and an outgroup species are equal or not. The formula used to test the difference of two genetic distances is

$$U = \frac{D_{xz} - D_{yz}}{\sqrt{Var(D_{xz}) + Var(D_{yz}) - 2Cov(D_{xz}, D_{yz})}}$$

where the notations x and y stand for two within-group species and z for the third outgroup species. If U is significantly larger than zero, the two genetic distances D_{xz} and D_{yz} can be thought different. Here the distribution of U is assumed to be normal. The probability level chosen is 0.001 because of the large number of tests involved. Nei and Roychoudhury (1974) and Mueller and Ayala (1982) derived the variance and covariance of genetic distances as follows

$$Var(D_{xz}) = \frac{Var(J_x)}{4J_x^2} + \frac{Var(J_z)}{4J_z^2} + \frac{Var(J_{xz})}{J_{xz}^2} + \frac{Cov(J_x, J_z)}{2J_xJ_z} - \frac{Cov(J_x, J_{xz})}{J_xJ_{xz}} - \frac{Cov(J_z, J_{xz})}{J_zJ_{xz}}$$

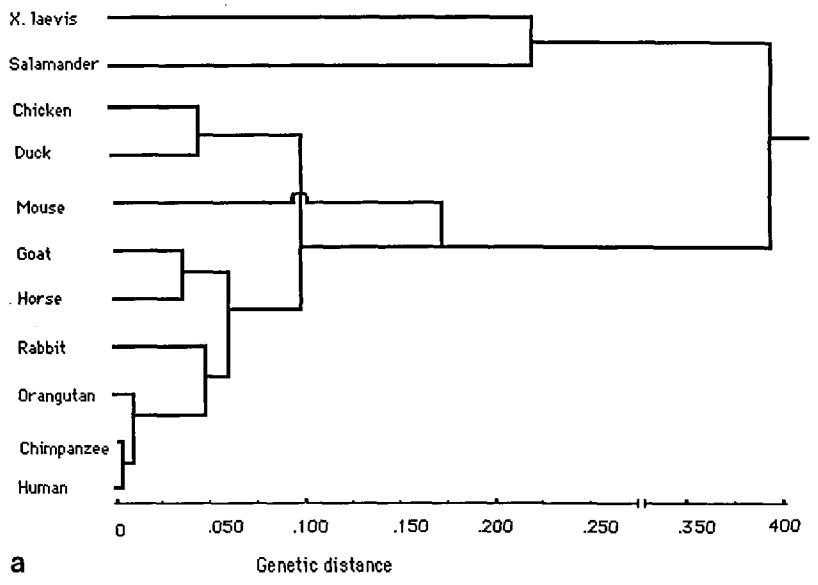
where $Var(J_x) = Var(j_x)/s$, $Var(J_{xz}) = Var(j_{xz})/s$, $Cov(J_x, J_y) = Cov(j_x, j_z)/s$, the j_x and j_{xz} are J_{jxx} and J_{jxz} in the context, respectively, and

$$Cov(D_{xz}, D_{yz}) = \frac{Cov(J_x, J_z)}{4J_xJ_z} + \frac{Cov(J_x, J_y)}{4J_xJ_y} - \frac{Cov(J_x, J_{zy})}{2J_xJ_{yz}} + \frac{Var(J_z)}{4J_z^2} + \frac{Cov(J_z, J_y)}{4J_zJ_y} - \frac{Cov(J_z, J_{yz})}{2J_zJ_{yz}} - \frac{Cov(J_{xz}, J_z)}{2J_{xz}J_z} - \frac{Cov(J_{xz}, J_y)}{2J_{xz}J_y} + \frac{Cov(J_{xz}, J_{yz})}{J_{xz}J_{yz}}$$

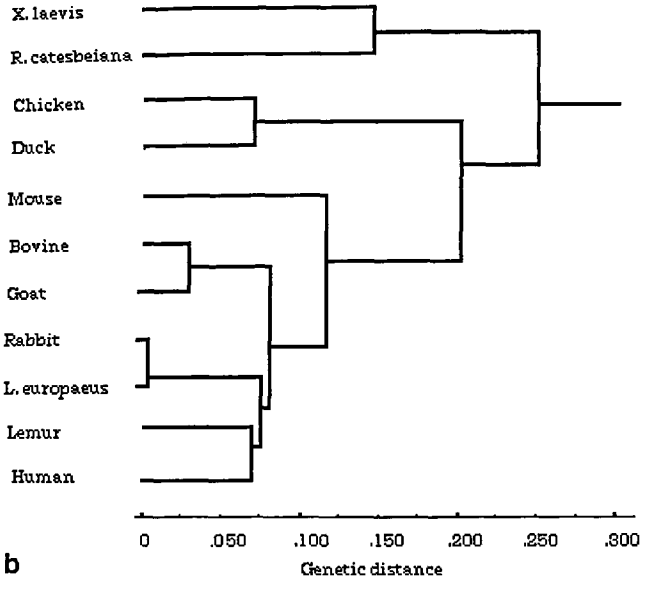
One problem, which could be asked is about the sample size, namely, the number of occurrences of each amino acid in the protein. In alpha-globin and beta-globin, the average occurrence number is about 8.5. Is the variance formula above still valid for this relatively small sample size? Nei and Roychoudhury (1974) discussed the sample size in detail after they completed their derivation of sampling variances of genetic distance. They came to the conclusion that the sampling size could be relatively small. Even so, the deviation from the original statistical model cannot be rejected completely. We can just use the formulae above to do an approximate test rather than an exact test.

Results

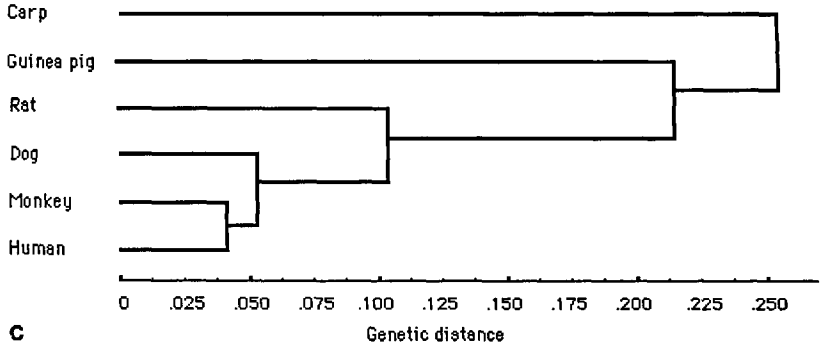
The codon usage tables of three homologous genes, i.e., alpha-globin, beta-globin, and preproinsulin, were made (not shown). From these tables, we can compare codon usage similarity among species qualitatively and see some common usage patterns. For instance, we can easily find that the codon usage in human alpha-globin is very close to chimpanzee. Their similarity is far greater than human and other nonprimate species. This is not a surprising obser-



a



b



c

Fig. 1. Hierarchical clustering graph. **a** Alpha-globin, **b** beta-globin, **c** preproinsulin.

vation, as their silent substitution rate itself was found to be very low (Li et al. 1987). Orangutan is a bit more different from human than the chimpanzee is from human, but this difference is not so big as human with a nonprimate species. Additionally, we can see that for almost every amino acid of the three genes, most species share some common choice pattern with slightly different usage frequency. In the alpha-globin table, e.g., we see that every species favors codon CTG for leucine, TCC for serine, ACC for threonine, and so on. These initial observations, which appear trivial, imply that we

can trace back evolutionary footsteps of species from the codon usage of homologous genes.

Based on the codon usage tables, we calculate genetic distances of three homologous genes among species and synonymous codon selection intensity, i.e., the standardized synonymous codon bias degrees, of each homologous gene (Table 2a–c). Some interesting things appear when we examine the results in this table. It seems that taxonomically close species usually have small genetic distances or higher identity. For instance, in Table 2a, the average distance within primate species is only 0.0053, whereas the primate has a 15 times larger distance value (= 0.083) with other mammal species (horse, goat, and rabbit), 17 times greater average distance value (= 0.089) with chicken and duck, and 74 times greater average distance (= 0.393) with the two amphibian species.

The results of hierarchical clustering analyses based on the distances among species are shown in Fig. 1a–c. On the whole, these results agree with Goodman et al. (1985) except mouse in alpha-globin and show the phylogenetic history of the species. Thus, basically, there is a close correlation between synonymous codon choice and taxonomic origin. In codon usage of alpha-globin, mouse is very different from mammals. Bernardi et al. (1988) have already observed that there is a genome compositional shift in murids (see their Fig. 1) so that mouse and other mammal alpha-globins are grouped into different isochores. For this reason, Saccone et al. (1989) believed that in evolution of molecules such as alpha-globin rodents just showed a pseudovelocity and hence rodents should be rejected in determinations of phylogeny. Also in Fig. 1c, guinea pig and rat (both rodents) are not grouped together. They show an unusually high divergence (their D = 0.229) against each other and with other mammal species, and guinea pig shows a very high evolutionary speed. This observation at codon usage level may not be coincident, recalling the well known high evolutionary rate of insulin of guinea pigs at the protein level (e.g., Kimura 1987).

Furthermore, we observed an approximately clocklike pattern of synonymous codon substitutions in alpha-globin, beta-globin, and insulin (Fig. 2) when we plotted the average genetic distance values of various species with their within-group species against their divergence time. The three regression equations obtained by the least-squares method are, approximately,

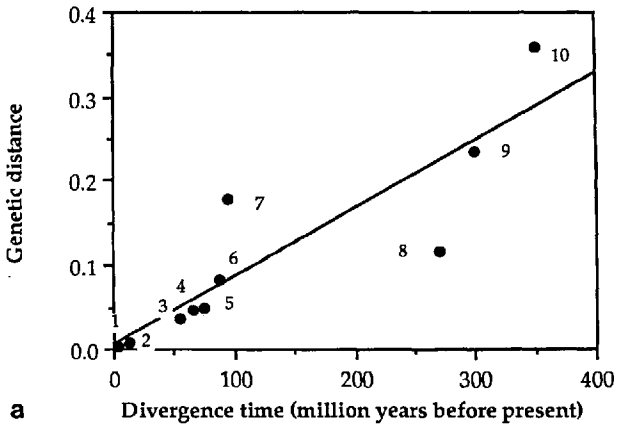
$$D_1 = 8.0 \times 10^{-4} T_1,$$
$$D_2 = 6.4 \times 10^{-4} T_2, \text{ and}$$
$$D_3 = 0.11 + 6.3 \times 10^{-4} T_3$$

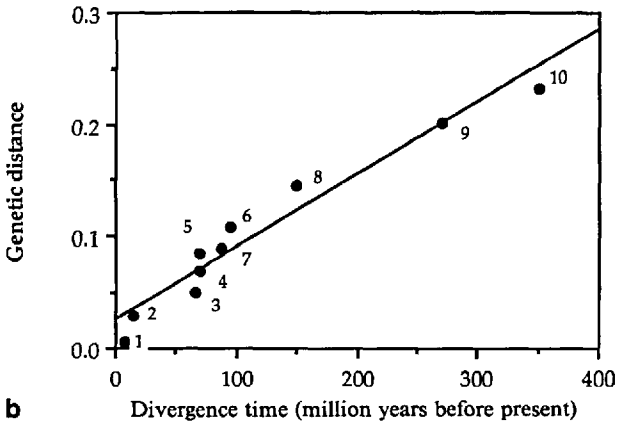
where D_1, T_1, D_2, T_2, and D_3, T_3 are divergence time and genetic distance of alpha-globin, beta-globin, and insulin, respectively. The former two equations give the evolutionary rates of silent codon usage of two proteins as 0.80×10^{-9} and 0.64×10^{-9} per year, with alpha-globin evolving faster than beta-globin. In addition, the coefficients of determination of regression in two equations are high ($R_1^2 = 0.775$ and $R_2^2 = 0.947$) such that the clocklike patterns are obvious even though we cannot make a statistical test of correlation because of the nonindependent distribution among the points in the graphs. Note that point 7 in Fig. 2a, consistent with the observation of Mouchiroud and Gautier (1988) of significant codon usage change between human and mouse, shows much bigger distance than the average evolutionary rate between mouse and other mammals. This is due to the effect of compositional shift (Bernardi et al. 1988) of rodents on the codon usage. Also point 8 in the same figure, the divergence between birds and mammals, is relatively low. The reason for this remains unknown. As to the third sequence, insulin, its clock behavior seems to be confirmed using more species data.

The relative rate tests were made on the alpha- and beta-globin, which show clocklike evolution, based on their phylogeny in Fig. 1a and b. Among 146 tests of alpha-globin not one showed a U value larger than 3.29 (i.e., 0.001 probability level). The tests of beta-globin show a similar result. The tests do not allow us to reject clocklike evolution, although there may not be much power for the test as evidenced by our failure to reject the mouse speedup.

Using the method we defined to examine the bias of synonymous codon preference, we find that the bias is large. In Table 2, the biases of alpha-globin in all species have B values from 0.597 to 0.165 with an average of 0.467, whereas the biases of beta-globin have B values from 0.513 to 0.155 with an average of 0.350 and the bias of insulin is also high (B values from 0.615 to 0.176 with an average of 0.419). A statistical test of the bias based on a likelihood ratio (see G-test of Sokal and Rohlf 1981) shows that the bias is significantly high in all three cases.

Finally, we analyzed the relation of codon usage divergence with isochore evolution, defined by the G+C percentage changes at the third codon position of alpha-globin and beta-globin (Fig. 3). Two graphs show that codon usage evolution occurs within genome compositional constraints. In Fig. 3b, especially obvious is that the codon usages of mammals except mouse, of duck and chicken, and of toad and frog have considerable amounts of divergence even though there is almost no change in the G+C percentage. On the other hand, from the two figures, we notice that the G+C content difference at the





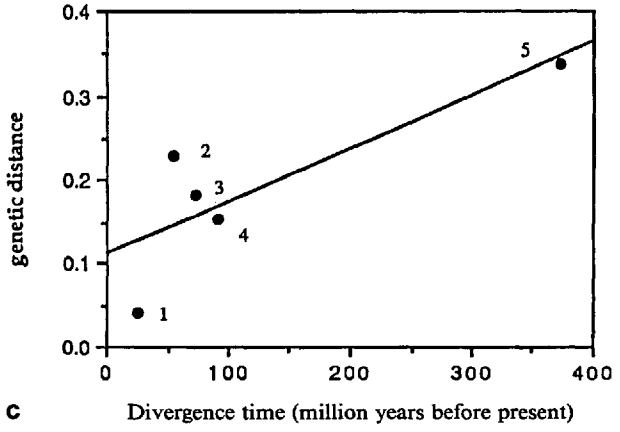


Fig. 2. The relationship between divergence time and genetic distance. **a** Alpha-globin. Point 1 shows human vs chimpanzee; 2, orangutan vs human and chimpanzee; 3, horse vs goat; 4, chicken vs duck; 5, rabbit vs primates; 6, horse, goat vs mouse, rabbit and primates; 7, mouse vs rabbit and primates; 8, chicken, duck vs mammals; 9, salamander vs *Xenopus laevis*; 10, amphibians vs chicken, duck, and mammals. **b** Beta-globin. Point 1 shows rabbit (*Oryctolagus cuniculus*) vs hare (*Lepus europaeus*); 2, bovine vs goat; 3, human vs lemur; 4, rabbit vs primates; 5, chicken vs duck; 6, mouse vs rabbit and primates; 7, bovine, goat vs mouse, rabbit, hare and primates; 8, bull frog (*Rana catesbeiana*) vs *Xenopus laevis*; 9, chicken, duck vs mammals; 10, amphibians vs mammals, chicken, and duck. **c** Insulin. Point 1 shows human vs monkey; 2, rat vs guinea pig; 3, rodents vs primates; 4, dog vs rodents and primates; 5, carp vs mammals.

third codon position indeed contributes to codon usage divergence among species. However, because of a broad range of codon usage variation available for the given G+C percentage, we also should be cautious when we attempt to use the G+C percentage at the third codon position to describe codon usage even if what we compare is a group of homologous sequences just as we did in this study.

Discussion

We observed codon usage clocks in alpha-globin and beta-globin of vertebrates. Seemingly, the codon usage clock we propose in this paper may be gene specific, because for some genes, for instance, mitochondrial genes, Lanave et al. (1984) found that the codon usage in humans is far apart from the one in other mammals. Also, we found that the synonymous codon preference biases in three molecules are large and differ among species. Thus each gene cannot be in a steady state for the bias degree in different species. These results appear when we employed homologous loci across species and described quantitatively the synonymous codon usage divergence of species.

Traditionally, the molecular clock is given as evidence for the neutral allele theory, although it is not clear whether the theory predicts a clock-time or a generation-time dependent clock. Thus, the observation of the codon usage clock would be thought to conform with the prediction of neutral theory. It is true that the neutrality in silent site evolution cannot be rejected for some reasons (Gillespie 1989). However, clocklike behavior is also possible even when natural selection is operating (Gillespie 1986, 1989). As far as we know, there are two kinds of explanations for a molecular clock on the grounds of natural selection. One is Lewontin's (1974) argument that the apparant constant rate of evolution is just an average of a changing evolutionary rate over millions of generations. This is possible if we assume a stationary distribution of evolutionary rate. Another is the Red Queen hypotheses (Van Valen 1974), although this does not provide a quantitative prediction about the clock. Thus, how to explain the clock quantitatively still remains unsolved even though we believe the clock is accountable from the viewpoint of natural selection. Meanwhile, it may be helpful to emphasize that the codon usage clock and corresponding codon usage distance are quite different from the general molecular clock and divergence based on the number of substitutions obtained from alignment of homologous sequences. The codon usage defined in this paper is used to describe what proportions of different synonymous

Table 2a. Estimation of D, I, and B for alpha-globin among vertebrate species

	Human	Chim-panzee	Orang-utan	Mouse	Rabbit	Horse	Goat	Chicken	Duck	Sala-mander	*X. laevis*
Human	**0.597**	0.003	0.005	0.183	0.047	0.052	0.045	0.083	0.095	0.488	0.304
Chimpanzee	0.997	**0.587**	0.008	0.179	0.052	0.059	0.047	0.093	0.089	0.496	0.298
Orangutan	0.995	0.992	**0.579**	0.179	0.052	0.053	0.051	0.078	0.094	0.475	0.296
Mouse	0.833	0.836	0.836	**0.434**	0.161	0.192	0.174	0.202	0.200	0.368	0.270
Rabbit	0.954	0.949	0.949	0.851	**0.559**	0.097	0.061	0.115	0.151	0.402	0.305
Horse	0.949	0.943	0.948	0.825	0.908	**0.555**	0.037	0.077	0.118	0.456	0.261
Goat	0.958	0.854	0.950	0.840	0.941	0.964	**0.518**	0.106	0.117	0.468	0.256
Chicken	0.920	0.911	0.925	0.814	0.891	0.926	0.899	**0.434**	0.047	0.351	0.272
Duck	0.909	0.915	0.910	0.819	0.860	0.889	0.890	0.954	**0.423**	0.390	0.282
Salamander	0.614	0.609	0.622	0.692	0.669	0.634	0.626	0.704	0.677	**0.165**	0.236
X. laevis	0.738	0.742	0.744	0.763	0.737	0.770	0.774	0.762	0.754	0.790	**0.284**

Table 2b. Estimation of D, I, and B for beta-globin among vertebrate species

	Human	Lemur	Rabbit	Mouse	Bovine	*L. euro-paeus*	Goat	Chicken	Duck	Bullfrog	*X. laevis*
Human	**0.429**	0.086	0.058	0.132	0.073	0.062	0.067	0.247	0.272	0.171	0.236
Lemur	0.918	**0.380**	0.035	0.079	0.095	0.043	0.070	0.173	0.226	0.196	0.282
Rabbit	0.944	0.966	**0.392**	0.113	0.093	0.006	0.074	0.186	0.233	0.173	0.292
Mouse	0.876	0.924	0.893	**0.332**	0.126	0.124	0.118	0.163	0.151	0.179	0.221
Bovine	0.930	0.909	0.911	0.882	**0.335**	0.099	0.029	0.177	0.247	0.185	0.214
L. europaeus	0.940	0.958	0.994	0.883	0.906	**0.411**	0.078	0.177	0.218	0.186	0.280
Goat	0.935	0.932	0.929	0.889	0.971	0.925	**0.332**	0.166	0.197	0.198	0.211
Chicken	0.781	0.841	0.830	0.850	0.838	0.838	0.847	**0.376**	0.069	0.271	0.257
Duck	0.762	0.798	0.792	0.860	0.781	0.804	0.821	0.933	**0.513**	0.362	0.288
Bull frog	0.843	0.822	0.841	0.836	0.831	0.830	0.820	0.762	0.696	**0.160**	0.146
X. laevis	0.79	0.754	0.747	0.802	0.807	0.756	0.810	0.773	0.750	0.864	**0.155**

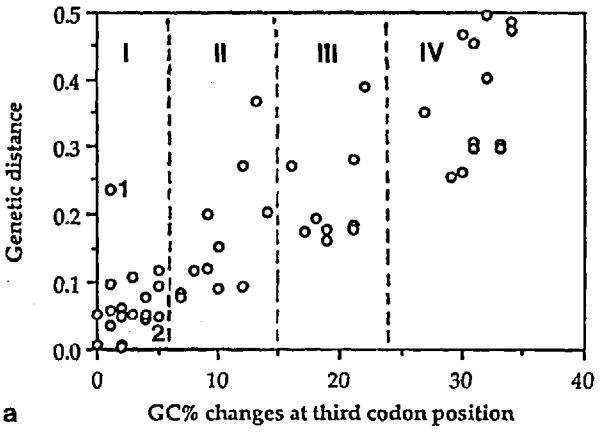
Table 2c. Estimation of D, I, and B for insulin among vertebrate species

	Human	Monkey	Rat	Guinea pig	Dog	Carp
Human	**0.454**	0.041	0.101	0.211	0.091	0.365
Monkey	0.960	**0.592**	0.149	0.276	0.054	0.392
Rat	0.904	0.862	**0.366**	0.229	0.174	0.253
Guinea pig	0.810	0.759	0.795	**0.312**	0.304	0.285
Dog	0.913	0.947	0.841	0.738	**0.615**	0.397
Carp	0.694	0.676	0.777	0.752	0.673	**0.176**

Note: in the table above, the numbers above the diagonal are genetic distances Ds, below the diagonal are similarity coefficients Is and the bold numbers on the diagonal are the standardized synonymous codon bias Bs

codons are chosen for a particular amino acid with a given number of occurrences in the sequence. However, the usual sequence divergence is referred to the number of different sites in the sequences aligned. Therefore, the two classes of clocks and distance values should not substitute for each other because they describe different things.

The reasons for codon bias have been extensively explored. It has been found that codon usage patterns are different between unicellular organisms and multicellular organisms. In the case of vertebrates, codon usage depends on base composition (Bernardi and Bernardi 1985, 1986). Thus, evolutionarily, we can ask what kinds of forces are responsible for the codon bias via the base composition. In the most recent studies, there are different explanations about the base composition. Sueoka (1988) insisted that directional mutation pressure is a major reason. This, per se, is a neutralism explanation because it rejected the role of natural selection. Similarly, in their paper, Wolfe et al. (1989) thought that the base composition in mammals resulted from variation in the



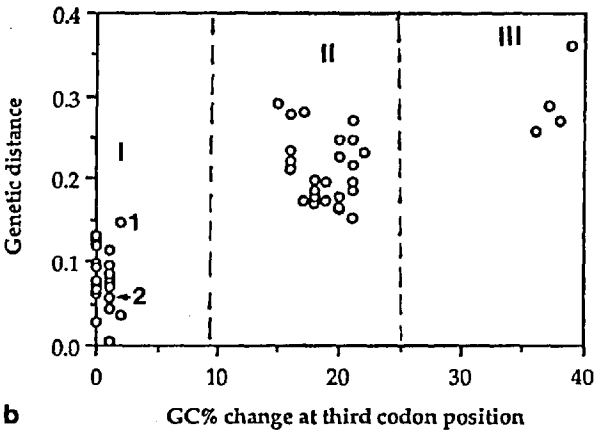


Fig. 3 The relation of genetic distance with GC% change at third codon position. **a** Alpha-globin.The points in region I represent mammals/mammals except point 1 is salamander/*Xenopus* and point 2 is chicken/duck; in region II are chick and duck/mammals; in region III are mouse/other vertebrates; in region IV are amphibians (salamander and *Xenopus*)/other vertebrates. **b** Beta-globin. The points in region I are mammals/mammals except that point 1 is toad (*Xenopus*)/frog (*Rana*) and point 2 is duck/chicken; in region II are mammals/amphibians (frog and toad) and birds (chicken and duck); in region III are chicken and duck/frog and toad.

process of mutation, rather than from natural selection. However, the observation of Bernardi and Bernardi (1986) about G+C content change of the fish *Tilapia* in the habitats with different temperatures (see their Fig. 7) revealed an obvious fact for the role of natural selection on base composition. On the other hand, for multicellular organisms such as *Drosophila melanogaster*, it has been shown that the base compositions in different genes are affected by varying selective constraints (Shields et al. 1988). From the population genetics, it can be shown that very small selection coefficients can lead to a highly biased codon usage (Kimura 1981; Shields et al. 1988). Therefore it is very hard to completely reject the role of natural selection when we try to account for the high codon bias as measured in this paper.

It is interesting to observe codon usage evolution under genome compositional constraint, especially in beta-globin. How this happens may be found in Bernardi et al. (1988). However, in the case of alpha-globin (Fig. 3a), even though we also see codon usage evolution under similar constraint, it seems that the distribution of G+C content changes is continuous. Therefore, we cannot help proposing a gradually evolving isochore in addition to the general compositional patterns proposed by Bernardi and his colleagues. Correspondingly, the gradual evolution of an isochore might add a parameter to codon usage evolution.

In their paper, Maruyama et al. (1986) tabulated pooled data from many genes for each species when they noticed that among taxonomically related organisms the codon choice patterns resemble each other. However, no comparison of homologous genes made them conclude that

> ... it is remarkable to note that the synonymous codon-choice patterns among the vertebrate, or at least among the mammals, are very similar, but clearly different from the pattern of a taxonomically distant organism such as yeast (*S. cerevisiae*) or of *E. coli*. It has been pointed out that the codon-choice pattern, known to be roughly common among the mammals, does not depend on the choice of genes .... (Ikemura 1985)

Thus, the pooled data lose considerable information, which diversifies mammal or vertebrate species. That is because the codon choice patterns of different individual genes are different (Maruyama et al. 1986 had noticed this difference). We clustered part of their pooled data (13 vertebrate species available with yeast and *Escherichia coli*). The clustering graph (not shown) is not good enough to reflect the phylogeny of concerned species except the taxonomically quite distant species yeast and *E. coli*. Thus, we could not reach a satisfactory conclusion until we had enough homologous sequence data. In our analysis, not only did we find considerable difference in synonymous codon choice within mammalian species of vertebrates, but we also found a close correlation between the codon choice and phylogenetic history within mammals.

It is essential to describe quantitatively the codon usage divergence in order to understand the relationship between codon usage divergence and taxonomic relationships among species. Correspondence analysis was applied for this purpose by Grantham et al. (1980, 1981) and brought about some significant results. The basic idea in this method is to define codon usage distance as the sum of the 61 codon frequency changes among species. Then, position the sequence (mRNA or DNA) under study by projection from the multidimensional space defined by the distance to a plane for simple visualization. When homology between the compared sequences is not small, this analysis may pro-

vide a reliable description of codon usage. But when the homology decreases, this method may suffer from a problem that the difference from the amino acid choice (not codon choice) would be incorporated into the distance. The second rapid but crude method is simply to use G+C content at the third position of a codon to describe the codon usage. When homology between the compared sequences is large, this method also can hold because the amino acid composition variation may be supposed to have a minor effect on the G+C content (Mouchiroud and Gautier, 1988). Obviously, if the homology is not considerable, this method will suffer from the same problem as the correspondence analysis does. Meanwhile, from Fig. 3, we know that there is a wide range of codon usage divergence for a given G+C percentage change or even for no G+C content change. Therefore, it appears that direct description with statistical tools such as the genetic distance -employed in present research is more powerful.

Acknowledgments. We thank Drs. Timothy Prout and Michael Turelli for valuable discussion, Dr. C. Gautier for his careful review and helpful comments, and one anonymous reviewer for his/her correcting some embarrassing mistakes made in our preliminary manuscript.

References

Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22:363–365

Bernardi G, Bernardi G (1986) Compositional constraints and genome. J Mol Evol 24:1–11

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FL, Rindone WP, Swindel CD, Tung CS (1986) The GenBank genetic sequence data bank. Nucleic Acids Res 14:1–4

Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398

Filipski J (1988) Why the rate of silent codon substitutions is variable within a vertebrate's genome. J Theor Biol 134:159–164

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284

Gillespie JH (1986) Rates of molecular evolution. Annu Rev Ecol Syst 17:637–665

Gillespie JH (1989) Lineage effects and the index of dispersion of molecular evolution. Mol Biol Evol 6:636–647

Goodman M, Romero-Herrera AE, Dene H, Czelusniak J, Tashian RE (1982) Amino acid sequence on the phylogeny of primates and other eutherians. In: Goodman M (ed) Macromolecualr sequences in systematic and evolutionary biology. Plenum, New York

Goodman M, Czelusniak J, Beeber JE (1985) Phylogeny of primates and other eutherian orders: a cladistic analysis using amino acid and nucleotide sequence data. Cladistics 1:171–185

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74

Hampe A, Therwarth A, Soriano P, Galibert F (1981) Nucleotide sequence analysis of a cloned duck β-globin cDNA. Gene 14:11–21

Hickman CP Sr, Hickman CP Jr, Hickman FM (1979) Integrated principles of zoology, ed 6. C.V. Mosby, St. Louis

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Kimura M (1981) Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. Proc Natl Acad Sci USA 78:5773–5777

Kimura M (1987) Molecular evolutionary clock and the neutral theory. J Mol Evol 26:24–33

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86–93

Lewontin RC (1974) The genetic basis of evolutionary change. Columbia University Press, New York

Li W-H, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330–342

Lingrel JB, Schon EA, Cleary ML, Shapiro SG (1983) Structure and evolution of the developmentally regulated globin genes of the goat. In: E Goldwasser (ed), Regulation of hemoglobin biosynthesis. Elsevier, New York

Maruyama T, Gojobori T, Aota S-I, Ikemura T (1986) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res 14:r151–r197

McLachlan AD, Staden R, Boswell DR (1984) A method for measuring the non-random bias of a codon usage table. Nucleic Acids Research. 12:9567–9575

Mouchiroud D, Gautier C (1988) High codon-usage changes in mammalian genes. Mol Biol Evol 5:192–194

Mueller LD, Ayala FG (1982) Estimation and interpretation of genetic distance in empirical studies. Genet Res Camb 40: 127–137

Nei M (1972) Genetic distance between population. Am Nat 106:283–292

Nei M (1975) Molecular population genetics and evolution. North-Holland, Amsterdam

Nei M, Roychoudhury AK (1974) Sampling variances of heterozygosity and genetic distance. Genetics 76:379–390

Romer AS (1966) Vertebrate paleontology, ed 3. University of Chicago Press, Chicago IL

Romero-Herrera AE, Lieska N, Friday AE, Joysey KA (1982) The primary structure of carp myoglobin in the context of molecular evolution. Phil Trans R Soc Lond B 297:1–25

Saccone C, Pesole G, Preparata G (1989) DNA microenvironments and the molecular clock. J Mol Evol 29:407–411

Sharp PM, Li W-H (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Sokal RR, Rohlf FJ (1981) Biometry. W.H. Freeman, San Francisco

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Van Valen L (1974) Molecular evolution as predicted by natural selection. J Mol Evol 3:89–101

Watson JD (1976) The molecular biology of the gene, ed 3. Benjamin/Cummings, Menlo Park CA

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

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Received February 5, 1990/Revised May 31, 1990