Journal of
Molecular Evolution
© Springer-Verlag New York Inc. 1987

Accumulation Pattern of Amino Acid Substitutions in Protein Evolution

Takashi Kunisawa, Katsuhisa Horimoto, and Jinya Otsuka

Department of Applied Biological Science, Faculty of Science and Technology, Science University of Tokyo, Noda (278), Japan

Summary. A simple method for the evolutionary analysis of amino acid sequence data is presented and used to examine whether the number of variable sites (NVS) of a protein is constant during its evolution. The NVSs for hemoglobin and for mitochondrial cytochrome c are each found to be almost constant, and the ratio between the NVSs is close to the ratio between the unit evolutionary periods. This indicates that the substitution rate per variable site is almost uniform for these proteins, as the neutral theory claims. An advantage of the present analysis is that it can be done without knowledge of paleontological divergence times and can be extended to bacterial proteins such as bacterial c-type cytochromes. It is suggested that the NVS of cytochrome c has been almost constant even over the long period (ca. 3.0 billion years) of bacterial evolution but that at least two different substitution rates are necessary to describe the accumulated changes in the sequence. This "two clock" interpretation is consistent with fossil evidence for the appearance times of photosynthetic bacteria and eukaryotes.

Key words: Molecular evolution — Hemoglobin — Cytochrome c — Number of variable sites — Clock hypothesis

Introduction

It has been suggested based on evolutionary comparisons between amino acid sequences derived from various species that the rate of amino acid substitution in a protein is constant over geologic time and that this rate differs greatly among proteins with different biological functions (Zuckerkandl and Pauling 1965; Dickerson 1971; Kimura 1982). The constancy of the substitution rate can be explained by assuming constant production of selectively neutral substitutions, since for neutral mutants the rate of fixation in the population is equal to the production rate (Kimura 1968, 1983). The difference in substitution rate between various proteins has been interpreted in terms of the ratio of neutral substitutions to the total ones, which depends strongly on the functional constraints on each protein (Kimura 1969; King and Jukes 1969).

In the present paper, we investigate the number of variable sites (NVS) of a protein molecule of the ancestor of living organisms on the basis of their amino acid sequences. By "variable sites," we mean portions of the molecule that allow amino acid substitutions. The other, "invariable" sites represent portions of the molecule that are closely related to its biological function and accept no substitutions. The NVS may be small for proteins interacting with large macromolecules in a well-regulated way, such as cytochrome c, but large for the rest, which interact with small molecules or have less-well-defined biological functions, such as fibrinopeptides. Thus, the rate of substitution will be higher in a protein with a larger NVS.

Some aspects of the variation of the NVS along the phylogenetic tree have been suggested. For example, the NVS may be decreased by an adaptive substitution, while it may be unchanged by a neutral substitution (Noguchi 1978). In contrast, the NVS may be expected to be increased by the loss of some functional constraints (Kimura 1983).

Offprint requests to: T. Kunisawa

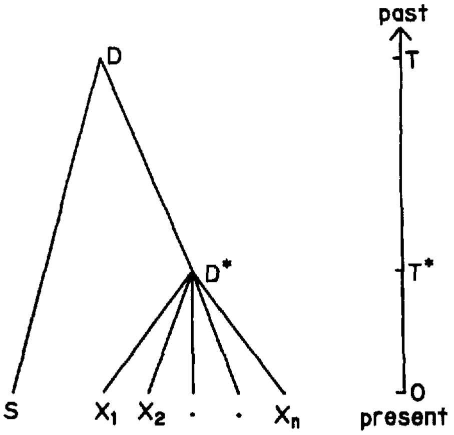

Fig. 1. The phylogenetic tree considered. At D* (T* years ago) n species, denoted by X_i (i = 1 to n), radiated, and at D (T years ago) the other species, denoted by S, diverged

A simple method is proposed here for examining whether the NVS of a protein is constant along its phylogenetic tree. An advantage of the method is that it requires no information on paleontological divergence times, which are often subject to some uncertainty. This analysis is applied to hemoglobin, mitochondrial cytochrome c, and bacterial c-type cytochrome c, whose amino acid sequences are now extensively available.

Method of Evaluating the NVS

Suppose two sequences (A and B) diverged from a common ancestral sequence (X). The number (n_{AB}) of amino acid differences between the two descendant sequences is in general not equal to the sum of the differences between X and A (n_{XA}) and between X and B (n_{XB}), because of accidental coincidence of substitutions at the same sites in the descendants. The form of the equation necessary for the correction of this coincidence is given by

$$n_{AB} = n_{XA} + n_{XB} - \gamma n_{XA} n_{XB} \qquad (1)$$

(for derivation, see Holmquist 1972a, b). Here, γ is expressed by

$$\gamma = (2 - \epsilon)/L \qquad (2)$$

where L is the NVS of the ancestral protein under consideration and ϵ is the probability of finding different amino acids at a variable site in the pairwise comparison between the descendant sequences under the condition that the amino acid of the ancestor is replaced in both of the descendants. Equation (1) is general in the sense that it holds without one's specifying the substitution process as long as all the variable sites are assumed to be equivalent and independent with respect to amino acid substitutions. The value of ϵ may depend on the structure of the genetic code and also on the amino acid composition of the protein. If all substitutions between the 20 amino acids take place with equal probability, ϵ is calculated to be $^{18}/_{19}$, which is close to unity. The values of n_{XA} and n_{XB} are resultant differences between the ancestor and its descendants after multiple substitutions at each of the variable sites. Below, we present a simple method for estimating the value of γ without knowing n_{XA} and n_{XB}.

We first consider the phylogenetic tree shown in Fig. 1, where n species ($X_1, X_2, \ldots, X_n$) radiated at the point D* after species S diverged. Using Eq. (1), the number of amino acid differences between the contemporary sequences S and X_i is written as

$$d_{SX_i} = z + (1 - \gamma z)x_i \qquad (3)$$

where x_i is the number of differences between D* and X_i, and z represents the number of differences between S and D*. If necessary, z can be expressed as

$$z = a + b - \gamma ab$$

where the differences between S and D and between D and D* are given by a and b, respectively. Here we have assumed that γ is constant.

We then consider an average of d_{SX_i} over the radiating species $\{X_i\}$ and obtain

$$\mu = z + (1 - \gamma z)\langle x_i \rangle \qquad (4)$$

and the variance of d_{SX_i} is written as

$$\sigma^2 = (1 - \gamma z)^2 \langle \Delta x_i^2 \rangle \qquad (5)$$

Here the brackets show the sample average over the radiating species. From Eqs. (4) and (5) we obtain

$$(\sigma^2)^{1/2} = \frac{\langle \Delta x_i^2 \rangle^{1/2}}{1/\gamma - \langle x_i \rangle}(1/\gamma - \mu) \qquad (6)$$

Equation (6) tells us that as long as the value of γ is constant, a plot of the average difference versus the square root of the variance should fall into a straight line for various species {S} with different divergence times. This plot can be readily drawn if the average and variance are known by comparison among contemporary amino acid sequences. The intercept of $\sigma^2 = 0$ gives the value for $1/\gamma$, from which the upper and lower limits of the NVS are estimated using Eq. (2), since $0 \leq \epsilon \leq 1$.

The phylogenetic tree given in Fig. 1 is applicable to the topology of divergence of extant species. From fossil records and morphological studies, it is believed that various orders of placental mammals radiated at nearly the same time (e.g., Gingerich 1977). Thus, the placental sequences can be regarded as the reference sequences $\{X_i\}$ and other organisms' sequences correspond to the sequences of interest {S}. Note that Eq. (6) is also valid if placentals diverging after the radiation (e.g., human and gorilla) are included in the reference set. This is because the present average and variance are not ensemble averages but merely sample averages with respect to animals diverging from a common ancestor D*, and their common evolutionary history in some period of time does not alter Eqs. (4)–(6) provided their values $\{x_i\}$ are defined as amino acid differences between D* and X_i. The strict condition that various orders of placentals radiated at a single time is not necessarily needed in the present analysis. In the case where the orders radiated, for example, at successive times, it is sufficient only to interpret D* as representing their oldest ancestor. The present analysis thus requires no detailed phylogeny of placentals. In the following, we have regarded all the available placental sequences as the reference set.

Analysis of Amino Acid Sequence Data

To obtain the average and variance, we first arranged the amino acid sequences following the alignments given by Dickerson and Geis (1983) for α- and β-hemoglobins and by Schwartz and Dayhoff (1978a) for the cytochrome c superfamily. The average number of differences and the variance were then calculated from the difference matrix obtained in pairwise comparisons of the aligned sequences.

The source of amino acid sequence information is the Protein Sequence Database (National Biomedical Research Foundation, Washington, DC) released in September 1985. A list of the species considered here is given in Table 1.

Hemoglobins and Mitochondrial Cytochromes c

The relationship between the average and the square root of the variance of the number of amino acid differences with respect to placental sequences is shown in Fig. 2 for α- and β-hemoglobins. The data points appear to fall in a straight line with fairly good accuracy. This implies that the NVS and ϵ have been approximately constant during vertebrate evolution. The intercepts of the solid lines drawn using an unweighted least-squares fit in Fig. 2 give about 160 and about 150 for the values of $1/\gamma$ for α- and β-hemoglobins, respectively. These values are slightly larger than the actual numbers of amino acid residues in the sequences. This may reflect a combination of the limited sensitivity of the present method and the limited data available on the sequences. With this reservation in mind, it appears that almost all of the sites are variable. Therefore, the value of ϵ is inferred to be near unity and random substitution between all kinds of amino acids may be useful as a first approximation. For simplicity we assume $\epsilon = 1$ hereafter; the NVS is then equal to $1/\gamma$, keeping in mind also that the NVS thus obtained is a minimum.

The same analysis was carried out on mitochondrial cytochromes c and the result is shown in Fig. 2C. A linear relationship is obtained also for mitochondrial ctyochromes c. The NVS is estimated to have a constant value of about 60.

It is of interest to compare the cytochrome c value of 60 with that of 160 or 150 for hemoglobin. The ratio between the NVSs is close to the ratio of ca. 3.4 between the unit evolutionary periods, which are estimated for vertebrate evolution using paleontological divergence times (e.g., Dickerson 1971). This shows that the substitution rate per variable site is almost uniform for hemoglobin and mitochondrial cytochrome c. We conclude that the NVSs of hemoglobin and mitochondrial cytochrome c are not significantly changed during their evolution and that the large difference between their substitution rates per amino acid site is mainly attributable to the difference between their NVSs, reflecting a difference in functional constraints on these proteins.

Cytochrome c Superfamily

The plot of $(\sigma^2)^{1/2}$ vs μ is shown in Fig. 3 for the members of the cytochrome c superfamily listed in Table 1, which includes bacterial and a few chloroplastic cytochromes c in addition to mitochondrial cytochromes c. In contrast to the situation for hemoglobins, it appears that a linear relationship between the square root of the variance and the average agrees relatively poorly with the data points for this superfamily. Rather, it seems that bacterial (plus chloroplastic) and mitochondrial cytochromes c lie on their own straight lines, each with its own slope and intercept; the intercept (and therefore the NVS) for bacterial ones is about 140, while that for mitochondrial ones is 60. At first glance, this seems to suggest that the NVS is not constant throughout the evolution from prokaryote to eukaryote and that the NVS of mitochondrial cytochromes c switched to a different value from that of bacterial or chloroplastic ones concomitantly with the appearance of eukaryotes. This possibility, however, is improbable for the following reason. Using the intercepts and slopes of the two assumed straight lines, we can obtain the individual values for $\langle x_i \rangle$ and $\langle \Delta x_i^2 \rangle$. The ratio $\langle \Delta x_i^2 \rangle / \langle x_i \rangle$ is thus estimated to be about 0.04. On the other hand, $\langle x_i \rangle$ and $\langle \Delta x_i^2 \rangle$ can also be estimated roughly by solving numerically a set of Eqs. (1). This method gives a ratio of ca. 3, which is in near agreement with the values reported by others (Kimura 1983; Gillespie 1984). Thus, there is a large discrepancy in the ratios $\langle \Delta x_i^2 \rangle / \langle x_i \rangle$ if we adopt a sudden change in the NVS only in the mitochondrial cytochromes c. The alternative explanation is that the NVS is constant but amino acid substitutions take place at different rates depending on the position in the cytochrome c molecule. The period of eukaryotic (mitochondrial) evolution is too short for slower substitutions to be detected, but both slower and faster substitutions appear distinctly in the long period of evolution from prokaryote to eukaryote. We examine this possibility below by taking account of the nonuniform variability of the variable sites.

The variable sites are, for simplicity, assumed to be grouped into fast and slowly varying sites. The average number of amino acid differences between S and $\{X_i\}$ is then given by

$$\mu = z_f + (1 - \gamma_f z_f)\langle x_{if} \rangle + z_s + (1 - \gamma_s z_s)\langle x_{is} \rangle \quad (7)$$

and the variance by

$$\sigma^2 = (1 - \gamma_f z_f)^2 \langle \Delta x_{if}^2 \rangle + (1 - \gamma_s z_s)^2 \langle \Delta x_{is}^2 \rangle \quad (8)$$

where the subscripts f and s indicate the quantities referring to the fast and slowly varying regions, respectively. Here, substitutions in the fast and slowly varying regions are assumed to be independent of one another and therefore both the average and variance become additive. Although Eqs. (7) and (8) are not as simply analyzed as Eqs. (4) and (5), the average and variance are readily estimated if we regard the amino acid substitutions as two Poisson processes with the faster rate being λ and the slower

Table 1. List of species whose amino acid sequences were analyzed

α-Hemoglobin	β-Hemoglobin	Cytochrome c superfamily[a]
1 Human	1 Human	1 Human
2 Gorilla	2 Gibbon	2 Spider monkey
3 Langur	3 Rhesus monkey	3 Horse
4 Rhesus monkey	4 Japanese monkey	4 Pig
5 Savannah monkey	5 Savannah monkey	5 Sheep
6 Spider monkey	6 Gelada baboon	6 Hippopotamus
7 Capuchin	7 Yellow baboon	7 Rabbit
8 Marmoset	8 Macaque	8 Mouse
9 Yellow baboon	9 Langur	9 Rat
10 Gelada baboon	10 Colobus	10 California gray whale
11 Sooty mangabey	11 Spider monkey	11 Dog
12 Gorilla alpha-3	12 Human delta	12 Elephant seal
13 Brown lemur	13 Spider monkey delta	13 Gray kangaroo
14 Slow loris	14 Slow loris	14 Mouse testis-specific
15 Slender loris	15 Slender loris	15 Chicken
16 Grand galago	16 Grand galago	16 King penguin
17 Tarsier	17 Tarsier	17 Emu
18 Tree shrew	18 Tree shrew	18 Ostrich
19 Mouse	19 Brown lemur	19 Snapping turtle
20 Mole rat	20 Ring-tailed lemur	20 Rattlesnake
21 Muskrat	21 Badger	21 Bullfrog
22 Golden hamster	22 Dog	22 Bonito
23 Rat	23 Rabbit	23 Carp (iso-1)
24 Guinea pig	24 Egyptian fruit bat	24 Carp (iso-2)
25 Rabbit	25 European hedgehog	25 Puget Sound dogfish
26 European hedgehog	26 Musk shrew	26 Pacific lamprey
27 Musk shrew	27 European mole	27 Starfish
28 European mole	28 Horse	28 Common brandling worm
29 Dog	29 Zebra	29 Freshwater prawn
30 Coyote	30 Tapir major	30 Garden snail
31 Badger	31 Tapir minor	31 Mediterranean fruit fly
32 Indian elephant	32 White rhinoceros	32 Fruit fly
33 African elephant	33 Pig	33 Screwfly
34 Egyptian fruit bat	34 Llama	34 Cynthia moth
35 Rock hyrax	35 Camel	35 Locust
36 Horse	36 Bovine	36 *Candida krusei*
37 Zebra	37 Gayal	37 Yeast
38 Wild ass 1	38 Elk	38 *Debaryomyces kloeckeri*
39 Wild ass 2	39 Sheep	39 Baker's yeast iso-1
40 Tapir	40 Goat A	40 Baker's yeast iso-2
41 White rhinoceros	41 Goat C	41 *Schizosaccharomyces pombe*
42 Armadillo	42 Barbary sheep C	42 *Humicola lanuginosa*
43 Pig	43 Fetal bovine	43 *Neurospora crassa*
44 Bovine	44 Fetal goat	44 Smut fungus
45 Gayal	45 Indian elephant	45 Rape
46 Goat	46 African elephant	46 Mung bean
47 Goat alpha-2	47 Rock hyrax	47 Hemp
38 Elk	48 Armadillo	48 Sesame
49 Deer	49 Mouse major	49 Rice
50 Llama	50 Mouse minor	50 Maize
51 Camel	51 Mole rat	51 Castor bean
52 Gray kangaroo	52 Rat	52 Cotton
53 Opossum	52 Vole	53 Indian mallow
54 Echidna	54 Muskrat	54 Tomato
55 Echidna alpha-2	55 Golden hamster	55 Elder
56 Platypus	56 Guinea pig	56 Box elder
57 Chicken	57 Human gamma	57 Leek
58 Stressed chicken	58 Macaque gamma	58 Arum
59 Pheasant	59 Rabbit gamma	59 Love-in-a-mist
60 Starling	60 Human epsilon	60 Nasturtium
61 Duck	61 Pig epsilon	61 Wheat
62 Greylag goose	62 Goat epsilon	62 Niger
63 Barhead goose	63 Rabbit epsilon	63 Sunflower
64 Magpie goose	64 Mouse epsilon	64 Parsnip

Table 1. Continued

α-Hemoglobin	β-Hemoglobin	Cytochrome c superfamily[a]
65 Flamingo	65 Goat epsilon	65 Buckwheat
66 Golden eagle	66 Gray kangaroo	66 Spinach
67 White stork	67 Potoroo	67 *Ginkgo biloba*
68 Ostrich	68 Opossum	68 *Enteromorpha intestinalis*
69 Rhea	69 Echidna	69 *Euglena gracilis*
70 Nile crocodile	70 Platypus	70 *Crithidia oncopelti*
71 Alligator	71 Chicken	71 *Crithidia fasciculata*
72 Caiman	72 Pheasant	72 *Tetrahymena pyriformis*
73 Chicken delta	73 Duck	73 *Rhodomicrobium vannielii* (c2)
74 Pheasant delta	74 Northern mallard	74 *R. viridis* (c2)
75 Duck delta	75 Greylag goose	75 *R. acidophila* (c2)
76 Starling delta	76 Barhead goose	76 *Rs. fulvum* (iso-1) (c2)
77 Painted turtle delta	77 Canada goose	77 *Rs. molischianum* (iso-1) (c2)
78 Turtle delta	78 Magpie goose	78 *Rs. molischianum* (iso-2) (c2)
79 Chicken pi	79 Flamingo	79 *Rs. rubrum* (c2)
80 Human zeta	80 Golden eagle	80 *Rs. photometricum* (c2)
81 Pig zeta	81 Ostrich	81 *R. palustris* (strain 2.1.6) (c2)
82 Viper	82 White stork	82 *R. sphaeroides* (c2)
83 Bullfrog tadpole	83 Starling	83 *R. capsulata* (c2)
84 *Xenopus* major	84 Chicken rho	84 *Paracoccus denitrificans* (c550)
85 *Xenopus* minor	85 Chicken epsilon	85 *Rs. tenue* (c2)
86 Newt	86 Nile crocodile	86 *Pseudomonas aeruginosa* (c551)
87 Axolotl	87 Alligator	87 *Pseudomonas fluorescens* (c551)
88 Carp	88 Caiman	88 *Pseudomonas stutzeri* (c551)
89 Desert sucker	89 *Xenopus* major	89 *Pseudomonas mendocina* (c551)
90 Goldfish	90 Frog	90 *Pseudomonas denitrificans* (c551)
91 Trout I	91 Bullfrog	91 *Azotobacter vinelandii* 0 (c551)
92 Lungfish	92 Bullfrog tadpole	92 *R. gelatinosa* (c2)
93 Shark	93 *Xenopus* tadpole	93 *Monochrysis lutheri* (c6)
	94 Lungfish	94 *Porphyra tenera* (c6)
	95 Carp	95 *Alaria esculenta* (c6)
	96 Goldfish	96 *Plectonema boryanum* (c6)
	97 Trout IV	97 *Euglena gracilis* (c6)
	98 Trout I	98 *Spirulina maxima* (c6)
	99 Shark	99 *Prosthecochloris aestuarii* (c555)
		100 *Chlorobium limicola* (c555)
		101 *Pseudomonas mendocina* (c5)

[a] The genera *Rhodospirillum* and *Rhodopseudomonas* are abbreviated *Rs.* and *R.*, respectively

rate being λ/q ($q > 1$). At a slowly varying site, allowable amino acids may be more restricted than at a fast varying site. Such effects on the substitution rate are represented by the factor q. The probabilities of finding different amino acids after a time T at fast and slowly varying sites are thus given, respectively, by

$$P_f(T) = 1 - \exp(-\lambda T) \qquad (9)$$

$$P_s(T) = 1 - \exp(-\lambda T/q) \qquad (10)$$

Here, the probability of reverting to the original amino acid is neglected for simplicity. Using Eqs. (9) and (10), we can estimate the values of z_f, $\langle x_{if}\rangle$, z_s, and $\langle x_{is}\rangle$ for various values of λT and λT^*. The numbers of sites in the fast and slowly varying regions, $1/\gamma_f$ and $1/\gamma_s$, can be read from the intercepts of the two assumed straight lines. The values of $\langle \Delta x_{if}^2\rangle$ and $\langle \Delta x_{is}^2\rangle$ are estimated to fit the slopes.

The relationship between the average and variance thus obtained as a function of λT is shown by a solid curve in Fig. 3. In this way, the feature of lying on two straight lines, found in the cytochrome c superfamily, is not incompatible with a constant NVS, and is well reproduced by using the two substitution processes without any change of the NVS. The value of the NVS ($=1/\gamma_s + 1/\gamma_f$) is thus estimated to be about 140, and very few codon sites are invariable. This is consistent with the result of Holmquist et al. (1983), who estimated only three codons to be invariable over the evolution of the cytochrome c superfamily based on their fitting of mutation distribution over codons to the negative binomial distribution. The large q value of 12 in our simulation may in part be attributable to the slow rate of insertion or deletion events. Such an event is formally regarded as a substitution in the present analysis, since we obtained a similar biphasic feature even when the positions with gaps in the aligned sequences were all neglected.

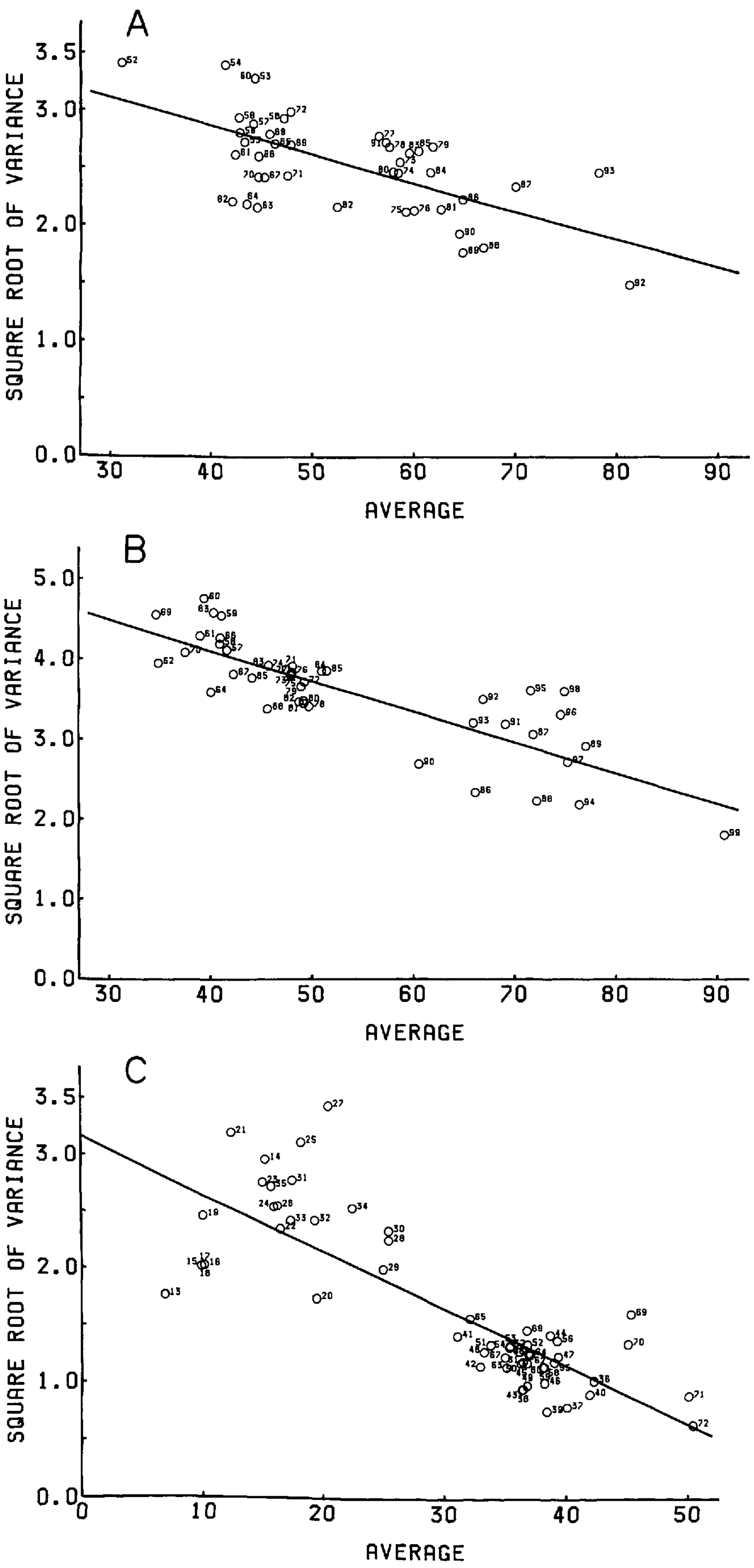

Fig. 2A–C. The linear relationship between the square root of the variance and the average number of amino acid differences with respect to placental sequences: **A** α-hemoglobins; **B** β-hemoglobins; **C** mitochondrial cytochromes c. The numbers attached to the data points are the species numbers given in Table 1. The straight line is the best linear regression fit to the data points. In counting the amino acid differences, positions with gaps or deletions in one aligned sequence when paired against real amino acids in the other aligned sequence were neglected

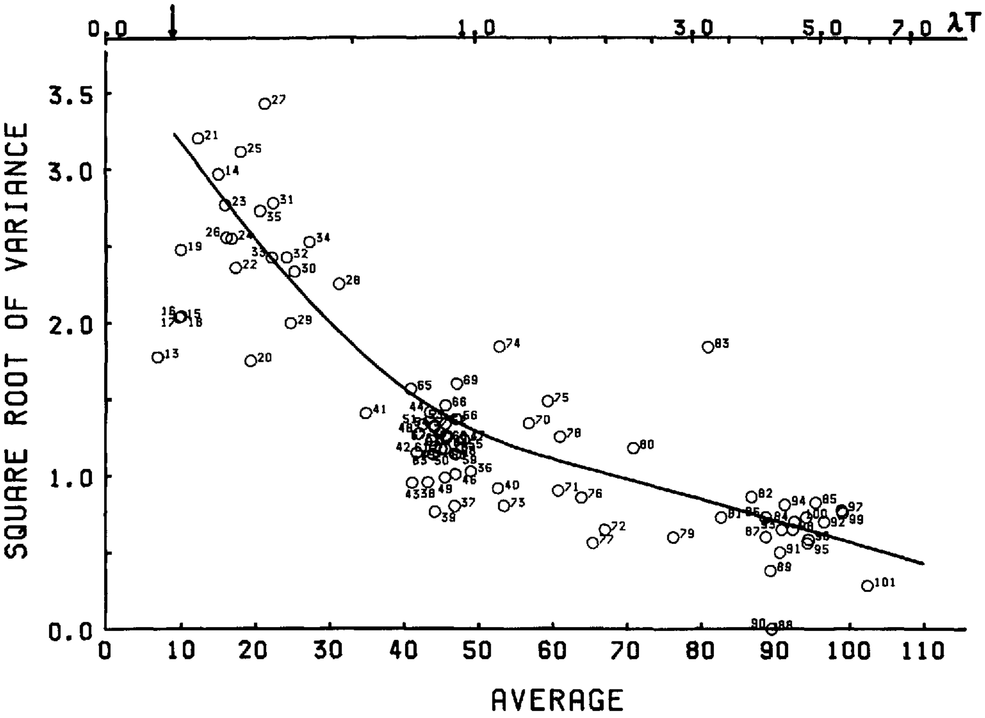

Fig. 3. The relationship between the square root of the variance and the average number of amino acid differences for the cytochrome c superfamily. In counting the amino acid differences, a position with a gap in one sequence when paired against a real amino acid in the other sequence is counted as difference. A total of 156 sites was compared in obtaining the number of amino acid differences. The solid curve was drawn using Eqs. (7)–(10) for various divergence times of the species S in Fig. 1. The number of fastly varying sites is taken to be 40 and that of slowly varying sites is 100. The radiation point, taken at $\lambda T^* = 0.1$, is shown by the arrow. $\langle \Delta x_{if}^2 \rangle = 10.6$, $\langle \Delta x_{is}^2 \rangle = 1.94$, $q = 12$

Discussion

Quantities analogous to the present NVS of a protein have attracted some interest and can be found in the literature. It is important to distinguish between our NVS and T_2, the number of varions in the random evolutionary hit theory developed by Holmquist and his co-workers (Holmquist et al. 1972; Jukes and Holmquist 1972; Holmquist and Pearl 1980). T_2 is a quantity concerning contemporary sequences, as Holmquist and Jukes (1981) have explicitly written, and is obtainable independent of the topology of species divergence. By contrast, our NVS is a quantity that can be defined at any stage of phylogenetic tree and our analysis is aimed at examining whether the NVS is constant during evolutionary development.

Fitch and Markowitz (1970) have reported that the number of invariable codons for a "wide range" of species (which ranges, in practice, from fungi to mammals) is significantly less than that for a "narrow range" of species (which means, in practice, mammals). They interpreted this result as indicating interconversions between variable and invariable codons that result from stereochemical interactions on the protein molecule, and they proposed the concept of concomitantly variable codons (covarions). It should be noted, however, that although in their view a codon that is variable in one period could become invariable in another, this conversion is not taken into account at all in their analysis, which assumes the superposition of Poisson distributions with fixed intensities representing the size of each group of codons. Thus, their method of analysis seems too primitive to reproduce their view of protein evolution. They inferred the minimum number of mutations from the parsimony principle, and the error in such an estimation may be large for a small number of sequences, as in the case of their "narrow range" of species. Furthermore, it would be very difficult to draw a distinction between a truly invariable codon and a merely unvaried codon from the consideration of only sequences that have recently diverged. Their number of variable codons for the "narrow range" of species is probably underestimated, since for cytochrome c in eukaryotes their method gives nearly the same number of variable codons as is obtained by the present method (Fitch 1976). Our result of a constant NVS indicates that conversions between the variable and invariable sites are very rare, even if a substitution at one site could change the number of allowable amino acids at another site. This may be mostly because of the degeneracy in physicochemical properties of the 20 amino acid residues. The residues at the invariable sites are expected to be closely related to protein function because of their specific physicochemical properties and there may therefore be no room for substitution. Under such circumstances, the invariable sites would remain invariable and the variable sites variable.

Along with the uniformity of the substitution rate per variable site, the constant NVS supports the neutral theory. This is not to say that adaptive substitutions have never occurred. It means merely that we cannot detect them by taking into account only the total number of amino acid differences. To resolve the ratio of adaptive substitutions to the total number, it may be necessary to investigate amino acid sequences themselves, since, for instance, the cooperativity in oxygenation of hemoglobin is known to be altered drastically by a single mutation (e.g., Kunisawa and Otsuka 1978). At any rate, the amino acid substitution rate for the entire protein seems to be regulated by the NVS and by the number of allowable amino acids at each variable site.

We have estimated the value of γ on the basis of Eq. (1). The last term in Eq. (1) represents an average of correction to n_{AB} for given n_{XA} and n_{XB}. This correction is calculated on a probabilistic basis, and it has a large variance for small values of n_{XA} and n_{XB}. In accordance with this large variance is the recognizably larger scatter in data points for cytochrome c as the divergence time of the sequence of interest S approaches the radiation point. Accordingly, we cannot completely exclude the possibility of a linear relationship in the cytochrome c superfamily. As far as the available sequence data are concerned, however, that the data lie on two straight lines seems more likely, and the two-clock interpretation proposed in this paper is consistent with fossil evidence for the appearance times of the photosynthetic bacteria and the eukaryotes, as we will discuss below.

The constancy of amino acid substitution rate reported so far relies mainly on proteins derived from vertebrates for the calibration of divergence times, and the validity of this clock for bacterial proteins has been questioned (Schwartz and Dayhoff 1978b). In contrast to the usual plot of the Poisson corrected number of substitutions vs the divergence time, our present plot can be constructed without knowledge of the divergence times, which can in fact be estimated from the plot. An example of such estimates is shown in Fig. 3 for the cytochrome c superfamily. From the time scale given in the upper abscissa of that figure, one can trace the anaerobic photosynthetic bacteria and cyanobacteria, with $\mu = 95$, back to about 3.2 billion years ago, taking the radiation of the placentals at 65 million years ago. Similarly, the appearance of eukaryotes, with $\mu = 65$, goes back 1.3 billion years. These estimates of the divergence times are consistent with fossil evidence suggesting that traces of both anaerobic photosynthetic and oxygen-releasing photosynthetic bacteria are recognizable more than 3.0 billion years ago and that the ancestor of the eukaryotes is 1.4–1.0 billion years old (e.g., Margulis 1981). Note that if we accept the usual "one clock" hypothesis and take the unit evolutionary period to be 20.0 million years, as estimated from vertebrate cytochromes c (e.g., Dickerson 1971), then the appearance of the anaerobic photosynthetic bacteria goes back only 2.0 billion years. Thus the one-clock hypothesis based on the vertebrate sequences underestimates the bacterial divergence time, since the clock runs too fast in the period of bacterial evolution. A "multiclock" theory as an extension of the present two-clock model seems to be needed for the construction of detailed and reliable phylogenetic trees from amino acid sequence data.

Acknowledgment. We would like to express our sincere gratitude to Professor Masao Kotani for stimulating discussions.

References

Dickerson RE (1971) The structure of cytochrome c and the rates of molecular evolution. J Mol Evol 1:26–45

Dickerson RE, Geis I (1983) In: (ed) Hemoglobin: structure, function, evolution, and pathology. Benjamin, Menlo Park, California

Fitch WM (1976) The molecular evolution of cytochrome c in eukaryotes. J Mol Evol 8:13–40

Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579–593

Gillespie JH (1984) The molecular clock may be an episodic clock. Proc Natl Acad Sci USA 81:8009–8013

Gingerich PD (1977) Patterns of evolution in the mammalian fossil record. In: Hallam A (ed) Patterns of evolution, as illustrated by the fossil record. Elsevier, Amsterdam, p 469

Holmquist R (1972a) Theoretical foundations for a quantitative approach to paleogenetics. Part I: DNA. J Mol Evol 1:115–133

Holmquist R (1972b) Theoretical foundations for a quantitative approach to paleogenetics. Part II: Proteins. J Mol Evol 1:134–149

Holmquist R, Jukes TH (1981) The current status of REH theory. J Mol Evol 18:47–59

Holmquist R, Pearl D (1980) Theoretical foundations for a quantitative paleogenetics. Part III: The molecular divergence of nucleic acids and proteins for the case of genetic events of unequal probability. J Mol Evol 16:211–267

Holmquist R, Cantor C, Jukes T (1972) Improved procedures for comparing homologous sequences in molecules of proteins and nucleic acids. J Mol Biol 64:145–161

Holmquist R, Goodman M, Conroy T, Czelusniak J (1983) The spatial distribution of fixed mutations within genes coding for proteins. J Mol Evol 19:437–448

Jukes TH, Holmquist R (1972) Estimation of evolutionary changes in certain homologous polypeptide chains. J Mol Biol 64:163–179

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1969) The rate of molecular evolution considered from the standpoint of population genetics. Proc Natl Acad Sci USA 63:1181–1188

Kimura M (1982) The neutral theory as a basis for understanding the mechanism of evolution and variation at the molecular level. In: Kimura M (ed) Molecular evolution, protein polymorphism and the neutral theory. Springer-Verlag, Berlin Heidelberg New York, p 3

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

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Kunisawa T, Otsuka J (1978) Co-operative ligand binding in a protein composed of subunits. J Theor Biol 74:559–578

Margulis L (1981) Symbiosis in cell evolution. Life and its environment on the early Earth. WH Freeman, San Francisco Oxford

Noguchi T (1978) A hybrid model of molecular evolution and an evolutionary clock. In: Matubara H, Yamanaka T (eds) Evolution of protein molecules. Japan Scientific Society Press, Tokyo, p 61

Schwartz RM, Dayhoff MO (1978a) Cytochromes. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington, DC, p 29

Schwartz RM, Dayhoff MO (1978b) Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science 27:395–403

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, p 97

Received June 25, 1986/Revised September 29, 1986