J. Mol. Evol. 3, 161—177 (1974)

An Examination of the Constancy of the Rate of Molecular Evolution*

Charles H. Langley** and Walter M. Fitch

Departments of Medical Genetics and Physiological Chemistry, University of Wisconsin, Madison, Wisconsin 53706

Received February 4, 1974

Summary. The vertebrate evolution of four proteins (α and β hemoglobins, cytochrome *c*, and fibrinopeptide A) is examined via a maximum likelihood procedure. The fundamental hypothesis is that the process of nucleotide substitution as revealed by the minimum phyletic distance procedure (Fitch, 1971) is Poisson with a constant time average for each protein. The method allows the simultaneous estimation of the relative times of divergence of all common ancestors while utilizing the information from all four proteins. It also affords the possibility of statistically testing several biologically meaningful hypotheses. The results are the following:

1. The total rate (sum of all proteins) of nucleotide substitution is not constant in time throughout the evolution of vertebrates.

2. The relative rates (among proteins) of nucleotide substitution are not constant throughout vertebrate evolution.

3. Despite the variation in the rates of nucleotide substitution the procedure employed provides estimates of the relative time of divergence which correlate well with paleontological dates.

4. The overall rate of nucleotide substitution within the primates is again found to be less than the rest of the mammals.

Key words: Rate of Molecular Evolution — Neutral Mutations — Protein Sequences — Primate Evolution — Mammalian Phylogeney.

Introduction

The evolutionary analysis of amino acid sequence data has been approached from several directions. The fundamental conclusion of even the most rudimentary analysis is that amino acid sequence differences correlate well with morphological and paleontological considerations. The amino acid sequence differences seem to accumulate with time, thus reflecting phylogenetic relationships. This observation led some to view the evolutionary process on the molecular level as a simple random process with a constant average rate (Kimura, 1968, 1969; King and Jukes, 1969).

* This is paper No. 1679 of the Laboratory of Genetics, University of Wisconsin-Madison.

** Present address National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709.

Kimura has suggested that the apparent uniformity of the rate of molecular evolution concurs with the prediction of the population genetics theory of selectively neutral mutations (Kimura, 1968). He and others have argued that data from amino acid sequences indicate that most substitutions were, in fact, selectively neutral (Kimura, 1968, 1969; King and Jukes, 1969; Jukes and King, 1971). Utilizing the model of a uniform random process we have re-examined the vertebrate sequence data of four proteins in an effort to better describe the historical facts and clarify the issue of constancy of rates. We find that the rates of evolution (as seen through the minimum phyletic distance analysis) change significantly in time and among lines of descent. These findings are consistent with our preliminary results reported previously (Langley and Fitch, 1973).

Analysis

Hypothesis: Uniform Rates

We take as the test hypothesis the assumption that the sequence differences among species in various proteins are the manifestation of the fixation through random drift of selectively neutral substitutions. The following simple algebra indicates why the rates of substitution of selectively neutral mutations should be independent of population size.

$$\lambda = 2N\mu(1/2N) = \mu$$

where the number of newly arising selectively neutral mutants each generation is $2N\mu$ ($2N$ = number of genes in a diploid population of size N; μ = the per gene per generation mutation rate) and the probability of fixation of a newly arising neutral mutant is $1/2N$. As a further assumption, let μ be constant over time and among lines of descent. More precisely our hypothesis is that the sequence differences are the manifestation of a set of independent though logically related homogeneous Poisson stochastic processes. In the Appendix is a more thorough development of this statement.

Model

From the above hypothesis we can construct a simple model of the evolutionary process of protein sequence differentiation. The expected number of substitutions in the m^{th} protein in a particular evolutionary interval (the i^{th}) is $\lambda_m(t_k - t_i)$ where t_k and t_i are the times of occurrence of the beginning and end of the evolutionary interval (measured in total substitutions in all proteins being considered) and λ_m is the proportionate rate of substitutions of the m^{th} protein ($\Sigma\lambda_m = 1$). Following the assumption above, the likelihood of observing $x_{m,i}$ substitutions in the m^{th} protein in

the i^{th} interval is

$$L(m, i) = \text{Exp}[-\lambda_m(t_k - t_i)] \left\{ \frac{[\lambda_m(t_k - t_i)]^{x_{m,i}}}{(x_{m,i})!} \right\}$$

$x_{m,i}$ has a Poisson probability distribution. From the assumption of homogeneity (or constancy) we assume that λ_m is constant. Thus, we can write the likelihood of observing a whole phylogeny by simply multiplying over interval and proteins:

$$L = \prod_m \prod_i L(m, i).$$

Maximum likelihood estimators of λ and t can be obtained by taking the logarithm of L and differentiating with respect to each variable. In the Appendix the uniqueness of the estimators is pointed out and the procedures for handling missing observations are discussed. Given a set of observed substitutions in each evolutionary interval for various proteins the estimators can be applied to obtain the maximum likelihood estimates, $\hat{\lambda}$ and $\hat{t}$. These are the values of λ_m and t_i that maximize the probability of obtaining the observed data under the assumed model.

Observations

There is presently only one possible source of data at all appropriate to the maximum likelihood estimators presented above: the minimum phyletic distance procedure as outlined in Fitch (1971). This procedure determines the minimum number of nucleotide substitutions required to explain the amino acid sequence differences, given an *a priori* phylogenetic relationship. It also distributes these substitutions over the branches (or intervals) of the phylogenic tree in an appropriate probabilistic manner (Fitch, 1971). Thus, for each protein in each interval the minimum phyletic distance is determined uniquely. Figs. 1–3 are the phylogenies of the proteins we have studied. We chose this dendrogram as the most reasonable estimate of the species relationships. Fig. 1 depicts the phylogenetic relationships among the hemoglobins as indicated by the minimum phyletic distances. Note the lamprey sequences are included in the hemoglobin and cytochrome *c* phylogenies. Table 1 contains the references for the sequences used in deriving the minimum phyletic distances.

Estimates

By utilizing the minimum phyletic distances as the observed substitutions, $x_{m,i}$ (in the maximum likelihood estimators), we have obtained the estimates $\hat{\lambda}$ and $\hat{t}$. Fig. 4 depicts the temporal relationships as given by the maximum likelihood estimates of the t_i. The abscissa of Fig. 4 is in units of expected observable nucleotide substitutions in all 4 proteins. The heights of the various nodes are placed to correspond to the maximum likelihood

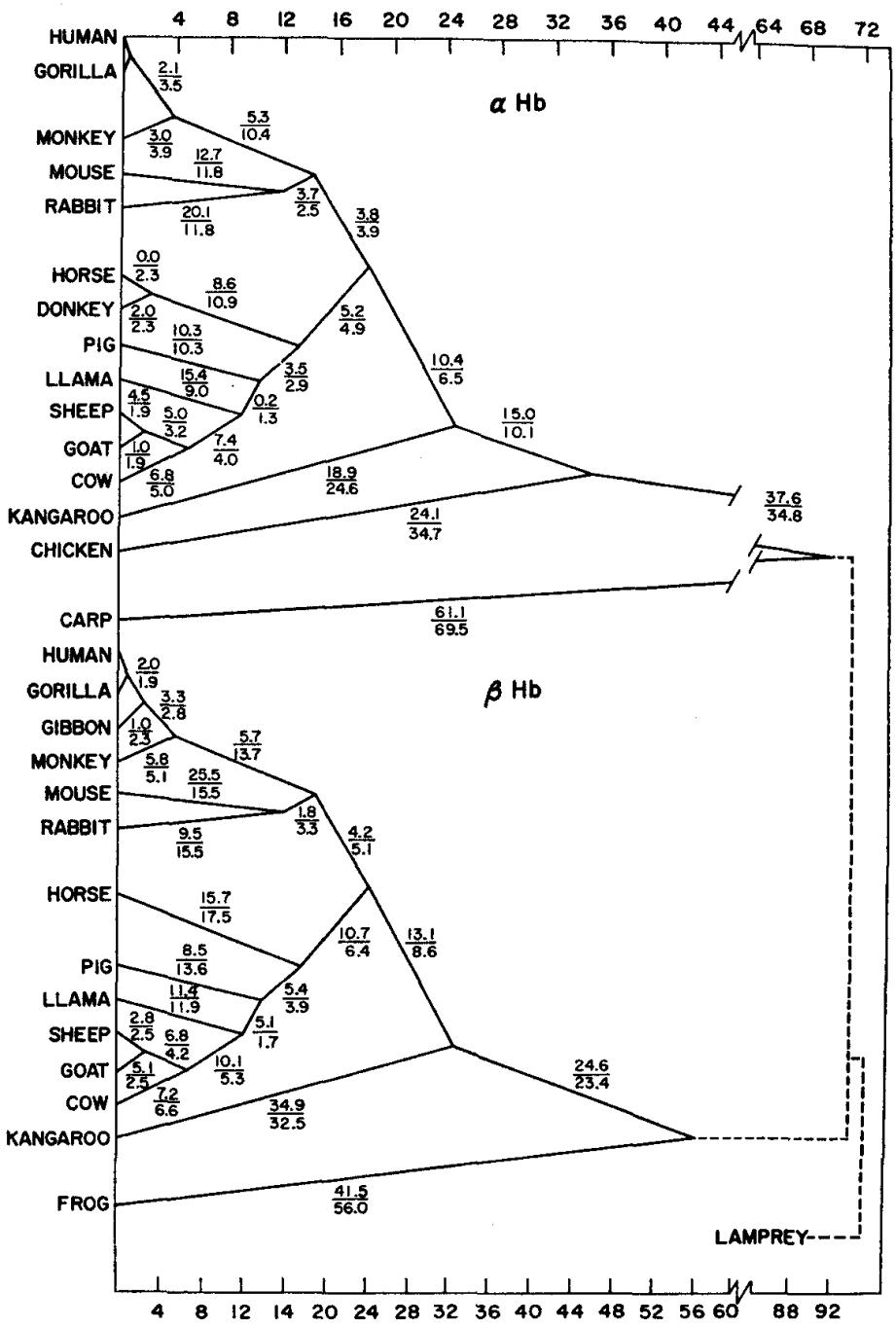

Fig. 1. Phylogenies for α and β hemoglobins. The dendrogram was assumed and was the same regardless of the protein considered. The upper number of each ratio is the observed number of substitutions assigned to that leg by the minimum phyletic distance method of Fitch (1971). The lower number of each ratio is the expected number of substitutions according to the procedures of this paper. All nodes are plotted at heights equal to the expected number of substitutions (abscissa) in the designated protein and are directly proportional to those of the overall solution presented in Fig. 4. Where a sequence was not available, that species and its ancestral node, as shown in Fig. 4, were omitted. Since the α and β hemoglobins were the presumptive results of a gene duplication after the divergence from the lamprey, the complete set of sequences was utilized in determining the observed substitutions by the minimum phyletic distance procedure. The dashed line indicates that portion of the tree not included in the further analysis. The ratios not shown are to Human (for α hemoglobin 1.0/0.3 and for β hemoglobin 0.0/0.4) and to Gorilla (for α hemoglobin 1.0/0.3 and for β hemoglobin 1.0/0.4). The values along the upper abscissa are for α hemoglobin, those along the lower abscissa are for β hemoglobin

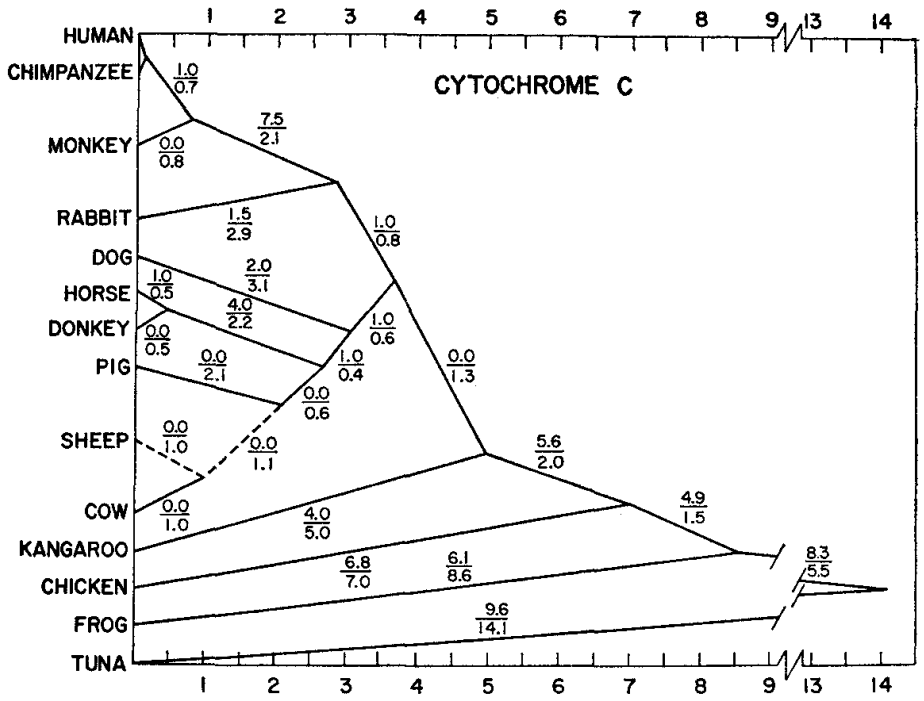

Fig. 2. Phylogeny for cytochrome *c*. All nodes are plotted at heights equal to the expected number of substitutions (abscissa) in the designated protein and are directly proportional to those of the overall solution presented in Fig. 4. See Fig. 1 for general explanation. The lamprey sequence was included in the minimum phyletic procedures. The dashed lines leading from the sheep to the ancestral artiodactyl indicate specially treated data. Vis-a-vis the complete phylogeny of Fig. 4, these two lines are missing nodes formed from the goat and llama. Since no substitutions occurred in the legs shown, it was assumed for the computations in the tests of the hypothesis that the division of each of these two lines into two segments, also possessing no substitutions, introduced no unreasonable biases. The ratios not shown are to Human (0.0/0.1) and to Chimpanzee (0.0/0.1)

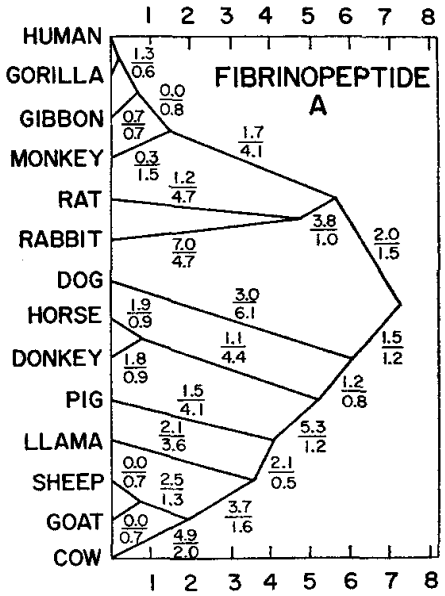

Fig. 3. The phylogeny for fibrinopeptide A. All nodes are plotted at heights equal to the expected number of substitutions (abscissa) in the designated protein and are directly proportional to those of the overall solution presented in Fig. 4. See Fig. 1 for general explanation. The ratios not shown are to Human (0.0/0.1) and to Gorilla (0.0/0.1)

Table 1. References for protein sequences used. Table 1a contains the number for the various references. These references are listed by these numbers in Table 1b

Table 1a. References to sequences used

		α Hemoglobin	β Hemoglobin	Cytochrome *c*	Fibrinopeptide A
Human	*(Homisapiens)*	1	1	2	3
Pongid[a]		4	4	5	6
Gibbon	*(Hylobates lar)*	—	7	—	6
Monkey	*(Macaca mulatta)*	8	8	9	10
Rodent[a]		11	12	—	10
Rabbit	*(Oryctolagus cuniculus)*	13	14	15	10
Dog	*(Canis familiaris)*	—	—	16	10
Horse	*(Equus caballus)*	17	18	19	10
Donkey	*(Equus asinus)*	17	—	20	10
Pig	*(Sus scrofa)*	21	22	23	24
Llama	*(Llama peruana)*	25	25	—	10
Sheep	*(Ovis aries)*	26	27	28	24
Goat	*(Capra* sp.*)*	29	30	—	24
Cow	*(Bos taurus)*	31	32	33	34
Kangaroo[a]		35	35	36	—
Chicken	*(Gallus gallus)*	37	—	38	—
Frog[a]		—	39	40	—
Fish[a]		41	—	42	—

[a] A single species was used for each protein, with the following exceptions: For rodent α and β hemoglobins, C57BL mouse (*Mus musculus*) was used, but rat (*Ratus norvegicus*) was used for fibrinopeptide A. For α and β hemoglobins the kangeroo *Macropus giganteus* was used, but for cytochrome *c Macropus cangura* was used. The frog *Rana esculenta* was used for β hemoglobin, while *Rana catesbiana* was used for cytochrome *c*. Carp (*Cyprinus carpio*) was used for α hemoglobin and tuna was used for cytochrome *c*. Among the pongids, *Gorilla gorilla* was the source of the hemoglobins and fibrinopeptide A, while the cytochrome *c* was from *Pan troglydytes*.

Table 1b

1. Braunitzer, G., Gehring-Muller, R., Hilschmann, N., Hilse, K., Hobom, G., Rudloff, V., Wittmann-Liebold, B.: Z. Physiol. Chem. **325**, 283 (1961)
2. Matsubara, H., Smith, E. L.: J. Biol. Chem. **237**, PC3575 (1962)
3. Blomback, B., Blomback, M., Edman, P., Hessel, B.: Biochim. Biophys. Acta **115**, 371 (1966)
4. Zuckerkandl, E., Schroeder, W. A.: Nature **192**, 984 (1961)
5. Needleman, S. B., Margoliash, M.: Unpublished
6. Doolittle, R. F., Wooding, G. L., Lin, Y., Riley, M.: J. Mol. Evol. **1**, 74 (1971)
7. Boyer, S. H., Crosby, E. F., Noyes, A. N., Fuller, G. F., Leslie, S. E., Donaldson, L. J., Vrablik, G. R., Schaefer, E. W., Jr., Thurmon, T. F.: Biochem. Genet. **5**, 405 (1971)
8. Matsuda, G., Maita, T., Takei, H., Ota, H., Yamaguchi, M., Miyauchi, T., Migita, M.: J. Biochem. (Tokyo) **64**, 279 (1968)
9. Rothfus, J. A., Smith, E. L.: J. Biol. Chem. **240**, 4277 (1965)
10. Blomback, B., Blomback, M., Grondahl, N. J., Guthrie, C., Hinton, M.: Acta Chem. Scand. **19**, 1788 (1965)

Table 1 b (continued)

11. Popp, R. A.: J. Mol. Biol. **27**, 09 (1967)
12. Rifkin, D. B., Rifkin, M. R., Konigsberg, W.: Proc. Natl. Acad. Sci. U.S. **55**, 586 (1966)
13. Braunitzer, G., Flamm, U., Best, J. S., Schrank, B.: Z. Physiol. Chem. **349**, 1073 (1968)
14. Braunitzer, G., Best, J. S., Flamm, U., Schrank, B.: Z. Physiol. Chem. **347**, 207 (1966)
15. Needleman, S. B., Margoliash, E.: J. Biol. Chem. **241**, 853 (1966)
16. McDowall, M. A., Smith, E. L.: J. Biol. Chem. **240**, 4635 (1965)
17. Kilmartin, J. V., Clegg, J. B.: Nature **213**, 269 (1967)
18. Smith, D. B.: Canad. J. Biochem. **46**, 825 (1968); **42**, 755 (1964)
19. Margoliash, E., Smith, E. L., Kreil, G., Tuppy, H.: Nature **192**, 1121 (1961)
20. Walasek, O. F., Margoliash, E.: Unpublished
21. Yamaguchi, Y., Horie, H., Matsuo, A., Sasakawa, S., Satake, K.: J. Biochem. (Tokyo) **58**, 186 (1965)
22. Braunitzer, G., Kohler, H.: Z. Physiol. Chem. **343**, 290 (1966)
23. Stewart, J. W., Margoliash, E.: Canad. J. Biochem. **43**, 1187 (1965)
24. Blomback, B., Doolittle, R. F.: Acta Chem. Scand. **17**, 1819 (1963)
25. Braunitzer, G., Hilse, K., Rudloff, V., Hilschmann, N.: In: Advances in protein chemistry, Vol. 19, p. 34. New York: Academic Press 1964
26. Beale, D.: Biochem. J. **103**, 129 (1967)
27. Boyer, S. H., Hathaway, P., Pascasio, F., Bordley, J., Orton, C., Naughton, M. A.: J. Biol. Chem. **242**, 2211 (1967)
28. Chan, S. K., Needleman, S. B., Stewart, J. W., Margoliash, E.: Unpublished
29. Hisman, T. H. J., Brandt, G., Wilson, J. B.: J. Biol. Chem. **243**, 3675 (1968)
30. Huisman, T. H. J., Adams, H. R., Dimmock, M. O., Edwards, W. E., Wilson, J. B.: J. Biol. Chem. **242**, 2534 (1967)
31. Schroeder, W. A., Shelton, J. R., Shelton, J. B., Robberson, B., Babin, D. R.: Arch. Biochem. Biophys. **120**, 1 (1967)
32. (B-chain) Schroeder, W. A., Shelton, J. R., Shelton, J. B., Robberson, B., Babin D. R.: Arch. Biochem. Biophys. **120**, 1 (1967)
33. Nakashima, T., Higa, H., Matsubara, H., Benson, A. M., Yasunobu, K. T.: J. Biol. Chem. **241**, 1166 (1966)
34. Folk, J. E., Gladner, J. A., Levin, Y.: J. Biol. Chem. **234**, 2317 (1959)
35. Air, G. M., Thompson, E. O. P., Richardson, B. J., Sharman, G. B.: Nature **229**, 391 (1971)
36. Nolan, C., Margoliash, E.: J. Biol. Chem. **241**, 1049 (1966)
37. Matsuda, G., Takei, H., Wu, K. C., Mizuno, K., Shiozawa, T.: Eighth International Congress in Biochem., Abstr. 4 (1970)
38. Chan, S. K., Margoliash, E.: J. Biol. Chem. **241**, 507 (1966)
39. Chauvet, J. P., Acher, R.: FEBS Letters **10**, 136 (1970)
40. Chan, S. K., Wallasek, O. F., Barlow, G. H., Margoliash, E.: Fed. Proc. **26**, 723 (1967), Abstract
41. Hilse, K., Braunitzer, G.: Z. Physiol. Chem. **349**, 433 (1968)
42. Kreil, G.: Z. Physiol. Chem. **334**, 154 (1963)

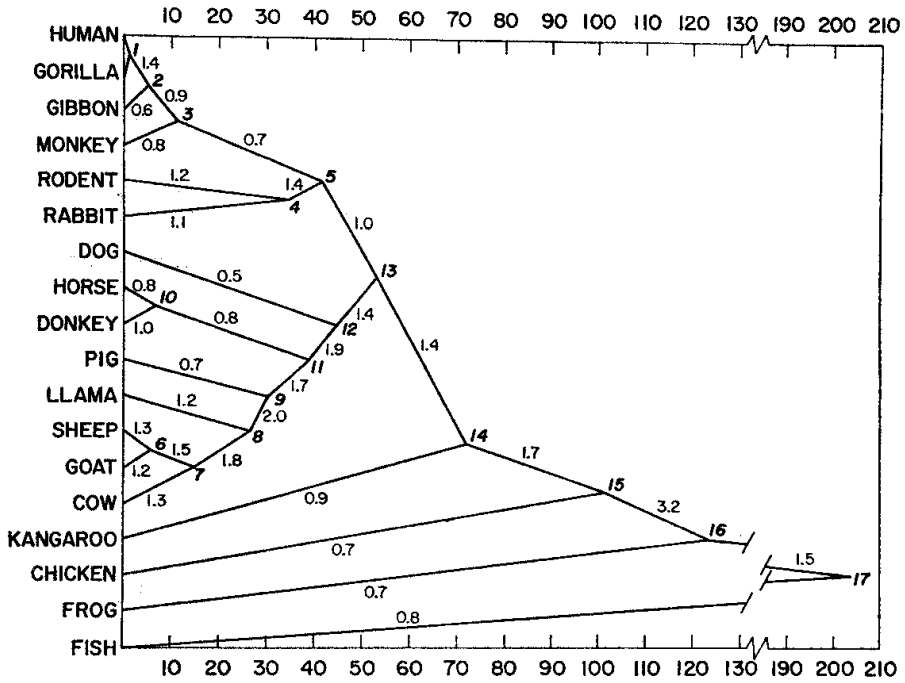

Fig. 4. Composite evolution of hemoglobins α and β, cytochrome c and fibrinopeptide A. The nodes representing common ancestors are placed at a height equal to the predicted total number of observable nucleotide substitutions in all 4 proteins in the descent from a common ancestor to any present day descendants. The numbers along each leg give the ratio of observed/expected substitutions for the proteins examined. The ratios not shown are to Human (0.0) and to Gorilla (2.1). This figure is taken from Langley and Fitch (1973). The italic number at the nodes designate the ancestors (see Fig. 6)

estimates of the node heights, *i.e.* the expected number of observable nucleotide substitutions in all four proteins in the time period from the occurrence of the ancestor to the present day.

The maximum likelihood estimates of proportionate rates of nucleotide substitution are

α hemoglobin	0.342
β hemoglobin	0.452
cytochrome c	0.069
fibrinopeptide A	0.137

Test of Hypothesis

In this section we will analyze the appropriateness of the hypothesis and model. We have obtained concise descriptions of the quantitative relationships in the evolution of these proteins. This has been possible because a simple hypothesis was assumed. As a further and perhaps more interesting question, we can test whether the model is correct or adequate. Essentially

the test is one in which we ask whether the observed number of substitutions $x_{m,i}$ is sufficiently close to the expected $\hat{\lambda}_m(\hat{t}_k - \hat{t}_i)$.

Previously we (Langley and Fitch, 1973) utilized a simple goodness of fit χ^2 analysis: (Observed—Expected)^2/Expected. This is useful in analyzing particular proteins and evolutionary intervals as to their deviation from expectation. Here we shall apply a likelihood ratio test which is more appropriate and powerful. The conclusions are not affected by this alternative analysis.

Under both of these hypothesis-testing procedures it is possible to break down or partition the test into component subtests. Our hypothesis is that the probability of observing $x_{m,i}$ is simply a function of $\lambda_m(t_k - t_i)$. The subtests mentioned above are simply tests of weaker hypotheses: 1. λ_m is constant for all lines of descent; and 2. the total rate is constant for different lines of descent. To elaborate we notice that $\lambda_m(t_k - t_i)$ is the product of the expected proportion of substitutions in the m^{th} protein and $(t_k - t_i)$, the expected substitutions in all proteins in the leg between nodes k and i. Thus we can separately test both hypotheses: 1. the relative rates of substitution are constant, and 2. the total rate of substitution is constant. We have designated these partitions: "among proteins within legs" and "among legs over proteins". The Appendix contains a more thorough development of the testing procedures and a discussion of the effect of missing data points.

Table 2 contains the χ^2 values for the various likelihood ratio tests performed. The "corrected" columns refer to procedures presented in the next section. These χ^2 values agree quite well with those obtained by the goodness of fit analysis (Langley and Fitch, 1973). It is quite clear that the hypothesis of overall constant evolutionary rate for each protein or even overall constancy for this group of proteins as a unit must be rejected. Similarly the relative rates are not constant as shown by the small probability in the test for the partition among proteins within legs.

Table 2. Tests of hypotheses

	Uncorrected			Corrected		
	χ^2	(df)	$P<$	χ^2	(df)	$P<$
Among proteins within legs (relative rates)	102.7	(62)	10^{-3}	102.7	(62)	10^{-3}
Among legs over proteins (total rates)	63.0	(26)	10^{-4}	48.7	(24)	0.002
Total	165.7	(88)	10^{-5}	151.4	(86)	10^{-4}

Correction of Bias

We have utilized the minimum phylectic distance (Fitch, 1971) as our "data" in this analysis. One of the obvious criticisms of this choice is that there is a clear bias in this procedure. Since the minimum phyletic distance is the minimum number of substitutions required to explain the observed sequence difference, it will more severely underestimate the number of substitutions in longer evolutionary intervals (Holmquist, 1972a, b). An examination of Fig. 4 clearly shows that the longer legs are more often shorter than expected. Fig. 5 is a plot of the number of observed substitutions in each leg (summed over all the observed proteins) against the expected number of substitutions (in those proteins). In this figure we note a tendency for the larger values to fall below the expected-equals-observed line and for smaller values (less than 30) to lie above the line. This impression is confirmed by a least squares regression (the dashed line) which does have a significant quadratic coefficient. The inherent bias in the minimum phyletic distance has affected the maximum likelihood estimations so that long legs are overestimated and short legs are underestimated.

The bias undoubtedly enters into the hypothesis test. Judging from the regression analysis, we can expect that a substantial proportion of the χ^2 for the test among legs over proteins is due to the bias. In order to estimate this we simply "corrected" the expected totals for each leg according to the regression formula graphed in Fig. 5. This left the test, among proteins within legs unaltered but it did significantly lower the χ^2 for the test, among legs over protein, and therefore the total χ^2 (see Table 2).

Only 2 parameters were estimated in fitting the regression line since it was forced through the origin. Thus with a loss of only 2 degrees of freedom

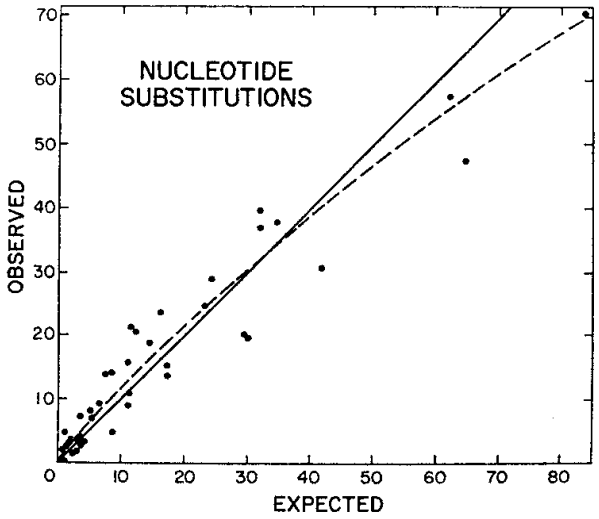

Fig. 5. Observed total nucleotide substitutions by legs plotted against expected nucleotide substitutions in those proteins. The dashed line is the least square quadratic regression of these points (constrained to intercept 0.0). Expected = Observed [0.8214 + Observed (0.0052)]

the χ^2 is reduced over 14 units. This is, however, not enough to affect the previous conclusion. Even after correcting for the bias in the minimum phyletic distance procedure both hypotheses must be rejected.

Comparison with Geological Dating

Throughout this analysis we have not had to refer to any geological dating whatsoever. The times of occurrence of various common ancestors are in units of nucleotide substitutions. We had assumed that this is a constant phenomenon and thus an appropriate unit of time. In the last 2 sections, however, we have seen that nucleotide substitutions do not occur at a constant rate. This weakens our confidence in the "dates" (relative distance from the origin) in Fig. 4. We note that despite these important shortcomings the dating in Fig. 4 may be superior to any alternative procedure based on amino acid sequencing. It is from this point of view that we compare these dates with the geological dating.

The estimates of the last common ancestor of mammals and fish, amphibians, or birds are very poor; each is based on only one hemoglobin and cytochrome *c*. Because of this and controversy in the geological dating, we hesitate to compare them. Most of the data are for the mammals and here there are fairly good geological dates. Fig. 6 is a plot of the estimates

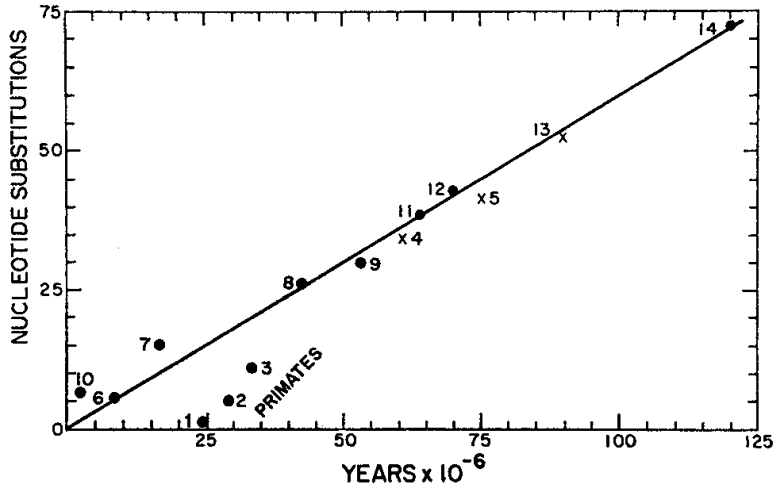

Fig. 6. Maximum likelihood estimates of times of occurrence of various mammalian common ancestors (from Fig. 4) plotted against geological dates. The geological dates were provided by Dr. L. von Valen (pers. comm.). *X* indicates a geological date provided by Dr. van Valen despite his reservations about the exact phylogenetic relationship. The slope of the line is 0.6 observable nucleotide substitutions per million years. Since these gene products are encoded by a total of 1230 nucleotides, this is equivalent to 0.49×10^{-9} substitutions/yr/nucleotide for those substitutions changing the amino acid. If those not changing the amino occur at least as often, the actual rate is $> 0.65 \times 10^{-9}$ substitutions/yr/nucleotide. This may be compared to Kohne *et al.*'s (1972) rate of 0.9×10^{-9} substitutions/yr/nucleotide obtained from hybridization studies of primate non-repeated (unique) DNA (their Table IV excluding Δ's, D, E, K, and L). Numbers refer to the comparably labelled nodes of Fig. 4 and the ordinate is the same as in that figure

in Fig. 4 against several geological dates (L. Van Valen, pers. commun.). The fit is remarkably good except among the primates where the rate of substitution appears to be slower. This retardation among primates lines of descent is apparent in the protein data alone and was originally noted by Goodman *et al.* (1971) and Barnabas *et al.* (1971) for the hemoglobins. Indeed, it is the hemoglobins, which account for nearly 80% of the observed substitutions, that are responsible for the observation. The ratios of observed to expected among the primates in Fig. 4 are consistently less than one. The geological dating apparently substantiates this observation.

Discussion

Procedures and Methods

Since the maximum likelihood equations are nonlinear in t and λ, nonlinear programming techniques were utilized. The solutions presented here were obtained by the GPM nonlinear programming package of the Madison Academic Computing Center of the University of Wisconsin, Madison. There are 2 available criteria upon which the accuracy of the solution can be judged. 1. The magnitudes of the derivatives at the solution were all quite small (less than 10^{-4}). This is not a critical criterion since some of the derivatives may be small in a large neighborhood of the solution. 2. Several maximizations were carried out using grossly different initial conditions. All of these gave the same solution to at least the 3 significant digit in each variable. Thus we can conclude that the solution is perhaps accurate to that level (3 significant digits). Considering the numbers in the data, this is more than adequate.

As noted in the Appendix, the independence of the two subtests (among proteins within legs and among legs over proteins) is assured only when observations are present for all proteins in all the legs, *i.e.* there are no missing data. The results presented above were obtained from incomplete data. We therefore examined the strength of the results with respect to this possible criticism. To accomplish this we analyzed these same 4 proteins over that portion of the phylogeny for which the observations are complete. The results of that study are essentially identical to those presented above. The estimates of λ and t are all quite similar and, most importantly, the hypothesis tests are qualitatively the same although the numerical values are different. This leads us to conclude that no serious errors have been introduced by the analysis of these incomplete data.

The greater variation in total rates compared to variation in relative (proportionate) rates was noted before (Langley and Fitch, 1973). This is reflected in the ratios of χ^2/degrees of freedom (among proteins within legs, $\chi^2/df = 1.66$; among legs over proteins, $\chi^2/df = 2.25$). One interpretation might be that most of the variation in rates of nucleotide substitution is associated with differences among species or lines of descent, *i.e.*, background

mutation rates, generation times, etc. Thus proteins evolving in the same species would tend to vary together. An alternative explanation of the difference in χ^2/df ratios is the simple fact that almost 80% of the substitutions are in two proteins (α and β hemoglobins) that not only have similar function but form a heterologous tetramer together. With only 4 proteins and 2 comprising 80%, the power of the partition is less because of the small number of classes and their skewed distribution. The 2 hemoglobins could well show positive correlation in their rates since they are similar components of the same functional molecule. Such a correlation would tend to effect the total rate more than the relative rates since the 2 hemoglobins account for almost 80% of the total nucleotide substitutions observed. These considerations lead us to interpret the difference in χ^2/df ratios between the two partitions as reflecting differences in power and possibly effects of correlations between hemoglobins. We do not interpret this as direct evidence that variation in rates is predominantly due to species differences, although it it clearly consistent within such an interpretation.

Previously we (Langley and Fitch, 1973) noted that the large χ^2 for the partition among proteins within legs is due to a small number of observations, some with small expected values. At face value it appeared that most of the observations did indeed fit the hypothesis of constant relative rates of nucleotide substitution and the large χ^2 was due to these few anomalous occurrences. Further consideration suggests an alternative interpretation which we have adopted. The rate of molecular evolution is clearly a bounded and buffered process, *i.e.* an autocorrelated stochastic process. If it is indeed not constant, the variability in rate from time to time must be bounded. This can lead to an averaging process such that the longer an evolutionary interval is, the more closely its overall rate will approach the "average". Suppose for example the rate cycled with time such that the rates between short intervals might differ by a factor of 2. Yet over long intervals the rates would approach the same average. If a sample of observations over various intervals were examined, the longer intervals would be expected to show constancy of rate while the short intervals would demonstrate the non-constancy.

We have no reason to believe that the rate of molecular evolution is cyclic. But we can safely assume that the rate changes from time to time within a bounded range. This leads us to expect a similar pattern: longer intervals showing less statistical deviation from the "average" than shorter intervals. This is what we, in fact, observe. Most of the observations which appear to be significantly deviant have small expected values. The obvious example is cytochrome *c* in the interval between the common ancestor of primates and the common ancestor of primates and rodents (see Fig. 2). It is this consideration which has led us to accept the hypothesis tests at face value.

The Constancy of the Rate of Nucleotide Substitution and the Neutral Hypothesis

On the basis of these results we reject the null hypothesis and its underlying assumption of uniformity of rates of nucleotide substitution. This could be criticized on the grounds that the minimum phyletic distances used as data are biased. However, after attempting to correct this bias we reject the hypothesis. The deviations of the observations from their expectations can in no reasonable way be accounted for completely by the inherent bias in the minimum phyletic procedure.

The neutral theory predicts that the rate of substitution is equal to the neutral mutation rate. The "apparent" constancy of molecular evolution has been sighted as evidence for this theory (Kimura, 1968). An implicit assumption of this type of analysis is that the neutral mutation rates are fairly uniform in geological time and among lines of descent. We cannot judge the validity of this assumption. However, it is clear that the *total* rate of substitution (as observed through the minimum phyletic distance procedure) varies markedly in geological time and among divergent lines of descent. This does not rule out the possible selective neutrality of substitutions that did in fact occur. It simply rejects the uniformity of rates in time and among lines of descent.

The rejection of the hypothesis of constancy of relative rates of substitution deals a second and more severe blow to the hypothesis that the amino acid sequence differences are the product of the random drift of neutral mutations. The relative rates of nucleotide mutation (selective and neutral) must be simply a function of the relative number of encoding nucleotide positions since the genes did evolve in the genomes of the same individuals and were replicated and repaired under what must have been identical circumstances. We have utilized only amino acid positions common to all lines of descent. Therefore it is reasonable to assume that the relative rates of total mutation (selective and neutral) were constant over the entire evolutionary history of the genes. The observed relative rates of substitution are not constant. Two interpretations are available. 1. Almost all substitutions are neutral, but the number of possible neutral mutations per gene varies significantly in time and independently of other genes. This interpretation is tantamount to maintaining that whatever variation in rate is observed it is simply a reflection of variation in number of possible neutral mutations per gene. Or, put another way, the study of rates of molecular evolution is not cirtical in the validation or rejection of the neutral theory. We can only note that this interpretation contributes no real understanding of nature and, more importantly, it is *ex post facto*. After all, one of the primary sources of evidence for the neutral hypothesis was the "apparent" constancy of rates of molecular evolution.

2. The second (and to us more reasonable) interpretation is that a large proportion of the observed substitutions were selectively fixed and can be expected to reflect environmental changes. There is no reason to expect relative rates of selective substitution to be constant in time or among lines of descent. We also note that this interpretation is well known and thus does not have the stigma of being an after-the-fact explanation.

Phylogenetic Dating

Despite the significant variation in rates of substitution among legs, the node heights in Fig. 4 correlate well with the paleontological dates provided by Dr. Van Valen. This is noteworthy on three counts. 1. Dr. Van Valen had no knowledge of our results when he gave us the dates. 2. No other attempt at such a correlation has, to our knowledge, performed as well. 3. We have made no correction for hidden substitutions. This result is perhaps less surprising when we note that the individual t_i or node height is directly affected by all 3 connecting intervals and indirectly by all intervals. Thus there is considerable information in the determination of a given t_i. Deviations among connecting intervals could easily compensate giving a better estimate of the true t_i value. Note also that the distribution of intervals is fairly uniform over the mammals thus reducing any distortion due to the minimum phyletic distance. A third important factor is the number of proteins analyzed simultaneously. The combining of several proteins tends to minimize the effect of deviation in particular proteins within an interval. As protein sequences from other loci become available, they can be included to give even more accurate estimates. For this introductory purpose we feel these four proteins adequately demonstrate the utility of these procedures in constructing phylogenies.

Acknowledgements. We wish to thank Mr. Frank Iltis for his help on the computer work and Drs. James F. Crow and Warren J. Ewens for their considerable advice. We also thank Jerome L. Kreuser of the Academic Computing Center of University of Wisconsin, Madison, for his assistance in solving the maximum likelihood equations and Peter Burrows for his suggestions concerning hypothesis testing. This work was supported by grants NSF 144-C711 (WMF) and NIH GM15422 (CHL).

Appendix

Likelihood Equations

Let $x_{m,i}$ be the observed number of events (nucleotide substitutions) in the m^{th} type (protein) during the evolution from the k^{th} common ancestor to the i^{th} common ancestor. The time of occurrence of a given most recent common ancestor (k^{th}) is t_k and t_i is the time of occurrence of most recent common ancestor (i^{th}) of a smaller phylogenetic subset, or $t_i = 0$ (present day). Assuming that the probability distribution of events ($x_{m,i}$) can be described by a Poisson probability distribution we can write

the likelihood of observing $x_{m,i}$ in terms of t_k, t_i, and λ_m:

$$L(m, i) = \frac{e^{-\lambda_m(t_k - t_i)} [\lambda_m (t_k - t_i)]^{x_{m,i}}}{(x_{m,i})!}$$

where λ_m is the mean proportion of events of the m^{th} type, $\sum_m \lambda_m = 1$; and t_i is now in units of events in all m types. Since we are assuming mean rates of event are constant over time and space it is appropriate to measure time in units of events. The expected number of events in all types in that particular evolutionary interval is $(t_k - t_i)$.

Assuming that the processes are independent across types and over legs of the phylogenetic tree (intervals), we can simply multiply probabilities to obtain a likelihood function for a complete phylogeny:

$$L = \prod_m \prod_i L(m, i).$$

Estimators

By differentiating $\ln L$ with respect of each t_i and λ_m and setting these derivatives equal to zero we obtain the maximum likelihood estimators of t and λ. Further differentiation shows that for all acceptable values of t_i and λ_m, L is globally convex. Thus the estimators must uniquely determine the maximum, if such a maximum exists. We note that this type of estimator was given by Cavalli-Sforza and Edwards (1967) in reference to the study of the evolution of gene frequencies at polymorphic loci.

Hypothesis Tests

The assumption of independence and constant rates can be tested by standard likelihood ratio methods. To test all the assumptions simultaneously the following expression is evaluated:

$$-2 \ln \frac{L(\hat{\lambda}, \hat{t})}{L(x)}$$

where

$$L(x) = \prod_m \prod_i \frac{e^{-x_{m,i}} (x_{m,i})^{x_{m,i}}}{(x_{m,i})!}$$

and

$$L(\hat{\lambda}, \hat{t}) = \prod_m \prod_i \frac{e^{-\hat{\lambda}_m(\hat{t}_k - \hat{t}_i)} [\hat{\lambda}_m (\hat{t}_k - \hat{t}_i)]^{x_{m,i}}}{(x_{m,i})!}$$

and $\hat{\lambda}$ and $\hat{t}$ are the maximum likelihood estimates. This expression is approximately distributed as a χ^2 with degrees of freedom equal to the number of observations minus the number of quantities estimated.

Two interesting subtests are also available. By exchanging $\hat{\lambda}_m (\sum_m x_{m,i})$ for $\hat{\lambda}_m (\hat{t}_k - \hat{t}_i)$ in each term of the numerator, we test the hypothesis of independence among types (proteins). We have designated this test "among proteins within legs." The degrees of freedom in this test are the total number of observations minus the number of estimated $\hat{\lambda}_m$'s and minus the number of intervals. Additionally we can set $\hat{\lambda}_m (\hat{t}_k - \hat{t}_i) = (x_{m,i} / \sum_m x_{m,i}) (\hat{t}_k - \hat{t}_i)$ in the numerator of the ratio. This tests the constancy of total rates of events (nucleotide substitutions). We have designated this subtest "among legs over proteins." The degrees of freedom in this test are the number of intervals minus the number of $\hat{t}_i$'s estimated.

Missing Data

When observations are not present for each type in each interval the initial formulation is the same except that these missing terms are omitted from L:

$$L = \prod_{p} \prod_{i} L(p, i),$$

where p indicates only those combinations (p, i) for which there are observations. The estimators are obtained by the standard differentiation of the logarithm of L.

The primary test of hypothesis is the same as before except that missing terms are not considered. In the subtest "among proteins within legs" $(\hat{\lambda}_m / \sum_p \hat{\lambda}_p)(\sum_p x_{p,i})$ replaces $\hat{\lambda}_m(\sum_m x_{m,i})$ in the numerator of the ratio. The degrees of freedom are the total number of observations minus the number of estimated $\hat{\lambda}_m$'s, minus the number of intervals. In the subtest "among legs over proteins" $(x_{m,i}/\sum_p x_{p,i})[\sum_p \hat{\lambda}_p(\hat{t}_k - \hat{t}_i)]$ replaces $(x_{m,i}/\sum_m x_{m,i})(\hat{t}_k - \hat{t}_i)$ in the numerator of the ratio. The degrees of freedom in this case are the number of intervals minus the number of estimated $\hat{t}_i$'s and $\hat{\lambda}_m$'s.

References

Barnabas, J., Goodman, M., Moore, G. W.: Comp. Biochem. Physiol. **39**B, 455–482 (1971)
Corvalli-Sforza, L. L., Edwards, A. W. F.: Evol. **21**, 550–580 (1967)
Fitch, W. M.: Syst. Zool. **20**, 406–416 (1971)
Goodman, M., Barnabas, J., Matsuda, G., Moore, G. W.: Nature **233**, 604–613 (1971)
Holmquist, R.: J. Mol. Evol. **1**, 115–133 (1972a)
Holmquist, R.: J. Mol. Evol. **1**, 134–149 (1972b)
Jukes, T. H., King, J. L.: Nature **231**, 114–115 (1971)
Kimura, M.: Nature **217**, 624–626 (1968)
Kimura, M.: Proc. Nat. Acad. Sci. **63**, 1181–1188 (1969)
King, J. L., Jukes, T. H.: Sci. **164**, 788–798 (1969)
Kohne, D. E., Chiscon, J. A., Hoyer, B. H.: In: Proc. of the Sixth Berkeley Symposium on Mathematical Statistics & Probability, L. M. LeCam, J. Neyman, E. L. Scott, Eds., pp. 193–209. Berkeley: University of California Press 1972
Langley, C. H., Fitch, W. M.: In: Genetic structure of populations, N. E. Morton, Ed., pp. 246–262. Honolulu: University Press of Hawaii 1973

Dr. Walter M. Fitch
Dept. of Physiological Chemistry
University of Wisconsin-Madison
Madison, Wisconsin 53706, USA