

## □ CHAPTER FOUR □

# EVOLUTIONARY CHANGE OF AMINO ACID SEQUENCES

Although the basic genetic change occurs in DNA, it is important to know the evolutionary change of proteins, since proteins are molecules essential for building morphological characters and carrying out physiological functions. It should also be noted that the rate of amino acid substitution in proteins is approximately constant, so that information on amino acid substitution is useful for elucidating the evolutionary relationship of organisms. The mathematical formulation of the evolutionary change of amino acid sequence is also simpler than that of nucleotide substitution, since backward and parallel mutations occur with a lower probability for amino acid sites than for nucleotide sites. Technically, nucleotide sequencing is less time-consuming than amino acid sequencing. However, once the coding regions of DNA are sequenced, the corresponding amino acid sequence can easily be inferred.

### Proportion of Different Amino Acids and the Number of Amino Acid Substitutions

The study of the evolutionary change of proteins or polypeptides starts from the comparison of two or more amino acid sequences of a given polypeptide from different organisms. A simple quantity to measure the evolutionary divergence between a pair of amino acid sequences is the proportion ( $p$ ) of different amino acids between the two sequences. This proportion is estimated by

$$\hat{p} = n_d/n, \quad (4.1)$$

where  $n$  is the total number of amino acids compared and  $n_d$  is the number of different amino acids. If all amino acid sites are subject to substitution with an equal probability,  $n_d$  follows the binomial distribution. Therefore, the variance of  $\hat{p}$  is given by

$$V(\hat{p}) = p(1-p)/n. \quad (4.2)$$

When  $p$  is small, it is approximately equal to the number of amino acid substitutions per site. When  $p$  is large, however, it is no longer a good measure of this number, because there might have been two or more amino acid substitutions at sites where amino acids are different between the two sequences. To estimate the number of amino acid substitutions from  $p$ , we need a mathematical model.

### *Poisson Process*

A simple mathematical model that can be used for relating  $p$  to the expected number of amino acid substitutions per site is the Poisson process in probability theory. Let  $\lambda$  be the rate of amino acid substitution per year at a particular amino acid site and assume for simplicity that it is the same for all sites. This assumption does not necessarily hold in reality, but, as will be seen later, the error introduced by this assumption is small unless a very long evolutionary time is considered. The mean number of amino acid substitutions per site during a period of  $t$  years is then  $\lambda t$ , and the probability of occurrence of  $r$  amino acid substitutions at a given site is given by the following Poisson distribution:

$$P_r(t) = e^{-\lambda t} (\lambda t)^r / r!. \quad (4.3)$$

Therefore, the probability that no change has occurred at a given site is  $P_0(t) = e^{-\lambda t}$ . Thus, if the number of amino acids in a polypeptide is  $n$ , the expected number of unchanged amino acids is  $ne^{-\lambda t}$ .

In reality, we generally do not know the amino acid sequence for an ancestral species, so that (4.3) is not applicable. The number of amino acid substitutions is usually computed by comparing homologous polypeptides from two different organisms that diverged  $t$  years ago. Since the probability that no amino acid substitution occurs during  $t$  years is  $e^{-\lambda t}$ , the probability ( $q$ ) that neither of the homologous sites of the two polypeptides undergoes substitution is

$$q = e^{-2\lambda t}. \quad (4.4)$$

This probability can be estimated by  $\hat{q} \equiv 1 - \hat{p} = n_i/n$ , where  $n_i$  is the number of identical amino acids between the two polypeptides. The

equation  $q = e^{-2\lambda t}$  is approximate because backward mutations and parallel mutations (the same mutations occurring at the homologous amino acid sites in two different evolutionary lines) are not taken into account. But the effects of these mutations are generally very small unless a long evolutionary time is considered.

If we use (4.4), the total number of amino acid substitutions per site for the two polypeptides ( $d = 2\lambda t$ ) can be estimated by

$$\hat{d} = -\log_e \hat{q}. \quad (4.5)$$

Therefore, if we know  $t$ ,  $\lambda$  is estimated by  $\hat{\lambda} = \hat{d}/(2t)$ . On the other hand, if we know  $\lambda$ ,  $t$  is estimated by  $\hat{t} = \hat{d}/(2\lambda)$ . The large-sample variance of  $\hat{d}$  is

$$\begin{aligned} V(\hat{d}) &= \left[ \frac{d\hat{d}}{dq} \right]^2 V(q) \\ &= (1-q)/(qn), \end{aligned} \quad (4.6)$$

since  $V(\hat{q}) = q(1-q)/n$  (see Elandt-Johnson 1970). Obviously, the variances of  $\hat{\lambda}$  and  $\hat{t}$  are given by  $V(\hat{d})/(2t)^2$  and  $V(\hat{d})/(2\lambda)^2$ , respectively.

It should be noted that if we knew the numbers of amino acid substitutions for all amino acid sites, the variance of the number of amino acid substitutions per site would have been  $2\lambda t/n$  under the Poisson process. (The variance of a Poisson variable is equal to the mean.) In practice, it is impossible to know these numbers, so we must estimate  $d$  by equation (4.5). Since (4.5) is based on incomplete information on amino acid substitutions, (4.6) gives a variance larger than  $2\lambda t/n$ .

In the above formulation, we assumed that the rate of amino acid substitution is the same for all amino acid sites. This assumption usually does not hold, since the rate is higher at functionally less important sites than at functionally more important sites (see next section). Indeed, Fitch and Margoliash (1967b) and Uzzell and Corbin (1971) have shown that the distribution of the number of amino acid substitutions has a larger variance than the Poisson variance. However, equation (4.5) is quite robust and approximately holds even if the rate varies considerably from site to site. This can be seen by considering the extreme case, where proportion  $a$  of amino acid sites is invariable and proportion  $1-a$  is subject to the Poisson change. The expected proportion of identical amino acids for this case is given by

$$q = a + (1 - a)e^{-2\lambda t}$$

Strictly speaking, therefore,  $d \equiv -\log_e q$  is not linear with evolutionary time. However, as can be seen from figure 4.1,  $d$  is approximately linear when  $2\lambda t \leq 1$ . It should be noted that when  $2\lambda t \ll 1$ ,  $e^{-2\lambda t} \approx 1 - 2\lambda t$ . Therefore,  $d \approx -\log_e[1 - 2(1 - a)\lambda t] \approx 2(1 - a)\lambda t$ . Since  $(1 - a)\lambda$  is the average rate ( $\bar{\lambda}$ ) of amino acid substitution for all amino acid sites,  $d$  can be written as  $2\bar{\lambda}t$ . Namely, when  $\lambda$  varies with amino acid,  $d = 2\bar{\lambda}t$  still holds unless  $2\bar{\lambda}t$  is large.

Of course, if  $a$  is large and  $2\lambda t > 1$ , the relationship between  $d$  and  $q$  is no longer linear. In this case, a curve similar to figure 5.3 may be obtained. Recently, Lee et al. (1985) obtained a curvilinear relationship between evolutionary time and  $d$  for superoxide dimutase for yeast, fruitfly, horse, cow, and man. It is possible that this is caused by variation in the rate of amino acid substitution among different sites.

In the above formulation, we assumed that the two amino acid se-

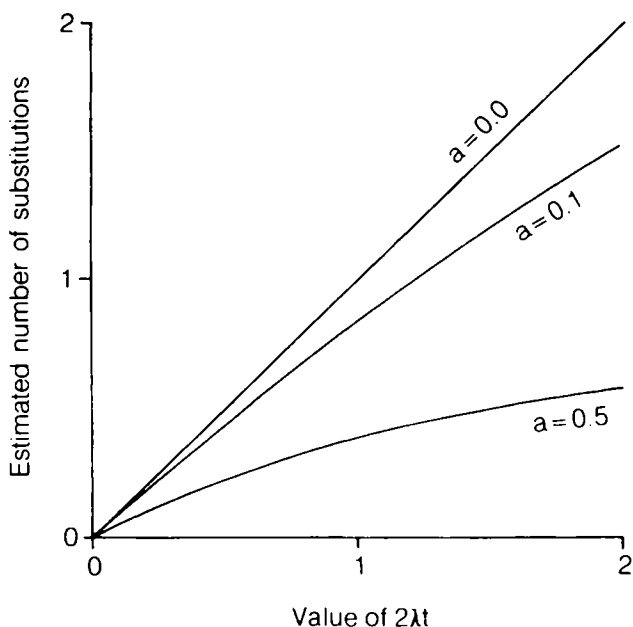


Figure 4.1. Effects of invariable amino acids on the estimate of amino acid substitutions ( $d$ ).  $a$  = proportion of invariable amino acids in a polypeptide.

quences to be compared have the same number of amino acids and that the divergence between them occurred solely by amino acid substitution. When the two sequences are distantly related, however, insertions and deletions are often involved. In this case, we must first identify the locations of insertions and deletions. When the number of insertions or deletions involved is small, as in the case of the following example, this can be done relatively easily. When the number of insertions and deletions is relatively large, however, the alignment of amino acid sequences is quite troublesome. Since this problem is usually more serious with DNA sequences and the problem is nearly identical for the two types of data, I shall discuss this problem in the next chapter.

#### EXAMPLE

Figure 4.2 shows the amino acid sequences of hemoglobin  $\alpha$  chains from the human, horse, bovine, and carp. Here, amino acids are represented by one-letter codes rather than by usual three-letter codes (see table 3.2). The three mammalian hemoglobins consist of 141 amino acids, whereas the carp hemoglobin has 142 amino acids. Comparison of these sequences suggests that deletions or insertions occurred at three different positions after the divergence between fish and mammals. If we ignore these deletions/insertions, the proportion of different amino acids

Human	VLSPADKTNVKAAGWKVGAHAGEYGAEALERMF <del>LS</del> FPTTKTYFPHF-DLSHGSAQVKGHG
Horse	VLSAADKTNVKAAWSKVGGHAGEYGAEALERMF <del>LG</del> FPTTKTYFPHF-DLSHGSAQVKAHG
Bovine	VLSAADKGNVKAAGWKVGGHAAEYGAEALERMF <del>LS</del> FPTTKTYFPHF-DLSHGSAQVKGHG
Carp	SLSDKDKAAVKIAWAKISPKADDIGAEALGRMLTVYPQTKTYFAHWADLSPGSGPVK-HG

Human	KKVA-DALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFT
Horse	KKVA-DGLTLAVGHLLDLP <del>GA</del> LS <del>DL</del> SNLHAHKLRVDPVNFKLLSHCLLVTLAVHL <del>PND</del> FT
Bovine	AKVA-AALTKAVEHL <del>LD</del> LP <del>GA</del> LS <del>EL</del> SDLHAHKLRVDPVNFKLLSHSL <del>LV</del> TLASHL <del>PSD</del> FT
Carp	KKVINGAVGDAVSKIDDLVGG <del>LA</del> SLSELHASKLRVDPANFKILANHIVVGIMFYLP <del>GD</del> FP

Human	PAVHASLDKFLASVSTVLTSKYR
Horse	PAVHASLDKFLSSVSTVLTSKYR
Bovine	PAVHASLDKFLANVSTVLTSKYR
Carp	PEVHMSVDKFFQNLALALSEKYR

Figure 4.2. Amino acid sequences in the  $\alpha$  chains of hemoglobins from four vertebrate species. Amino acids are expressed in terms of one-letter codes. The hyphens indicate the positions of deletions or insertions.

( $\hat{p}$ ) and the estimate of the number of amino acid substitutions for each pair of organisms becomes as given in table 4.1. For example, in the case of the human and carp  $\hat{p} = 68/140 = 0.486$ , and  $\hat{d} = -\log_e(1 - \hat{p}) = 0.666$ . On the other hand, the variance of  $\hat{d}$  becomes  $V(\hat{d}) = 0.486/(0.514 \times 140) = 0.006754$ , the standard error being 0.082.

Table 4.1 indicates that  $\hat{d}$  is nearly the same for the three pairs of mammalian species, whereas the  $\hat{d}$  values between the carp and the mammalian species are considerably larger. This observation is in agreement with the view that the number of amino acid substitutions is roughly proportional to evolutionary time, since the human, horse, and bovine diverged about 75 MY ago, whereas the carp (bony fish) and mammals diverged about 400 MY ago (figure 2.2). The average  $\hat{d}$  for the three pairs of mammalian species is 0.135, whereas the average for the pairs of the carp and the three mammalian species is 0.642, the latter being about five times the former. This ratio is close to the ratio of the corresponding divergence times.

### *Amino Acid Substitution Matrix*

The above Poisson process method for estimating the number of amino acid substitutions gives quite an accurate estimate as long as the evolutionary time considered is relatively short. When the amino acid sequences from distantly related organisms are compared, however, the effects of backward and parallel mutations cannot be neglected, and the Poisson process method is expected to give an underestimate. Unequal

Table 4.1 Numbers of amino acid differences (above the diagonal) between hemoglobin  $\alpha$  chains from the human, horse, bovine, and carp. Deletions and insertions were excluded from the computation, the total number of amino acids used being 140. The figures in parentheses are the proportions of different amino acids. The values given below the diagonal are estimates of the average number of amino acid substitutions per site between two species ( $\hat{d}$ ).

	<i>Human</i>	<i>Horse</i>	<i>Bovine</i>	<i>Carp</i>
Human		18(0.129)	17(0.121)	68(0.486)
Horse	$0.138 \pm 0.032$		18(0.129)	66(0.471)
Bovine	$0.129 \pm 0.031$	$0.138 \pm 0.032$		65(0.464)
Carp	$0.666 \pm 0.082$	$0.637 \pm 0.080$	$0.624 \pm 0.079$	

rates of substitution at different amino acid sites would also contribute to the inaccuracy of the estimate obtained. To take care of these problems, Dayhoff et al. (1978) proposed another method. In this method, the amino acid substitution matrix for a relatively short period of time is considered, and the relationship between the proportion of identical amino acids and the number of amino acid substitutions is derived empirically. The amino acid substitution matrix Dayhoff et al. used was derived from empirical data for many proteins such as hemoglobins, cytochrome c, fibrinopeptides, etc. They first constructed an evolutionary tree for closely related amino acid sequences and then inferred the relative frequencies of substitutions among various amino acids. From these data, they constructed an empirical amino acid substitution matrix ( $\mathbf{M}$ ) for the twenty amino acids (see figure 82 of Dayhoff et al. 1978).

An element ( $m_{ij}$ ) of this substitution matrix gives the probability that the amino acid in column  $i$  will be replaced by the amino acid in row  $j$  during one evolutionary time unit. The time unit used in the matrix is the time during which on the average one amino acid substitution per 100 residues occurs. Dayhoff et al. (1978) measured the number of amino acid substitutions in terms of "PAM," which is an acronym for *accepted point mutations* and represents one amino acid substitution per 100 residues. Therefore, their substitution matrix gives the amino acid substitution probabilities for one PAM.

The amino acid substitution matrix  $\mathbf{M}$  can be used for predicting amino acid changes for any evolutionary time if we know the initial amino acid frequencies. Let  $\mathbf{g}_0$  be a column vector of the relative frequencies of 20 amino acids for a polypeptide at time 0. The amino acid frequencies at time  $t$  or for  $t$  PAMs are then given by

$$\mathbf{g}_t = \mathbf{M}_t \mathbf{g}_0, \quad (4.7)$$

where  $\mathbf{M}_t = \mathbf{M}^t$ . Here, we note that element  $m_{t(ij)}$  of matrix  $\mathbf{M}_t$  gives the probability that the amino acid in column  $i$  at time 0 will be replaced by the amino acid in row  $j$  at time  $t$ . In particular,  $m_{t(ii)}$  represents the probability that the  $i$ th amino acid at time  $t$  is the same as the original one. This latter probability can be used for relating the proportion of different amino acids between homologous polypeptides ( $p$ ) to the number of amino acid substitutions per site ( $d_D$ ). Namely,  $p$  is given by

$$p = 1 - \sum_i g_i m_{2t(ii)}, \quad (4.8)$$

where  $g_i$  is the initial frequency of the  $i$ th amino acid in the polypeptide under investigation. Here, we use  $m_{2t(ii)}$  instead of  $m_{i(ii)}$  because we are considering a pair of polypeptides which diverged  $t$  time units ago. Since  $d_D = 0.01 \times t$  ( $= 0.01$  PAMs) and  $m_{2t(ii)}$  can be obtained from  $M_{2t}$ ,  $p$  can be related to  $d_D$ .

In practice,  $g_0$  may vary from polypeptide to polypeptide or in a given polypeptide from time to time. To avoid this difficulty, Dayhoff et al. used the amino acid frequencies averaged over many different proteins. This does not take into account the specificity of each polypeptide, but it certainly makes the method applicable for many different proteins. Furthermore, if we note that many different proteins have rather similar amino acid frequencies (Dayhoff et al. 1976), this procedure seems to be acceptable for obtaining a rough estimate of the number of amino acid substitutions.

Using the above method, Dayhoff et al. (1978) derived the relationship between  $p$  and  $d_D$ . This relationship is given in table 4.2. This table also includes the value of  $d = -\log_e(1-p)$  in (4.5). It is clear that when  $p$  is smaller than 0.2,  $d_D$  is very close to  $d$ . However, as  $p$  increases, the difference between  $d_D$  and  $d$  gradually increases. Kimura (1983a) produced an approximate formula for  $d_D$ . It is given by

$$d_K = -\log_e(1-p-0.2p^2). \quad (4.9)$$

This formula gives a good approximation as long as  $p \leq 0.7$  (table 4.2). Since  $p$  rarely exceeds 0.7 in actual data, the above formula is very useful.

**Table 4.2** Relationships of the proportion of different amino acids between two homologous polypeptides ( $p$ ) to  $d_D = 0.01$  PAMs,  $d$  in equation (4.5), and  $d_K$  in equation (4.9).

$p$	$d_D$	$d_K$	$d$	$p$	$d_D$	$d_K$	$d$
.01	.01	.01	.01	.45	.67	.67	.60
.05	.05	.05	.05	.50	.80	.80	.69
.10	.11	.11	.11	.55	.94	.94	.80
.15	.17	.17	.16	.60	1.12	1.11	.92
.20	.23	.23	.22	.65	1.33	1.33	1.05
.25	.30	.30	.29	.70	1.59	1.60	1.20
.30	.38	.38	.36	.75	1.95	1.98	1.39
.35	.47	.47	.43	.80	2.46	2.63	1.61
.40	.56	.57	.51				



### Pattern of Amino Acid Substitution

The pattern of amino acid substitution in evolution was first studied by Zuckerkandl and Pauling (1962) and Margoliash (1963) in hemoglobin and cytochrome c, respectively. In these studies, the approximate constancy of the rate of amino acid substitution was apparent. Later, many authors have studied detailed aspects of amino acid substitution. In this section, I will discuss a few important aspects of amino acid substitution.

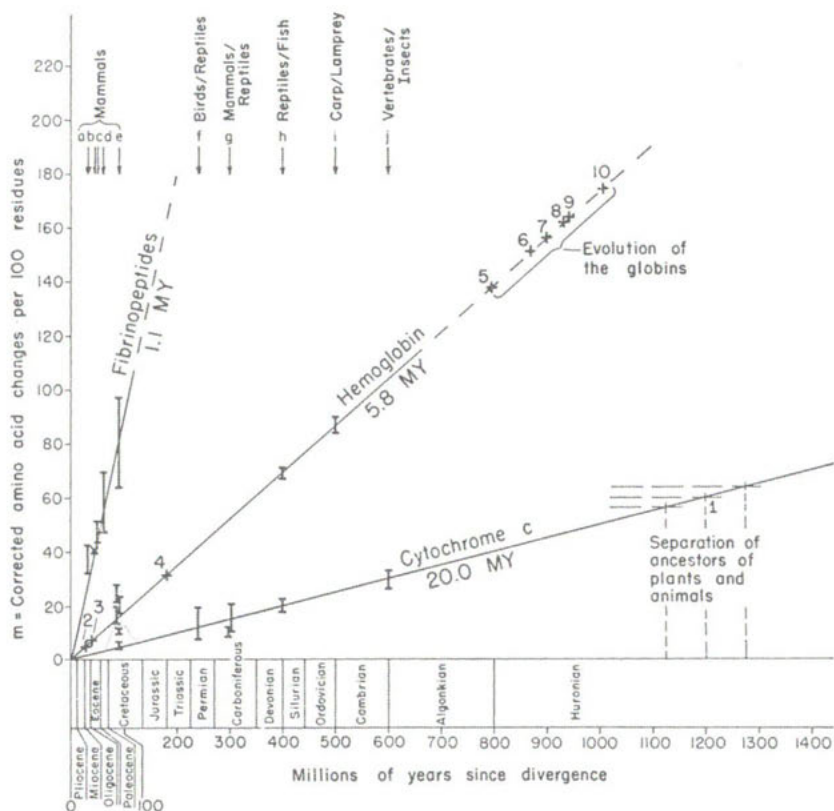
#### *Rates of Amino Acid Substitution*

In the past two decades, the relationship between the number of amino acid substitutions and evolutionary time has been studied for a large number of proteins. In most of these proteins, the number of amino acid substitutions has been shown to increase with increasing evolutionary time. Table 4.3 shows the proportions of amino acid differences between cytochrome c sequences obtained from various groups of organisms. It is clear that the cytochromes c from closely related organisms are more similar than those from distantly related species. The similarity is such that the difference between any two organisms depends almost entirely on the time after divergence. For example, the difference between bacterial cytochrome  $c_2$  (this is homologous to cytochrome c in eukaryotes) and cytochrome c of any other higher organism, plant or animal, is virtually the same (62–72%). Similarly, the cytochrome c in the fungus and yeast group is almost equally related to any other higher organism, the amino acid difference being 41 to 50 percent.

Using equation (4.5), Dickerson (1971) studied the relationship between the number of amino acid differences ( $d$ ) and divergence time ( $t$ ) for cytochrome c, hemoglobins, and fibrinopeptides. The results obtained are presented in figure 4.3. The  $d$  value increases almost linearly with  $t$  in all three groups of polypeptides. A similar linear relationship has been obtained in many other proteins, though the linearity is not always as good as in the case of the three polypeptides in figure 4.3. However, there is enormous variation in the rate of substitution among different polypeptides, as is clear from table 4.4. The highest rate of substitution so far observed is that ( $9 \times 10^{-9}$  per site per year) for fibrinopeptides, and this is 900 times higher than the lowest rate observed for histone H4. The mean and median of the rate are  $1.2 \times 10^{-9}$  and  $0.74 \times 10^{-9}$ , respectively.

Table 4.3 Amino acid differences (%) in cytochrome *c* and *c*<sub>2</sub> between different organisms. The number of positions compared varies with the pair of organisms. All positions are used in computation except those in which both sequences have a gap. Cytochrome *c*<sub>2</sub> in bacteria is known to be homologous to cytochrome *c* in eukaryotes. From Dayhoff (1972).

	Human	Pig	Horse	Chicken	Turtle	Bullfrog	Tuna	Carp	Lamprey	Fruit fly	Screw-worm	Silkworm	Sesame	Sunflower	Wheat	<i>C. krusei</i>	Yeast	<i>N. crassa</i>	<i>R. rubrum</i>
Human	0	10	12	13	14	17	20	17	19	27	25	29	35	38	38	46	41	44	65
Pig, bovine, sheep	10	0	3	9	9	11	16	11	13	22	20	25	38	40	40	45	41	43	64
Horse	12	3	0	11	11	13	18	13	15	22	20	27	39	41	41	46	42	43	64
Chicken, turkey	13	9	11	0	8	11	16	14	17	23	21	26	40	41	41	45	41	44	64
Snapping turtle	14	9	11	8	0	10	17	13	18	22	22	26	38	39	41	47	44	45	64
Bullfrog	17	11	13	11	10	0	14	13	20	20	20	27	41	42	43	46	43	45	65
Tuna fish	20	16	18	16	17	14	0	8	18	23	22	30	42	43	44	43	43	45	65
Carp	17	11	13	14	13	13	8	0	12	21	20	25	40	41	42	45	42	43	64
Lamprey	19	13	15	17	18	20	18	12	0	27	26	30	44	44	46	50	45	47	66
Fruit fly	27	22	22	23	22	20	23	21	27	0	2	14	42	41	42	43	42	38	65
Screw-worm fly	25	20	20	21	22	20	22	20	26	2	0	13	41	40	40	43	42	38	64
Silkworm moth	29	25	27	26	26	27	30	25	30	14	13	0	39	40	40	43	44	44	65
Sesame	35	38	39	40	38	41	42	40	44	42	41	39	0	10	13	47	44	48	65
Sunflower	38	40	41	41	39	42	43	41	44	41	40	40	10	0	13	47	43	49	67
Wheat	38	40	41	41	41	43	44	42	46	42	40	40	13	13	0	45	42	48	66
<i>Candida krusei</i>	46	45	46	45	47	46	43	45	50	43	43	43	47	47	45	0	25	39	72
Baker's yeast	41	41	42	41	44	43	43	42	45	42	42	44	44	43	42	25	0	38	69
<i>Neurospora crassa</i>	44	43	43	44	45	45	45	43	47	38	38	44	48	49	48	39	38	0	69
<i>Rhodospirillum rubrum c</i> <sub>2</sub>	65	64	64	64	64	65	65	64	66	65	64	65	65	67	66	72	69	69	0



**Figure 4.3.** Rates of amino acid substitution in fibrinopeptides, hemoglobin, and cytochrome c. Comparisons for which no adequate time coordinate is available are indicated by numbered crosses. Point 1 represents a date of  $1,200 \pm 75$  MY (million years) for the separation of plants and animals, based on a linear extrapolation of the cytochrome c curve. Points 2–10 refer to events in the evolution of the globin family. The  $\delta/\beta$  separation is at point 3,  $\gamma/\beta$  is at 4, and  $\alpha/\beta$  is at 500 MY (carp/lamprey). From Dickerson (1971).

The rate of amino acid substitution is usually measured by the number of substitutions per amino acid site per year ( $\lambda$ ), as given in table 4.3. Some molecular biologists (e.g., Dickerson 1971), however, have used the *unit evolutionary time* ( $T_u$ ). This is the average time required for one substitution per 100 amino acid sites. It is related to  $\lambda$  by

$$T_u = 1/(100\lambda). \quad (4.10)$$

**Table 4.4** Rates of amino acid substitutions per amino acid site per  $10^9$  years ( $\lambda \times 10^9$ ) in various proteins. Modified from Dayhoff (1978).

<i>Protein</i>	<i>Rate</i>	<i>Protein</i>	<i>Rate</i>
Fibrinopeptides	9.0	Thyrotropin beta chain	0.74
Growth hormone	3.7	Parathyrin	0.73
Ig kappa chain C region	3.7	Parvalbumin	0.70
Kappa casein	3.3	Protease inhibitors, BP1 type	0.62
Ig gamma chain C region	3.1	Trypsin	0.59
Lutropin beta chain	3.0	Melanotropin beta	0.56
Ig lambda chain C region	2.7	Alpha crystallin A chain	0.50
Complement C3a anaphylatoxin	2.7	Endorphin	0.48
Lactalbumin	2.7	Cytochrome b <sub>5</sub>	0.45
Epidermal growth factor	2.6	Insulin (exc. guinea pig and coypu)	0.44
Somatotropin	2.5	Calcitonin	0.43
Pancreatic ribonuclease	2.1	Neurophysin 2	0.36
Lipotropin beta	2.1	Plastocyanin	0.35
Haptoglobin alpha chain	2.0	Lactate dehydrogenase	0.34
Serum albumin	1.9	Adenylate kinase	0.32
Phospholipase A <sub>2</sub>	1.9	Triosephosphate isomerase	0.28
Protease inhibitor, PST1 type	1.8	Vasoactive intestinal peptide	0.26
Prolactin	1.7	Corticotropin	0.25
Pancreatic hormone	1.7	Glyceraldehyde 3-PO <sub>4</sub> dehydrogenase	0.22
Carbonic anhydrase C	1.6	Cytochrome c	0.22
Lutropin alpha chain	1.6	Plant ferredoxin	0.19
Hemoglobin alpha chain	1.2	Collagen (exc. nonrepetitive ends)	0.17
Hemoglobin beta chain	1.2	Troponin C, skeletal muscle	0.15
Lipid-binding protein A-II	1.0	Alpha crystallin B chain	0.15
Gastrin	0.98	Glucagon	0.12
Animal lysozyme	0.98	Glutamate dehydrogenase	0.09
Myoglobin	0.89	Histone H2B	0.09
Amyloid AA	0.87	Histone H2A	0.05
Nerve growth factor	0.85	Histone H3	0.014
Acid proteases	0.84	Ubiquitin	0.010
Myelin basic protein	0.74	Histone H4	0.010

Thus,  $T_{\lambda}$  for fibrinopeptides is  $1.1 \times 10^6$  years, whereas  $T_{\lambda}$  for histone H4 is  $10^9$  years.

### *Differences Among Proteins*

Why is the rate of amino acid substitution so different among different proteins? The answer to this question seems to be that the functional requirement of each protein determines the rate (Zuckermandl and Paul-

ing 1965; King and Jukes 1969; Dickerson 1971). For example, fibrinopeptides do not seem to have any particular function after they are cut out of fibrinogen when the latter is converted to fibrin for blood clotting. Thus, virtually any amino acid can be replaced by any other amino acid. That is, almost all mutations occurring in fibrinopeptides seem to be selectively neutral or nearly neutral. The rate of amino acid substitution is, therefore, expected to be close to the mutation rate per locus (chapter 13). The apparently functionless parts of ribonuclease also show a rate of amino acid substitution similar to that of fibrinopeptides (Barnard et al. 1972).

By contrast, cytochrome c seems to require a rather rigid arrangement of amino acids for the protein to function normally (Dickerson 1971). The polypeptide of this protein forms a shell, inside which the heme group is contained with one edge of the heme being exposed outside. Most of the interior amino acids are hydrophobic and apparently cannot be replaced by hydrophilic amino acids. The heme is attached covalently to the protein through cysteines at positions 14 and 17. The amino acids at the surface of this protein are less restrictive but still must form a certain structure to interact with cytochrome oxidase and reductase, both of which are macromolecules much larger than cytochrome c itself. This functional requirement apparently rejects many amino acid changes in this protein, and only at a limited number of amino acid sites are mutational changes freely accepted.

The functional constraint of hemoglobin is intermediate between fibrinopeptides and cytochrome c. This protein also contains the heme group, and its interior amino acids do not easily accept mutational changes. In hemoglobin  $\alpha$  chain, there are 19 amino acid sites that are involved in the heme pocket. Replacement of amino acids at these sites is known to cause abnormal function of the hemoglobin molecule in man (Perutz and Lehman 1968). The function of hemoglobin is to bind to  $O_2$  in the lung and interact with  $CO_2$  in the tissue, and the surface of the molecule has no essential function except for holding other important amino acids. Thus, the amino acids at the surface can easily be replaced by other amino acids. Kimura and Ohta (1973b) have shown that the rate of amino acid substitution at the surface is about ten times higher than that at the heme pocket.

As mentioned earlier, histone H4 is a highly conserved protein. There are only two amino acid differences in the sequence of 105 amino acids between calf and pea. If we assume that plants and animals diverged one billion years ago, the rate of amino acid substitution is computed to be

$1 \times 10^{-11}$  per site per year. This is about 1/100 of the rate of hemoglobin and about 1/40 of that for cytochrome c. This extremely slow rate of evolution in histone H4 is believed to be due to the important function this protein plays in controlling the expression of genetic information by binding to DNA in the nucleus. An equally slow rate of evolution has been observed for ubiquitin, which also seems to be important in controlling the expression of genetic information (P. M. Sharp and W.-H. Li, personal communication). In addition to these proteins, transfer and ribosomal RNAs are known to evolve very slowly. Since these RNAs play an important role in protein synthesis, many nucleotide substitutions seem to result in deleterious effects.

Some caution, however, should be exercised in the interpretation of the relationship between functional constraint and the rate of amino acid substitution. Because of the above argument, whenever we find a protein with a slow rate of amino acid substitution, we tend to believe that the protein has an important biological function even if its function is not known. Graur (1985) showed that there is a significant negative correlation between the proportion of glycines in proteins and the rate of amino acid substitution, irrespective of the protein function. This correlation is apparently generated by the fact that glycine is smallest among the twenty amino acids and the replacement of this amino acid by another generally impairs the function of the protein.

Nevertheless, a large proportion of the variation in the rate of amino acid substitution can be explained by the differences in functional constraint among proteins. Kimura (1983a) takes this finding as support for his neutral mutation theory. He has hypothesized that almost all amino acid substitutions in fibrinopeptides occur by random fixation of neutral or nearly neutral mutations, and thus the substitution rate per site ( $\lambda$ ) is equal to the mutation rate ( $\mu$ ). In most other proteins, however, a certain amount of purifying selection operates because of functional constraints. Therefore,  $\lambda$  may be expressed as

$$\lambda = (1 - f)\mu, \quad (4.11)$$

where  $f$  is the fraction of deleterious mutations. It would be interesting to test this hypothesis by examining the rates of nucleotide substitution at the three nucleotide positions of codons of the fibrinopeptide gene. If Kimura's hypothesis is correct, we would expect that all three positions show essentially the same rate.

At this point, one might argue that variation in the substitution rate among different proteins is caused by the differences in positive (advantageous) selection rather than negative (purifying) selection. In this argument, however, we must assume that the rate of occurrence of advantageous mutations is higher in rapidly evolving polypeptides such as fibrinopeptides than in slowly evolving polypeptides such as cytochrome c. This assumption does not seem to be reasonable.

Of course, this does not mean that there are no advantageous mutations at the amino acid level. There must be some. Otherwise, adaptive evolution can not occur. However, a relatively small number of advantageous mutations seem to be sufficient for producing an adaptive change of a molecule. For example, crocodilian hemoglobin has lost its old function (the binding of organic phosphate, chloride, and carbamino  $\text{CO}_2$ ) and gained a new function (bicarbonate binding). This functional change represents an adaptive response to the blood acidity that occurs during the prolonged stay of crocodiles under water and can be explained by 5 amino acid substitutions (Perutz et al. 1981). This is a small proportion of the total number of amino acid substitutions (123) between crocodiles and humans, who have not experienced such a change. In general, most amino acid substitutions in hemoglobins do not appear to be related to any significant functional change (Perutz 1983). The functional change of stomach lysozyme of ruminants can also be explained by a small proportion of amino acid changes (Jolles et al. 1984).

In bacteria, there are many examples in which an adaptive change in enzymes (change in substrate specificity) has occurred by one or a few amino acid substitutions (e.g., Clarke 1984). For example,  $\beta$ -lactam antibiotics kill bacteria by inactivating a set of penicillin-binding proteins (PBPs) that are essential for cell division. Some mutants of *E. coli* are resistant to these antibiotics because of the reduction in affinity between antibiotics and PBPs. Hedge and Spratt (1985) have shown that this reduction in affinity is caused by one to four amino acid substitutions in the active center of a PBP and that the majority of other amino acid substitutions do not affect antibiotic susceptibility.

### ***Is the Rate of Amino Acid Substitution Constant in a Given Protein?***

Although the pattern of amino acid substitution shows clocklike behavior, the accuracy of the molecular clock has been controversial. Some

workers discovered cases in which the clock apparently failed. This subject has been reviewed by Wilson et al. (1977a), Goodman et al. (1982), Dickerson and Geis (1983), and Kimura (1983a), and the reader may refer to them for the details. Here, I present only a brief summary of the subject.

It should first be remembered that the clock is empirical and its theoretical basis is not well founded, though the neutral theory can provide the basis once this is firmly established. Second, the molecular clock is stochastic and has a large variance when the number of amino acids examined is small. Third, there are many exceptions even if the stochastic factor is considered. A notable exception is guinea pig insulin, whose rate ( $5.3 \times 10^{-9}$ ) of amino acid substitution is more than ten times higher than that ( $0.33 \times 10^{-9}$ ) of other organisms (King and Jukes 1969). This high rate seems to have been caused by the relaxation of purifying selection, which resulted from a change in the tertiary structure of guinea pig insulin (Kimura and Ohta 1974). Fourth, in the computation of absolute rates of amino acid substitution, it is necessary to have geological dates of divergence times, but these dates are often inaccurate (Wilson et al. 1977a; Joysey 1981). Therefore, great caution is required in testing the constancy of absolute rates of substitution.

A simple method of testing the molecular clock is to estimate the number of amino acid or codon substitutions for each branch of the evolutionary tree for a group of organisms and relate the estimate to the evolutionary time. It is difficult to know the exact number of substitutions for each branch, but the number can be estimated by parsimony methods (e.g., Goodman et al. 1974) or some other technique (chapter 11). Romero-Herrera et al. (1978) used this method to study the rate of codon substitution in myoglobin, using amino acid sequence data from 19 vertebrate species (see also Joysey 1981). From information on fossils and anatomical differences, they first constructed a phylogenetic tree for these organisms. They then estimated the number of codon substitutions ( $n_i$ ) for all branches and related them to the evolutionary times known from the fossil records. They considered a time span of 80 MY with an interval of 10 MY. There were several independent estimates of codon substitutions at each of the 8 evolutionary times considered. Their results indicate that the average number of codon substitutions increases almost linearly with evolutionary time but that the variation around the linear relationship is quite large. Because of this large variation, the number of codon substitutions does not always give the correct estimate



of divergence time when it is used as a clock. For example, at the evolutionary time of 80 MY there were 14 observations of  $n_c$ , but the number varied from 11 to 34, the mean being about 22. Therefore, the largest value of  $n_c$  gives a time estimate which is about three times greater than the smallest. From this observation, Joysey (1981) concluded that myoglobin cannot be used as a molecular clock.

It should be noted, however, that myoglobin is a relatively small polypeptide (about 150 amino acids long) and that stochastic errors alone produce a large variance relative to the mean when the number of substitutions is small. Under the assumption of the Poisson process, the expected variance of  $n_c$  is equal to the mean. Therefore, the expected standard error for a mean of 22 is  $\sqrt{22} = 4.69$ . Thus, the expected range (confidence interval) of  $n_c$  with a probability of 95 percent is 12.62 to 31.38 (mean  $\pm 2$  standard errors). This covers all observations except the two extreme ones (11 and 34). Thus, only two of the 14 observations are significantly different from the expected number. Indeed, similar computations suggest that a large portion of the variance of  $n_c$  observed by Romero-Herrera et al. is due to stochastic errors. It should also be noted that Romero-Herrera et al.'s estimates of  $n_c$  are subject to estimation errors as well as to stochastic errors because they used a parsimony method (see chapter 11).

A statistical study of variation in the rate of amino acid substitution was conducted by Ohta and Kimura (1971a). They examined the rates of amino acid substitution in hemoglobins and cytochrome c in various vertebrate lines and showed that the variance of the rate is 1.9 to 4.2 times higher than that expected under the Poisson process. This suggests that the rate is not strictly constant. Their test, however, was not very accurate because the correlation between different pairs of distances ( $d$ 's) was ignored and some of the divergence times used were unreliable.

Kimura (1983a) devised an improved method for testing rate constancy. He considered the case where all species diverged at the same time. Suppose that  $s$  species are studied, and let  $x_i$  be the number of amino acid substitutions that occurred in the  $i$ th species after separation from the common ancestor. We assume that the number of amino acid substitutions between the  $i$ th and  $j$ th species is  $d_{ij} = x_i + x_j$ . The mean and variance of  $x_i$  are then given by  $\bar{x} = \bar{d}/2$  and  $V_x = (s+1)V_d/[2(s-1)]$ , respectively, where  $\bar{d}$  and  $V_d$  are the mean and variance of  $d_{ij}$  among  $s(s-1)/2$  comparisons. Since  $\bar{d}$  and  $V_d$  are observable quantities, we can estimate  $\bar{x}$  and  $V_x$ .  $V_x$  can then be compared with the expected variance

$\sigma_x^2 = \bar{x}$  under the Poisson process. Application of this method to data on hemoglobin  $\alpha$  and  $\beta$  chains, cytochrome *c*, myoglobin, and pancreatic ribonuclease from different orders of mammals (e.g., human, mouse, rabbit, dog, horse, bovine, etc.) has shown that  $V_x$  is significantly larger than  $\sigma_x^2$  in two polypeptides out of the five examined [see Gillespie (1984) for a somewhat different type of analysis]. It should be noted, however, that even this method is not very accurate because the different species used (mammalian orders in the present case) may not have diverged at the same time.

The problem of inaccurate dating of divergence times can be avoided if we use only three species. Consider the three species *A*, *B*, and *C* in figure 4.4, and assume that *A* and *B* are known to be more closely related to each other than to *C*. We denote by  $d_{AB}$ ,  $d_{AC}$ , and  $d_{BC}$  the numbers of amino acids substitutions between *A* and *B*, *A* and *C*, and *B* and *C*, respectively. We can then write  $d_{AB} = x + y$ ,  $d_{AC} = x + z$ ,  $d_{BC} = y + z$ , where  $x$ ,  $y$ , and  $z$  are the numbers of substitutions for the branches given in figure 4.4. Therefore, if we know  $d_{AB}$ ,  $d_{AC}$ , and  $d_{BC}$ , we can estimate  $x$ ,  $y$ , and  $z$ . That is,

$$x = (d_{AB} + d_{AC} - d_{BC})/2 \quad (4.12a)$$

$$y = (d_{AB} - d_{AC} + d_{BC})/2, \quad (4.12b)$$

$$z = (-d_{AB} + d_{AC} + d_{BC})/2. \quad (4.12c)$$

From figure 4.4, it is clear that if the rate of substitution is constant, the expectations of  $x$  and  $y$  are equal to each other. This method has been used by Sarich and Wilson (1973) and their colleagues to test whether or not the numbers of amino acid substitutions for the human and ape

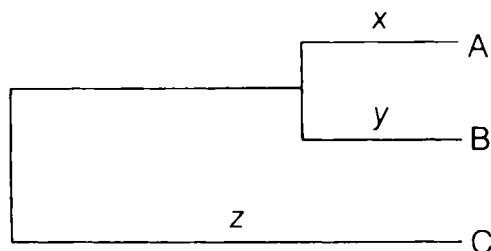


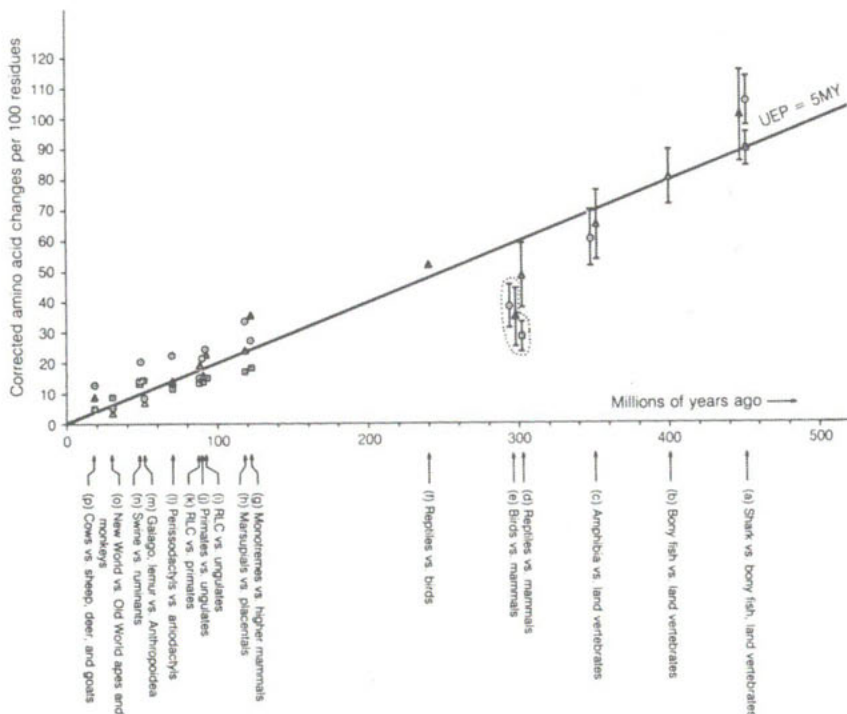
Figure 4.4. Phylogenetic tree for three species.  $x$ ,  $y$ , and  $z$  are the numbers of amino acid substitutions.

lineage and the Old World monkey lineage (e.g., baboon, macaque) are the same. Using New World monkeys and other placental mammals as the external reference species (C), these authors examined data on hemoglobins  $\alpha$  and  $\beta$ , fibrinopeptides A and B, carbonic anhydrase, cytochrome c, myoglobin, and lysozyme c (see Wilson et al. 1977a). However, they could not find a significant difference between the two groups of organisms. A similar test was also conducted for immunological distance data, which will be discussed in the next section, but the result obtained was the same.

A more complex statistical test of the molecular clock was conducted by Langley and Fitch (1974). In this test, no knowledge of evolutionary times is necessary, but we must know the branching pattern (topology) of the phylogenetic tree of the organisms used and the number of amino acid substitutions for each branch. For a given topology and a given set of observed numbers of amino acid substitutions for all branches, the expected branch lengths (expected numbers of amino acid substitutions) are estimated by using the maximum likelihood method, and the heterogeneity of the rate among different branches is tested with a  $\chi^2$  test (chapter 11). Langley and Fitch applied this method to data for hemoglobins  $\alpha$  and  $\beta$ , cytochrome c, and fibrinopeptide A from 18 vertebrate species, showing that the rate heterogeneity is statistically significant.

Uzzell and Corbin (1971) studied the distribution of the number of amino acid substitutions in cytochrome c for a group of vertebrate species and showed that observed data fit the negative binominal distribution better than the Poisson distribution, the variance of the number of substitutions being about two times greater than the Poisson variance. This observation is in agreement with Langley and Fitch's results but suggests that the rate of substitution ( $\lambda$ ) varies with amino acid site, following the gamma distribution [equation (9.57)]. Furthermore, in a statistical analysis of amino acid substitution data, Gillespie (1984, 1986) has suggested that the substitution rate varies randomly from time to time in each evolutionary lineage. If this is the case, the molecular clock may become episodic, with a cluster of substitutions being separated by periods with a few substitutions.

When extensive data on amino acid sequences are available and the times of divergence for the species are known, one can directly examine the relationship between evolutionary time and the number of amino acid substitutions. Figure 4.5 gives such a relationship for hemoglobins and myoglobins. The number of amino acid substitutions increases al-



**Figure 4.5.** Evolutionary changes of hemoglobin and myoglobin chains. The number of amino acid substitutions ( $d$ ) is plotted against the divergence time obtained from paleontological records. Vertical error bars for the older divergence points extend over  $\pm 2$  standard deviations or to the 95 percent confidence level. Triangles represent hemoglobin  $\alpha$ , circles are hemoglobin  $\beta$ , and squares are myoglobin. The straight line is the best linear regression fit to all data points. From Dickerson and Geis (1983).

most linearly if we exclude the comparisons between birds and mammals. Particularly in the mammalian group, the linear relationship holds very well, if we ignore the apparent random errors caused by stochastic factors. However, careful examination suggests that the rate of amino acid substitution for the period from 300 MY ago to 450 MY ago is higher than that for mammalian species even if we exclude the bird-mammal comparison. That is, the rate of substitution apparently deviates from constancy to some extent.

It should be noted that not only the three globin chains but also cytochrome *c* show a smaller number of amino acid substitutions in the bird-mammalian comparison than expected from the linear relationship (Dickerson and Geis 1983). At the present time, the mammalian line is

believed to have split from the reptile-bird line before these two groups of organisms diverged (see figure 2.2). However, fossil records of birds are notoriously poor, and the possibility that birds and mammals diverged after some group of reptiles (e.g., snakes) diverged cannot be excluded (Dickerson and Geis 1983). If this is the case, the anomaly concerning the amino acid substitution might disappear.

Studying the evolution of hemoglobins and myoglobins in vertebrates, Goodman et al. (1975) and Czelusniak et al. (1982) have concluded that the rate of amino acid substitution increased markedly following the gene duplication events separating hemoglobin and myoglobin and  $\alpha$  and  $\beta$  hemoglobins and that this increase was due to advantageous mutations, improving the function of hemoglobin and myoglobin. Wilson et al. (1977a) and Kimura (1981a) have challenged this view and contended that the increased substitution rate observed by Goodman and his associates is probably caused by inaccurate dating of gene duplication events. An increase in the rate of amino acid (nucleotide) substitution was also observed by Li and Gojobori (1983) in duplicate globin genes of the goat and the sheep. However, these authors concluded that it is largely due to the relaxation of purifying selection (functional constraints) in duplicate genes rather than to advantageous mutations. Nevertheless, it is likely that when a new gene is formed after gene duplication its function is improved by nucleotide substitutions within the gene (Dickerson 1971).

It is now clear that the rate of amino acid substitution is not strictly constant but varies substantially with organism. Yet, if we consider a long evolutionary time, it is approximately constant. As was noted by Fitch (1976), the molecular clock does not run as regularly as the regular clock or the isotope clock. It is quite "sloppy" but useful for obtaining a rough idea of evolutionary time when fossil records are absent or unreliable. Note also that for constructing a phylogenetic tree for a group of organisms the rate of amino acid substitution need not be constant. As long as parallel or backward mutations are rare, one can reconstruct a tree fairly easily (chapter 11).

### Immunological Distance and Evolutionary Time

Although amino acid sequence data are useful for estimating evolutionary time and clarifying the genetic relationship of organisms, acquisition of sequence data is time-consuming. As mentioned earlier, nucleotide sequencing is much easier than amino acid sequencing but, even

so, takes too much time. For taxonomic purposes, we need a simpler method. One such method is to use the intensity of immunological reaction between antigens and antisera prepared from different species. There are several methods for measuring the intensity of immunological reaction, but the simplest and most useful method seems to be that of quantitative microcomplement fixation of a purified protein, introduced by Sarich and Wilson (1966). The protein often used for this purpose is serum albumin (Champion et al. 1974). Briefly, the method is as follows. The antisera to be used are produced by immunization of rabbits with purified serum albumin from an organism of the group to be tested. The antisera produced strongly react with the albumin from this organism (homologous antigen) but less strongly with that from another species (heterologous antigen) for a given concentration of antisera. If the serum concentration is raised, however, the reaction with heterologous antigen increases to the level for homologous antigen. The degree of antigenic difference between two albumins is measured by the factor by which the antiserum concentration must be raised in order for a heterologous albumin to produce the same reaction as that with a homologous albumin. This factor is called the index of dissimilarity (*I.D.*). The antigen-antibody reaction is measured by a method called quantitative complement fixation. Sarich and Wilson (1967) have shown that the logarithm of *I.D.* is approximately linearly related to the time since divergence between the two organisms tested. They called  $d_i = 100 \times \log_{10} I.D.$  the immunological distance. (They used the notation *ID* for  $d_i$ .) In many proteins (albumin, lysozyme, ribonuclease, etc.), this  $d_i$  is linearly related to the proportion of different amino acids between the two sequences compared (Prager and Wilson 1971; Benjamin et al. 1984). The reason why  $d_i$  should be a linear function of the proportion of different amino acids is not well understood. However, this empirical property of  $d_i$  is very useful for measuring the genetic distance between species, since the technique is much simpler than amino acid sequencing. According to Maxson and Wilson (1974), one unit of  $d_i$  corresponds to roughly one amino acid difference in albumin.

The relationship between  $d_i$  and the time (*t*) since divergence between two species may be written as

$$t = cd_i, \quad (4.13)$$

where *c* is the proportionality constant and varies with the protein used and also to some extent with the group of organisms used. The *c* value

that has been used for albumin is  $(5.5 - 6) \times 10^5$  in mammals, reptiles, and frogs, and  $1.9 \times 10^6$  in birds (Prager et al. 1974; Wilson et al. 1977a; Collier and O'Brien 1985). This suggests that the evolutionary rate of albumin is about three times slower in birds than in other organisms. However, the estimate of the evolutionary rate for birds may be erroneous, since the fossil records used are quite unreliable. If Alvarez et al.'s (1980) asteroid impact theory (chapter 2) is correct, most orders of birds might have evolved relatively recently (Wyles et al. 1983). If this is the case, the difference in the rate of albumin evolution between birds and the other groups of vertebrates might disappear.

For technical reasons, albumin cannot be used for microcomplement fixation in invertebrates. Beverley and Wilson (1984, 1985), therefore, used a larval hemolymph protein in the study of the phylogenetic relationship of various species of Drosophilidae and higher Diptera. Using information on fossils in amber, continental drift, island formation, etc., they estimated that the proportionality constant for this protein is  $c = 8 \times 10^5$ . This is of the same order of magnitude as that for albumin.

The estimate of  $t$  obtained by equation (4.13) is subject to four different types of errors. The first type is experimental error. According to Sarich and Wilson (1966), this error is generally less than 2 percent of the estimate, even if the estimate is relatively small. The second type of error arises when the antigen (protein) used is polymorphic in the species examined. When distantly related species are compared, however, this type of error is relatively small. The third type of error is generated by the fact that amino acid substitution in antigenic proteins occurs stochastically rather than deterministically. As mentioned earlier, this type of error will make the variance of  $d_i$  larger than the mean. Thus, it is much more significant than the first two types. The fourth type of error occurs because  $d_i$  is not strictly proportional to the number of amino acid substitutions (Champion et al. 1974). This type of error could be as important as the third type.

Nei (1977b) examined the (total) variance of  $d_i$  empirically by using  $d_i$  values obtained for various groups of organisms. He concluded that the variance of  $d_i$  is at least twice as large as the mean when the mean is small and that the ratio of the variance to the mean increases with increasing mean. Examining more extensive data for both albumin and larval hemolymph protein, Beverley and Wilson (1984) also showed that the variance of  $d_i$  is at least four times as large as the mean. Therefore, the immunological distance clock seems to be a little less accurate than the amino acid sequence clock.

Despite this inaccuracy, application of this method has brought about many interesting results in the study of evolution. Sarich and Wilson (1966, 1967) used this method to clarify the phylogenetic relationship of humans and apes. The results obtained were surprising. Unlike the prevailing view at that time, they showed that chimpanzees and gorillas are more closely related to humans, which belong to a different family (Hominidae), than to orangutans and gibbons, with which they form one family (Pongidae). Furthermore, their data indicated that humans, chimpanzees, and gorillas diverged about 5 MY ago, as mentioned earlier (chapter 2).

Another interesting finding is that different frog species belonging to the same genus often have large immunological distances similar to the values for different families or orders of mammals (Maxson et al. 1975; Post and Uzzell 1981; Maxson 1984). This suggests that frog species diverged a long time ago but retained similar morphological characters. Similarly, some pairs of *Drosophila* species apparently diverged a long

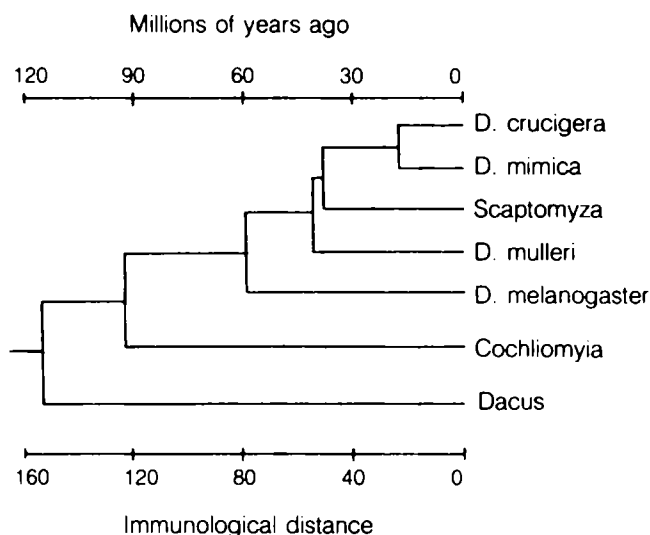


Figure 4.6. Phylogenetic tree reconstructed from immunological distance data for seven fly lineages. *D. crucigera* and *D. mimica* are Hawaiian drosophilids, whereas *D. mulleri* and *D. melanogaster* are continental drosophilids. *Scaptomyza* is a fly genus closely related to *Drosophila*. *Drosophila*, *Cochliomyia*, and *Dacus* belong to different families. From Beverley and Wilson (1985).



time ago. Beverley and Wilson (1984, 1985) estimate that the species belonging to subgenus *Drosophila* (e.g., *D. virilis*, *D. mulleri*) diverged from the *D. melanogaster* species group about 60 MY ago, whereas Hawaiian *Drosophila* species diverged from subgenus *Drosophila* about 40 MY ago (figure 4.6).

The Hawaiian archipelago is known to be no older than 5–6 MY. Therefore, one might expect that all Hawaiian *Drosophila* species diverged relatively recently. Immunological data, however, suggest that the picture-winged group (e.g., *D. crucigera*) and the modified-mouth-parts group (e.g., *D. mimica*) of Hawaiian species diverged about 20 MY ago. Furthermore, the Hawaiian flies belonging to a different genus, *Scaptomyza*, are closer to Hawaiian drosophilids than to continental drosophilids, indicating that the Hawaiian fly fauna originated about 40 MY ago. This paradoxical finding is understandable if we consider the history of the Koko Seamount–Midway–Hawaii Archipelago, which has witnessed the sequential rise and erosion of many islands during the past 70 MY (see Beverley and Wilson 1985).