J Mol Evol (1981) 18:47–59

Journal of Molecular Evolution

Letters to the Editor

The Current Status of REH Theory

Richard Holmquist and Thomas H. Jukes

Space Sciences Laboratory, University of California at Berkely, Berkeley, California 94720, USA

Summary. The recent evaluation by Fitch (1980) of REH theory for macromolecular divergence is a severely erroneous and distorted analysis of our work over the past decade. We reply to those distortions here. At present, there is no factual basis for believing Fitch's assessment that corrections which move evolutionary estimates of total mutations fixed closer to the true distance must do so at the expense of an increased variance sufficient to compromise the value of the improvement. By direct calculation the variance in the estimates of total mutations fixed given by REH theory is comparable to that of other models now in the literature for the case in which genetic events are equiprobable. A general argument is given that suggests that, as we consider more and more carefully the selective, functional, and structural constraints on the evolution of genes and proteins, this variance may be expected to decrease toward a lower bound.

Key words: Gene evolution – Protein evolution – Molecular evolution – REH theory – Evolutionary estimates

Introduction

In 1972 we proposed a model (Holmquist et al. 1972), designated REH theory, for the evolutionary divergence of proteins and nucleic acids. This model, like other models in the literature at that time (Zuckerkandl and Pauling 1965; Jukes and Cantor 1969; Holmquist 1972a; Kimura and Ohta 1972) assumed the equiprobability of genetic events of various types. No structural genes had then been sequenced and the precise relationship between such genes and the proteins for which they code was not known because of codon degeneracy. To have made more complicated assumptions would have been an exercise in speculation. Our motivation in proposing the REH model was several-fold. It was a simple extension of existing models - - an extension that gave computed values of the types of amino acid replacements as classified by their minimal base difference values in much better agreement with experiment than other models. Secondly, the model showed clearly that by considering the *type* of amino acid replacement as well as the total *number* of such replacements a more accurate estimate of the total mutations fixed in the gene resulted and that this more accurate estimate was much higher than the values then in vogue and calculated by the method of parsimony (Fitch 1971). It was not until 1976 (Moore et al. 1976: Holmquist et al. 1976) that the parsimony estimates of total fixed mutations approached the values calculated by us from REH theory in 1972. Thirdly, the model showed that to explain the experimentally observed pattern of amino acid replacements, it was necessary for some codon sites, in addition to those absolutely invariant (such as the methionine at residue position 80 in the cytochromes *c)* to be restricted in their ability to fix mutations and that an estimate of the number of such sites (T_1 in REH theory) could be made from the data rather than assumed. The absence of gene sequences made it impossible to know what was happening at the third position within codons and thus precluded a very accurate estimate of T_1. But the principle that the value of this parameter could be established from the data, rather than guessed at (as in all other models), was important.

The REH model for protein divergence was complete in 1972 (Jukes and Holmquist 1972). Two years ago

Offprint requests to: Dr Richard Holmquist, Space Sciences Laboratory RSSF, 1414 Harbour Way South, Richmond, CA 94804, USA

0022–2844/81/0018/0047/$02.60

(Holmquist 1978a) a correction for a statistical bias (Nei and Tateno 1978) was incorporated into the original model. To complete our understanding of how macromolecular systems evolve when genetic events are equiprobable we published the analogous theory for genes (Holmquist 1980), and applied that theory, giving numerical results, to the divergence of human, mouse and rabbit hemoglobin genes. Prior to the latter publication, Holmquist and Pearl (1980) had completed their work on REH theory for the situation in which genetic events were not equiprobable, but subject to arbitrarily large selective or functional constraints. A preprint of the Holmquist and Pearl (1980) paper had been sent to Fitch in May of last year, well in time for his consideration of its findings, which made his own criticism redundant. In addition Professor Fitch chaired a symposium at the Second International Congress of Systematics and Evolutionary Biology, in Vancouver in July 1980, at which the evolutionary analysis of the human, mouse and rabbit hemoglobin mRNAs under selective constraints was presented by one of us (R.H.). This presentation was ignored by Fitch.

To find grounds on which to criticize REH theory, Fitch (1980) finds it necessary to 1) use almost a decade of hindsight, 2) use experimental data that did not even exist until five years after the theory was published in its original form, 3) ignore known recent developments in the theory, and 4) set up a straw man by comparing estimations made from nucleic acid sequence data with a maximum information content per codon of 6 bits, with those made from protein data with a maximum information content inherently less, 4.32 bits/residue.

In this communication it is our purpose to summarize the correct REH calculations for the beta hemoglobin mRNAs of human, mouse and rabbit both in the absence and presence of selective constraints; 2) show that the alternative evolutionary analysis of these mRNAs by Fitch underestimates total fixed mutations, 3) show that the model used by Fitch to test for completeness of the count of total base substitutions incorporates all the assumptions of REH theory, with genetic events equiprobable, that he so strongly criticizes and is in fact a minor variant of that theory, and 4) comment on certain other aspects of Fitch's analysis that bear on the future of more adequate evolutionary models.

REH Calculations

The fixation intensity μ_2 is the average number of total base substitutions per codon able to fix mutations during the divergence of two species. In Fitch's (1980) paper (his Table 11, for example) this parameter is designated as "rate", an imprecise usage as the latter term implies a time derivative. The estimated numerical value of this parameter is listed in Table 1 for the three beta

Table 1. Comparison of an evolutionary analysis of the beta hemoglobin mRNAs from human, mouse and rabbit by REH theory with genetic events equiprobable, and with genetic events constrained by structural constraints imposed by function, with the analysis given by Fitch

Comparison	$\hat{\mu}_2$[a]	$\hat{T}_2$[b]	$\hat{REH} = \hat{\mu}_2\hat{T}_2$[c]	P_1[d]	P_2	P_3
Human/Rabbit						
Equiprobable[e]	0.71±0.24	84±22	59±9	0.33	0.33	0.33
Constrained[f]	1.45	60	87	0.12	0.09	0.79
Fitch[g]	0.56	97	56	0.19	0.17	0.64
Human/Mouse						
Equiprobable	1.09±0.25	93±14	102±13	0.33	0.33	0.33
Constrained	1.31	88	113	0.25	0.18	0.57
Fitch	0.91	97	88	0.27	0.22	0.51
Mouse/Rabbit[h]						
Equiprobable	0.97±0.23	104±17	101±12	0.33	0.33	0.33
Constrained	1.28	120	154	0.19	0.14	0.67
Fitch	0.97	97	94	0.26	0.21	0.53

a The fixation intensity $\hat{\mu}_2$ is the estimated average number of base substitutions per variation

b $\hat{T}_2$ is the estimated number of varions (codons free to fix mutations during the divergence)

c $\hat{REH}$ is the estimated total base substitutions

d The proportion of the total base substitutions that have occurred at the i^{th} (i = 1, 2 or 3) position within the codons is designated p_i. This quantity is assumed in the REH model with genetic events equiprobable, and estimated from the data in the REH model with genetic events constrained by structural restrictions imposed by function. In Fitch's method, the values were estimated from the data in his Table 1 as described in footnote *g* below

e These values are essentially those in Table 4 from Holmquist (1980). They differ slightly from those values because of the recent experimental clarification (Lawn et al. 1980) of some ambiguities in the human β-globin ►

gene. The tabulated parameter values give agreement of the expected number of base differences OBD and of the estimated number of codons having exactly one, two and three differences with the observed number of such codons. Because of the model assumptions the differences are apportioned equally among the three nucleotide loci within codons

f From Tables 15, 16 and 17 in Holmquist et al. 1981. These are the values given at the Vancouver Symposium for the Second International Congress of Systematics and Evolutionary Biology in July, 1980. The parameter values listed give agreement of the expected number of base differences with the observed number at each of the three positions within the codon and of the estimated number of codons showing exactly one, two and three base differences with the observed number

g The total fixed mutations separating the human and rabbit beta hemoglobin mRNAs was calculated from the data in Table 1 of Fitch (1980). At the first, second, and third position within the codons the human mRNA differs from the nodal sequence by 3 + (2/3)3 = 5 base differences, 2 + (2/3)3 = 4 base differences, and 10 + + (2/3)7 = 14.67 differences. These total 23.67 differences (Fitch gives 24.6 for this number in the last sentence of the legend to his Table 1, presumably due to arithmetic error). At these same positions the rabbit mRNA differs from the nodal sequence by 5, 5 and 19.67 base differences. As these differences are additive the human and rabbit hemoglobin mRNAs differ from one another by 10, 9 and 34.33 base substitutions. To these values must be added the additional 6.5 substitutions due to multiple substitutions in the same position within the codon. We have distributed these 6.5 substitutions first among the human, rabbit and mouse lineages, as suggested by Fitch on page 160 of his paper, in proportion to the number of observable differences between the taxon at the tip and the nodal sequence at the fork, and second in proportion to the number of differences at each position within the codon for those cases in which all three nucleotides differ among the three taxa. The human/rabbit thus has an estimated 10.70, 9.70, and 35.99 number of substitutions at the three codon positions. From this the total number of base substitutions = 56.39 was calculated, which when rounded to the nearest integer is the value of 56 entered into the table. $P_1 = 10.7/56.39 = 0.19$; $P_2 = 9.7/56.39 = 0.17$ and $P_3 = 35.99/56.39 = 0.64$. By its definition a varion is a codon with one or more variable nucleotides. The number of such nucleotides in best agreement with the hemoglobin mRNA data is calculated in the lower half of page 161 by Fitch (1980) and has the value 292. On the average the number of variations will be about 1/3 this value, or 97.33, the rounded value of which is entered as T_2 in our table. The fixation intensity $\mu_2 = REH/T_2 = 56/97 = 0.56$. The human/mouse and mouse/rabbit values for the Fitch model were calculated similarly. From Fitch's (1980) Table 2, the total substitutions separating the human/rabbit, human/mouse and mouse/rabbit mRNAs are 53, 83 and 89 respectively, which differ negligibly from those calculated by us and given in the table as 56, 88 and 94. Either set of values suffices for our discussion in the text

h For the mouse/rabbit beta hemoglobin mRNA divergence there are 84 unchanged codons. 44 with a single base substitution, 14 with two base substitutions, and 4 due to three base substitutions. Because of an error in hand counting, values of 83, 45, 14 and 4 were reported in Table 1 in Holmquist (1980). The correct count thus gives 62 codons changed, 84 base differences, $\hat{\mu}_2 = 0.97 \pm 0.23$, $\hat{T}_2 = 104 \pm 17$, and $\hat{REH} = 101 \pm 12$. These differ insignificantly from the values reported in Table 4 of Holmquist (1980)

hemoglobin divergences discussed by Fitch. To make meaningful comparisons possible all estimations were made from gene or mRNA sequence data. Three estimation procedures were used. The first, REH theory under the assumption of the equiprobability of genetic events (Holmquist 1980) retains all the unrealistic assumptions of our 1972 (Holmquist et al. 1972) model which Fitch (1980) lists on page 196 of his paper and proceeds to criticize for the next nine pages. In Table 1 this procedure is designated the "Equiprobable" model. The second estimation procedure is the procedure given on page 259 and 260 of Holmquist and Pearl (1980) and includes allowance for nonrandom base compositions, transition and transversion frequencies, unequal usage of degenerate codons, a nonPoisson density of fixed mutations among codons, and a uneven density of fixed mutations among the three nucleotide positions within codons. This second model is designated "Constrained" in Table 1. The third estimation procedure is that resulting from the parsimonious reconstruction of the nodal beta hemoglobin mRNA sequence for the human, rabbit and mouse lineages as described by Fitch on pages 157 to 161 of his paper. This is designated the "Fitch" model in Table 1. Table 1 also lists for each method the estimated number of varions $\hat{T}_2$, the estimated total mutations fixed $\hat{REH}$, and the distribution of these mutations among the three positions within the codon. For all three methods the estimated total mutations fixed is at least as large as the number of observable differences between the gene or mRNA pair. For all three methods the estimated number of varions is at least as large as the number of codons experimentally observed to have varied. The values of μ_2, T_2 and REH estimated by the method of parsimony differ negligibly from those determined by the far simpler computations of REH theory for equiprobable genetic events. Of the three sets of values, Fitch's are the smallest. The values given in Table 1 for $\hat{\mu}_2$, $\hat{T}_2$ and $\hat{REH}$ bear minimal resemblance to the values calculated from the protein sequence data by our 1972 model. These values, illustrating the difference between estimates made from mRNA and protein data, given in Table 11 of Fitch (1980) had also been independently published by us (Table 5 in Holmquist 1980). Finally on page 161 of his paper Fitch points out the essential completeness of his count of total fixed mutations concluding "Since the two different

models employed here both fit the data well and give the same estimate of s, there will be no need to concern ourselves with more complicated models that allow for multiple changes at a single site and non-equiprobable base replacements." Fitch's faith in the completeness of his count is not justified. This is obvious when the REH values for total fixed mutations estimated from the model in which the non-random aspects of gene structure and evolution are considered in detail (Table 1, constrained model). By neglecting these complicating factors Fitch has underestimated his count, as well as the number of third-base substitutions, leading in turn to his overestimation of the proportion of fixed mutations occurring at the fist two positions within the codon. The agreement of Fitch's estimates of the fixation intensity, number of varions and total mutations fixed with those given by REH theory with equiprobable genetic events is because the differences between the nucleic acid sequences of the beta hemoglobin mRNAs for human, rabbit and mouse are comparatively small. For more distant divergences estimates given by Fitch's method would be more in error.

The Near Equivalence of the Method Used by Fitch for Testing for Completeness of the Count of Total Fixed Mutations and REH Theory with Genetic Events Equiprobable

In Table 1, the numerical closeness of the values for the evolutionary parameters μ_2, T_2 and REH as estimated by REH theory for equiprobale genetic events (Holmquist 1980) and by the procedure of Fitch (1980) suggests a formal equivalence between REH theory and the method Fitch used to test for completeness of count in the lower-half on page 61 of his article. A careful comparison of the procedure outlined by Jukes and Holmquist (1972) and Holmquist (1980) with that described by Fitch reveals the following correspondences: what Fitch designates s approximates $\hat{\text{REH}}$; what Fitch designates v approximates $3\hat{T}_2$; and what Fitch designates r approximates $\hat{\mu}_2/3$. In REH theory the parameter μ_2 is estimated from the ratio of multiply substituted codons to singly substituted codons; in Fitch's procedure the approximate (for not too distantly related sequences) inverse of this ratio is used to estimate r. Fitch's procedure includes all the assumptions of REH theory as originally published. There is one important difference. In REH theory experimental data are used to determine the ratio of multiply to singly hit codon sites; in Fitch's procedure his parameter r is not the ratio of two experimentally measured quantities, but of the *inferred estimate* of the number of singly substituted nucleotide sites to the *inferred estimate* of the number of multiply substituted nucleotide sites. These inferred estimates, because they are made from a parsimony network, are irreducibly erroneous (Kimura 1981; King 1980; Holmquist 1979). Felsenstein (1978) demonstrated that for some simple, but practically relevant evolutionary situations, the parsimony method is increasingly certain to give the *wrong* answer. Although Fitch refers to these inferred estimates as observed values, they are not; confusing calculated quantities with experimentally measured values does not advance any field. Karon (1979) has made a similar observation. It is difficult enough to explain the real sequence data without having to explain the properties of imaginary nodal sequences

Can Amino Acid Sequence Data Be Used To Estimate Total Nucleotide Replacements?

Published findings of the nucleotide sequences of messenger RNA show that there are wide variations in the proportion between the number of amino acid replacements and the number of nucleotide substitutions. This is shown in Table 2. Obviously, amino acid sequence data alone cannot be used to estimate total nucleotide substitutions.

Among proteins of known sequence, the histones are most higly conserved in evolution. Bovine and pea histones IV differ by only two amino acid replacements in 102 residues following a period of divergence of 1.5 billion years. However, many silent nucleotide substitutions have taken place in the mRNA's for the histones of two sea urchins (Table 2). At the other end of the scale are human growth hormones and humanchorionic somatomammotropin, which have diverged recently from a common ancestral gene to carry out separate functions, more than two-thirds of the nucleotide substitutions in this divergence bring about changes in amino acid sequence.

The proportion between nucleotide substitutions and amino acid replacements in the mRNA's of the globin family varies among different comparisons. The smallest ratio is in the comparison of mouse beta major and minor globins, which are believed to be the product of a recent gene duplication. The ratio increases in comparisons of beta globins of different species, or of alpha globins of different species, in which the number of nucleotide substitutions is in the range of 2.5 to 3 times as great as the number of amino acid replacements. As the evolutionary divergence widens, as shown in comparisons of alpha chains with beta chains, the number of amino acid replacements rises more rapidly than the number of total nucleotide substitutions, so that the ratio is in the range of 2.1 to 2.3. Interestingly enough this proportion, which has resulted from an evolutionary divergence lasting 600 million years, is about the same as in the case of homologous genes of two influenza viruses that diverged during the seven

Table 2. Examples of proportions between total nucleotide substitutions and amino acid replacements in evolutionary comparisons of mRNA's and corresponding proteins

System used for comparison	Total nucleotide substitutions (a)	Amino acid replacements (b)	(a):(b)
Sea urchin histones[a]	161	11	15
Hormones[b]			
HGH, HCS[c]	49	32	1.5
HGH, RGH[c]	162	76	2.1
HCS, RGH	160	84	2.0
Globins			
Mouse β major and minor[d]	17	9	1.9
Human β, mouse β[d,e]	54	20	2.7
Human β, chicken β[e,f]	76	29	2.6
Rabbit α, mouse α[g]	58	18	3.2
Human α, human β[e,h]	120	54	2.2
Human α, rabbit β[h,i,j]	125	55	2.3
Human α, mouse β[d,h]	120	54	2.2
Human α, chicken β[h,f]	120	58	2.1
Preproinsulin			
Human, rat[k]	63	19	3.3
Hemagglutinins			
Influenza viruses[l]	63	28	2.2
øX 174 and G4 genes[m]			
Read in phase	1,660	550	3.0
Read by frameshift	178	116	1.5

a Grunstein M, Grunstein J (1977) The histone H4 gene of *Strongylocentrotus purpuratus:* DNA and mRNA sequences at the 5' end. Cold Spring Harbor Symp Quant Biol 42:1083–1092
Schaffner W, Kunz G, Daetwyler H, Telford J, Smith HO, Birnstiel ML (1978) Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing. Cell 14:655–671
Sures I, Lowry J, Kedes LH (1978) The DNA sequence of sea urchin (*S. purpuratus*) H2A, H2B and H3 histone coding and spacer regions. Cell 15:1033–1044

b Martial JA, Hallewell RA, Baxter JD, Goodman HM (1979) Human growth hormone: Complementary DNA cloning and expression in bacteria. Science 205:602–606
Seeburg PH, Shine J, Martial JA, Baxter JB, Goodman HM (1977) Nucleotide sequence and amplification in bacteria of structural gene for rat growth hormone. Nature (London) 270:486–494
Seeburg PH, Shine J, Martial JA, Ullrich A, Baxter JD, Goodman HM (1977) Nucleotide sequence of part of the gene for human chorionic somatomammotropin: Purification of DNA complementary to predominant mRNA species. Cell 12:157–165
Shine J, Seeburg PH, Martial JA, Baxter JD, Goodman HM (1977) Construction and analysis of recombinant DNA for human chorionic somatomammotropin. Nature (London) 270:494–499

c HGH = human growth hormone; HCS = human chorionic somatomammotropin; R = rat

d Konkel DA, Maizel JV, Jr, Leder P (1979) The evolution and sequence comparison of two recently diverged mouse chromosomal β-globin genes. Cell 18:865–873
Konkel DA, Tilghman SM, Leder P (1978) The sequence of the chromosomal mouse β-globin major gene: Homologies in capping, splicing and poly(A) sites. Cell 15:1125–1132

e Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechi AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ (1980) The structure and evolution of the human β-globin gene family. Cell 21:653–668
Forget BG, Cavallesco C, DeRiel JK, Spritz RA, Choudary PV, Wilson JT, Wilson LB, Reddy VB, Weissman SM (1979) Structure of the human globin genes. In: Axel R, Maniatis R, Fox CF (eds) Eucaryotic gene regulation. ICN-UCLA Symposium on Molecular and Cellular Biology XIV. Academic Press, New York, pp 367–381
Kafatos FC, Efstratiadis A, Forget BG, Weissman SM (1977) Molecular evolution of human and rabbit β-globin mRNAs. Proc Natl Acad Sci USA 74:5618–5622
Lawn RM, Efstratiadis A, O'Connell C, Maniatis T (1980) The nucleotide sequence of the human β-globin gene. Cell 21:647–651

f Richards RI, Shine J, Ullrich A, Wells JRE, Goodman HW (1979) Molecular cloning and sequence analysis of adult chicken β-globin cDNA. Nucl Acids Res 7:1147–1162

g Heindell HC, Liu A, Paddock GV, Studnicka GM, Salser WA (1978) The primary sequence of rabbit α-globin mRNA. Cell 15: 43–54

Nishioka Y, Leder P (1979) The complete sequence of a chromosomal mouse α-globin gene reveals elements conserved throughout vertebrate evolution. Cell 18:875–882

h Michelson AM, Orkin SH (1980) The 3' untranslated regions of the duplicated human α-globin genes are unexpectedly divergent. Cell 22:371–377

Proudfoot NJ, Maniatis T (1980) The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. Cell 21:537–544

i Hardison RC, Butler ET, III, Lacy E, Maniatis T, Rosenthal N, Efstratiadis A (1979) The structure and transcription of four linked rabbit β-like globin genes. Cell 18:1285–1297

j Efstratiadis A, Kafatos FC, Maniatis T (1977) The primary structure of rabbit β-globin mRNA as determined from cloned DNA. Cell 10:571–585

k Bell GI, Swain WF, Pictet R, Cordell B, Goodman HM, Rutter WJ (1979) Nucleotide sequence of a cDNA clone encoding human preproinsulin. Nature 282:525–527

Lomedico P, Rosenthal N, Efstratiadis A, Gilbert W, Kolodner R, Tizard R (1979) The structure and evolution of the two nonallelic rat preproinsulin genes. Cell 18:545–558

l Verhoeyen M, Fang R, Min Jou W, Devos R, Huylebroeck D, Saman E, Fiers W (1980) Antigenic drift between the haemagglutinin of the Hong Kong influenza strains A/Aichi/2/68 and A/Victoria/3/75. Nature (London) 286:771–775

m Godson GN, Barrell BG, Staden R, Fiddes JC (1978) Nucleotide sequence of bacteriophage G4 DNA. Nature (London) 276:236–247

Sanger F, Air G, Barrell B, Brown N, Coulson A, Fiddes C, Hutchinson C, III, Slocombe P, Smith M (1977) Nucleotide sequence of bacteriophage φX 174. Nature 265:687–695

year period between 1968 and 1975 (Verhoeyen et al. 1980).

The genes of the coliophages φX 174 and G4 may be placed in two groups. The first group consists of genes A, C, D, F, G, H and J. The second set consists of genes B, E and K. These are read as frame shifts in genes A, C and D, but have a much lower percentage of silent third-base codon changes, because such changes would often be rejected as causing unacceptable amino acid replacements in genes A, C, and D. The comparison in Table 2 shows that the ratio between nucleotide substitutions and amino acid replacements is twice as high in the case of the first set of genes as in the second.

The unambiguous conclusion that emerges from Table 2, and particularly from the influenza virus comparisons, is that the ratio between amino acid replacements and nucleotide substitutions is *not* strictly time dependent, nor even approximately so, but also strongly depends upon the extent to which amino acid sequences are conserved in evolution.

Fitch attempts to force the 1972 model of Jukes and Holmquist (1972) to explain data that it was never intended to explain. He puts protein data in and expects to get a gene out. That he fails is no test of any method, and does not consitute valid criticism of our own.

Conceivably, by incorporation knowledge gained from a study of nucleic acid sequences into theoretical models for the evolutionary divergence of proteins one might be able to use amino acid sequences to make useful inferences about total substitutions fixed in the genes coding for those sequences, the number of varions able to accept substitutions, and the fixation intensity. This approach can work for distant divergences as shown by Holmquist and Pearl (1980; see their Table 21) for rabbit alpha and beta hemoglobin. But those authors also pointed out that for less distant evolutionary comparisons one cannot expect good agreement between evolutionary parameters estimated from protein data on the one hand, and nucleic acid data on the other, because for less distant divergences it is impossible to obtain an accurate estimate of the number of variable codons T_2 from amino acid sequence data. Indeed, in the absence of a reasonable estimate of T_2 it is not possible for any stochastic model to make accurate evolutionary estimates. Simply assuming that all the codons compared are able to fix mutations (as is often done) is not an acceptable substitute for a sound estimate of T_2. Such estimates can only be obtained from nucleic acid sequence data.

Uncertainty in Evolutionary Estimates

Figure 3 in Fitch (1980) is a genuine contribution toward furthering the understanding of the accuracy with which evolutionary estimates can be made. The great virtue of Fig. 3 is that the plot is 100% experimental: both the ordinate and abscissa are determined by experimental measurements alone, undistorted by plausible (or implausible) theoretical models of evolutionay divergence. The scatter in Fig. 3 is thus real and not the result of a particular model's assumptions. Figure 3 tells un in quite unambiquous terms that the same observable result (amino acid sequence differences in Fig. 3, but the argument applies equally well to nucleic acid sequences) can be reached by many different evolutionary pathways. It is a fundamental limitation on our knowledge that we can in principle never know which of these pathways in fact occured. Fitch states (p. 178) that 'Thus REH is an estimate with a much

greater variance than more conventional corrections for multiple substitutions at the same site.' The variance of a Poisson process is equal to its mean; but if the true process is not Poisson, and there is much evidence that is not (Fitch abd Markowitz, 1970; Uzzell and Corbin, 1971; Holmquist et al. 1981) then that equality is biologically irrelevant. We do not agree with Fitch (p. 204) that an inaccurate (badly biased) estimator precisely known is acceptable. We have explicitly calculated the variance for the Equiprobable model in Table 1 and the results are shown as ± the square root of this quantity in that table. The relative precision ($\sigma_{\hat{REH}}/\hat{REH}$) in $\hat{REH}$ is (Table 1) about 0.13 for the comparisons shown.[1] That is the stochastic variation about the estimated mean is about 13% of that mean. This precision continues to hold even for the distant rabbit *a*/rabbit *β* mRNA comparison for which the variation is about 12% (not shown). The variance in REH is thus not 'much greater' (Fitch 1980, p. 178) than the more conventional corrections for multiple substitutions at the same site, but is comparable. For example, if the total substitutions are calculated from the usual $-(3L/4)\ln[1 - 4\lambda/3)]$, L being the total nucleotide pairs compared in the sequence, with λ the average number of base differences per nucleotide site, then for the human/rabbit *β* hemoglobin mRNAs, the stochastic variation is 15%. The necesarry formulas for verifying this statement are in Kimura and Ohta (1972).

The stochastic variation in $\hat{T}_2$ is somewhat greater (Table 1) than in $\hat{REH}$ but still is within accepatable limits considering the difficulty in estimating this parameter adequately. The stochastic variation in $\hat{\mu}2$ is the largest of the three. The magnitude of this variation is about 25% for the comparisons in Table 1.

In Table 1, we have not given estimates for the variation in the estimates made from the constrained model in which many selective factores have been taken into account. The computation is difficult and no results are yet available. However, on the basis that for independent Bernoulli trials with variable probabilities, the *lack* of uniformity in these probabilities *decreases* the magnitude of chance variations (for a simple example, the bionominal variance npq is maximal when p = q = 1/2), one might anticipate that the variances in estimates made from models in which selective constraints are considered might be less or at least not grossly greater than for the equiprobable model in Table 1 (allowing for the fact that in real biological systems, genetic events are not always independent of one another).

Estimates resulting from Fitch's model are also given without their variances. We are somewhat at a loss as to how one might go about calculating a meaningful variance for estimates made from imaginary ancestral sequences.

From the above considerations excessively large variations do not appear to be a major problem for any of the evolutionary models now available and there is no factual basis for believing Fitch's assessment (Fitch 1980, p. 202) that corrections which move our estimates closed to the true distance must do so at the expense of an increased variance sufficient to compromise the value of the improvement.

[1]

$$\sigma_{\hat{\mu}_2} \cong \frac{d\mu_2}{dr}\frac{(1+r)}{(1+n_1)}\sigma_{n_1}\left\{1 - \frac{P(0)n(1)}{(1+r)\sigma^2_{n_1}}\right\}^{1/2};$$

$$\sigma_{\hat{T}_2} \cong \frac{-T_2^2}{3\sum_{k=1} n_k}\frac{dP(0)(1+r)}{dr\,(1+n_1)}\sigma_{n_1} \times \left\{1 - \frac{b^2-1}{b^2}\frac{P(0)n(1)}{(1+r)\sigma^2_{n_1}}\right\}^{1/2};$$

$$\sigma_{\hat{REH}} \cong \hat{REH}\left\{\epsilon^2_{\hat{\mu}} + \epsilon^2_{\hat{T}_2} - \frac{2ab}{\hat{\mu}_2 \sum_{k=1} n_k}(1+r)^2\sigma^2_{n_1} \times \left[1 - \frac{P(0)n(1)}{(1+r)\sigma^2_{n_1}}\right]\right\}^{1/2}$$

Here $\epsilon_\theta \equiv \sigma_\theta/\theta$, $a \equiv \frac{1}{(1+n_1)}\frac{d\mu_2}{dr}$,

$$b \equiv \frac{-T_2}{(1+n_1)}\frac{dP(0)}{dr}, \text{ and } \sigma_{n_1} = n(1)\left[1 - \frac{n(1)}{T_2}\right].$$

In these expressions n(i) and n_i are the expected and observed number of codons with exactly i (i = 1, 2, or 3) base substitutions, respectively, and r = [n(2) + (3)]/n(1). The two derivatives can be found by finite differences from Table 3 in Holmquist, 1980. P(0) is the proportion of unsubstituted varions and is given as a function of r in that same table. In practice the n(i) are replaced by their sample values, and r, μ_2 and T_2 are replaced by their estimators $\hat{r} = (n_2 + n_3)/(1 + n_1)$, $\hat{\mu}_2$ and $\hat{T}_2$. Derivation of these formulas, though straightforward, is tedious and will be published separately.

Metrics, Additive Trees and Negative Internodal Distances

Friday (1980) points out: 'If data are subject to error, then, this may result in non-additive (or even non-metric) distances ....'. The statement also applies to transforms of that data. It would be presumptuous to believe that present models perfectly capture the details of evolution accurately. Even if they did, the models state relationships between expectations values, whereas the

data analyzed by those models are the result of a unique experiment by nature. Thus when Fitch (1980) applies an additive algorithm (Fitch and Margoliash 1967) to a non-additive set of REH distances (his Fig. 4) and comes up with negative branch lengths it is no surprise.[2] The negative branch length found by Fitch for cytochrome *c* in the lineage leading to pigeon is due to the fact that the REH estimates of total mutations fixed during the pig/pigeon and neurospora/pigeon divergences are 21 and 128, respectively, which sum to 149. This is less than the 199 estimated fixed mutations of the pig/neurospora divergence and hence a violation of the triangle inequality. These two numbers, 149 and 199, do not differ significantly from that expected from the sampling errors in the individual distance estimates. The sum of the estimated total fixed mutations for the pig/pigeon and pig/*Candida* divergences is 265, less than the 285 estimated fixations in the pigeon/*Candida* divergence and again a violation of the triangle inequality, and again not significant. Fitch's negative branch lengths thus arise from his trying to impose upon the estimates a precision not obtainable from the data.

Fitch's (1980) criticism does, however, bring into focus the more important general question as to how non-additive matrix data sets are to be handled. The alogrithm of Fitch and Margoliash (1967) is inappropriate for such sets, but appropriate algorithms exist. Sattath and Tversky (1977) have shown how non-additive similarity distance matrices may be transformed into additive (hence metric) matrices in such a manner that the resulting tree or network constructed from them will have all nonnegative branch lengths whose sums deviate minimally from those of the original similarity matrix. If we use the sum of the squared deviations as the minimization criterion, the correct solutions for the lineages given in Fitch's Fig. 4 are for the left-hand network zero mutations fixed in the pigeon lineage, 38 fixed in the pig lineage, and 145 fixed in the neurospora lineage. For the right-hand tree in Fig. 4, the solution is 251 mutations fixed in the *Candida* lineage, 28 fixed in the pigeon lineage, and none fixed in the pig lineage. The reason for the different branch lengths in the two reconstructions becomes obvious from the amino acid sequence data: the pigeon cytochrome *c* sequence differs from that of *Candida krusei* and neurospora by about the same number of amino acid replacements of the minimal one base type, 22 and 24, respectively; but the amino acid replacements requiring at least two base replacements in the corresponding codons are almost twice as frequent between pigeon and *Candida krusei* cytochromes *c* as between pigeon and neurospora, 20 and 12, respectively, Clearly *Candida* is more distantly related to the pigeon than is neurospora, as judged by the amino acid sequences of their cytochromes *c*. It is natural that this experimental difference be reflected in the REH values and the phylogenetic reconstruction. The anomaly is in the data chosen for analysis by Fitch. The REH values given in Fitch's (1980) Fig. 4 are biologically realistic interpretation of the *protein* data. His inappropriate analysis of the values, his failure to consider the known biology, and his argument by testimonial, do not constitute adequate criticism of our methods.

Even in the absence of error, non-divergent change also causes non-additive distances. Relying on an additive best-fit appoach, whether that of Fitch and Margoliash (1967), Sattath and Tversky (1977), or other (Shepard 1980), to corroborate the mainly divergent nature of one's data can be seriously misleading (Friday 1980).

Silent Substitutions

Given the 113 base differences between human, rabbit and mouse beta hemoglobin and the nodal sequence (see page 165 in Fitch 1980) Fitch calculates that we would expect 84.8 of these to be nonsilent and 28.2 of these to be silent based on the procedure outlined in his Table 4 in which the actual codon count is used to enumerate the number of ways that beta hemoglobin can mutate. The observed values reported by Fitch are 52.1 and 60.8, which as he points out are in obvious disagreement with expectation. We conclude the method he used to compute Table 4 is inadequate. REH theory on the other hand predicts, for a total of 113 base differences, that 51.7 of these should be nonsilent and 61.3 should be silent based on actual codon usage (Table 4, Holmquist and Pearl 1980). Even if the actual codon usage is not known REH theory predicts (Table 5, Holmquist and Pearl) 50.7 and 62.3 nonsilent and silent substitutions respectively. Either set of values explains the observations well. (The 'observed' values given by Fitch are actually values inferred by the method of parsimony. As in the present instance we are concerned with the relative numbers of two type of substitutions, the undercount of total mutations fixed by parsimony does not alter our conclusions).

On page 198 of his article Fitch (1980) states '...that the silent substitutions do not occur in the varions...' This is incorrect. By definition a varion is a codon which during the divergence of a sequence pair has one or more variable nucleotides. From Table 1 of the present paper

2 In Fitch (1980) the matrix of REH values given in his Figure 4 is inappropriate for evolutionary analysis. The REH values for pig/pigeon is for 104 compared residues, for *Candida*/neurospora 107 compared residues, and for the remaining pairs for 103 compared residues. Correcting Fitch's matrix for this error the pig/pigeon and *Candida*/neurospora distances should be 21 and 135, respecitvely. This does not change the tenor of Fitch's argument.

60 – 80 % of all base substitutions have occured at the third position within the varions. Most of these are silent.

The Expected Number of the Different Types of Base Replacements

The expected fraction of observable base interchanges of a given type $i \leftrightarrow j$ (i = A, C, G, U; $i \neq j$), without regard to direction, is calculated by Fitch (1980) as $(N_i + N_j)/3N$, where N_i/N is the mole fraction of the base i in the sequence. This expression is valid only if each of the four bases has an equal probability of mutating to and being fixed as one of the remaining three. This is almost never the case in practice. More seriously, if it were, asymptotically N_i would equal N_j, in contradiction to observed base compostitions $N_i \neq N_j$.

More fundamentally, it is in principle *not* possible to calculate unique base interchange frequencies from the base composition alone. Even if the base composition is stable during evolutionary divergence, which it may not be, there are an infinitude of base interchange probabilities consistent with that stability. Among these, that set of twelve probabilities which is minmally prejudiced in the sense of admitting the fewest unjustified assumptions can be found as described by Holmquist and Cimino (1980) (see also Jaynes 1979). This minimally prejudiced set, for beta hemoglobin genes is not symmetrical, that is the probability for the interchange $i \rightarrow j$ is not equal to that for the interchange $j \rightarrow i$ (Table 5 in Holmquist and Cimino 1980). This method has the virtue of internal consistency, and the assumption of the base compostional stability can be replaced with a more realistic one, or ones, as such become known experimentally. The REH method does not assume each codon is used equifrequently (Compare Holmquist and Pearl 1980, Section 3.2 with the statements on page 199 in Fitch 1980).

On the Concordance of REH and Augmented Distance Values

In 1976 Moore et al. (1976) and Holmquist et al. (1976) reported that the REH values obtained from stochastic theory using the 1972 REH model of Jukes and Holmquist and the augmented maximum parsimony distance values of Moore (1977) and his colleagues correlated with a slope near unity for 641 pairwise comparisons of cytochrome *c* sequences and over a thousand inter- and intrafamily comparisons of alpha hemoglobin, beta hemoglobin, myglobin and leghemoglobins. These reports were important because for the first time, estimates obtained by parsimony were as large as those predicted four years earlier by the REH method. We were careful to qualify our observations with the comment (Holmquist 1976) that this concordance 'is not meant to imply that the present values of REH and AD are absolute final estimates of genetic divergence. AD will increase as additional experimental sequence information becomes available, and REH will increase when full account is taken of the differential ability of each nucleotide site to fix mutations. If the explanation of the convergence given ... is correct, then the limiting [3] values of REH and AD will still be such that their ratio is, on the average, unity.' The augmented distance values AD have increased (Baba et al. 1981) as have the REH values (Holmquist and Pearl 1980); Holmquist et al. 1981). What is critical in the 1976 papers (Moore et al. 1976; Holmquist et al. 1976) on REH theory and the augmented distances was not the coincidence of the numerical values of the two estimates, each a function of the state of the art at that time (and in that specific sense superficial), but the fact that this coincidence, despite the limitations of both models, led us to the generalization that there was no fundamental dichotomy between parsimonious and stochastic methodology and that for adequate models the concordance should continue to hold, not for superficial reasons, but as a basic principle. Thus, if the AD estimates are made from mRNA or gene nucleic acid sequence data and dense data sets (so that base replacements at the third position within the codon are detected), and if the REH estimates are made on assumptions that model evolution appropriately, the slope between the augmented distances and the REH distances should still, on the average be unity.

REH and PAMS

When all codons in a protein, or all nucleotide sites in a nucleic acid, are able to fix mutations the REH and PAM values are in approximate agreement (Holmquist 1972b, Table 1 and Fig. 1, for amino acid sequences; Holmquist 1973, Table 1 and Fig. 1, for tRNA sequences). This agreement persists even in the face of the most recent published data (Dayhoff et al. 1978, Table 23). When there are restrictions on codon mutability the Jukes and Holmquist (1972) REH values are much larger than the PAM values in most cases (Compare Dayhoff et al. 1972, Figs. 2–1 with Table 5 in Jukes and Holmquist 1972). The PAM values are often less than the unaugmented maximum parsimony distances (Compare Dayhoff et al. 1972, Fig. 2–1, to Fig. 2 in Baba et al. 1981) even when

3 By limiting is meant as REH theory includes in its formalism those functional and structural constraints on gene and protein structure that limit available evolutionary pathways, and as the data set from which the augmented distances are calculated becomes maximally dense (Holmquist 1978b).

both are estimated from the same protein data. Fitch's (1980) discussion of these measures on page 200 of his article is thus incorrect, perpetuating the error (Holmquist, 1978a) in his earlier review (Fitch 1973), and obsolete.

Variations and Covarions

Fitch (1980, pages 200 and 201) maintains that the number of varions, for equal time intervals, should be equal, at least within the estimation error. This is wrong because it ignores the fact that different numbers of codons may be free to fix mutations in different lineages. Using the estimates in Table 1 for the equiprobable model as illustration, an average of 84 ± 22, 93 ± 14 and 104 ± 17 varions separate the human/rabbit, human /mouse and rabbit/mouse β hemoglobin mRNA sequences (these numbers of varions are between the contemporary sequences, not as misstated by Fitch, between the contemporary sequences and their common ancestors). This is compatible with there being 56 varions during the divergence of mouse and the common ancestor of rabbits and humans, and with the existence, since the divergence of rabbit and humans from their common ancestor, of 36 varions in the human lineage and 48 in the rabbit lineage. There are an estimated 72 base subsitutions among the 56 varions during the divergence of mouse and the common ancestor of rabbits and humans; an estimated 29 base substitutions among the 48 varions in the lineage leading to the rabbit; and an estimated 30 base substitutions among the 36 varions leading to humans. These values are consistent with both the experimentally observed number of base differences between the mRNAs of the extant species pairs and the REH estimates of total base substitutions separating those pairs (Table 1); 59 ± 9 (human/rabbit), 102 ± 13 (human/mouse) and 101 ± 12 (mouse/rabbit). This analysis proves that Fitch errs in his basic logic. There is nothing in REH theory requiring different lineages to have the same number of varions. Because the number of varions cannot be estimated accurately from protein data, analyses of the above type are best limited to mRNA or DNA sequence data.

The discussion of the preceding paragraph has an important biological qualitative implication; the total number of base substitutions leading to two contemporary sequences from a common ancestor can differ for two reasons. The first, a change in the fixation intensity, the average number of base replacements per varion; the second, a change in the number of varions, that is the number of codons able to fix mutations in the two lineages. In the published literature, most of the focus has been on the former, primarily because it is easier to estimate. But an accurate estimate of total mutations fixed cannot be obtained without a knowledge of the latter.

An estimate of the number of covarions is not an acceptable substitute for an estimate of the number of varions, because the particular codon sites that form the set of covarions change as evolution proceeds. Further, Ratner et al. (1977) have shown that Fitch's estimates for the number of covarions for alpha and beta hemoglobin are incompatible with the experimental sequence data (Compare Fitch 1980, p. 160).

When the number of varions is estimated from the best data (gene or mRNA sequences) using the most adequate REH theory, the number of varions equals or exceeds the number of codons experimentally observed to have varied. As one example, from Table 1 of the present paper, the number of varions estimated for the rabbit and human beta hemoglobin mRNA divergence is 64 (not the 29.2 given in Table 11 and on page 201 of Fitch's paper). This is greater than the 42 codons changed, and quite within the bounds of biological reason.

Nonlinearities in the Estimation of Total Number of Substitutions

In a final section (p. 205) Fitch (1980) discusses the attempts by various authors (Kimura, Ohta, Dayhoff, Goodmann, ourselves) to obtain an adequate correction leading to a valid estimate of the total number of base substitutions or total amino acid replacements, dismissing these efforts with: 'But all of these concerns may be unimportant. Unless there is some bias in these procedures that causes the error in the estimate of total number of substitutions to be non-linear with respect to the true number, the relative times of divergence will be reasonably estimated regardless of how far the actual estimate is from the truth.' This shows a lack of concern about agreement between calculation and fact. Those who have studied the problem carefully agree that the correction for the more distant relationships is greater, nonproportionally so, than for the nearer relationships. The theoretical models, despite their various shortcomings, are in accord on this point, and computer simulations uniformly confirm this aspect of theory. Examination of published models shows that as the asssumptions going into the models have become more realistic, the estimates of total mutations fixed for the distant divergences have become progressively larger, both in the absolute sense, and *relative to* those estimates for less distant relationships. The bias Fitch would like to ignore exists, and has become progressively more important as our understanding of macromolecular evolution has advanced.

Inaccuracies of Fact

There are numerous errors of simple fact in Fitch's analysis. Here we point out only the more important. In

Table 7, codon position 40 is occupied by an arginine in human, mouse and rabbit mRNA. The number of minimal base differences for this position is thus 0, not 1. In both rabbit and mouse beta hemoglobin mRNA codon position 112 is occupied by isoleucine. Again the number of minimal base differences is zero. These errors progagate all the way through to Table 11 in Fitch's analysis. In citing history Fitch neglects to mention that the work of Jukes (1963) and Doolittle and Blombäck (1964) on the use of parsimony in a phylogenetic context foreshadowed his own by several years.

On page 203 Fitch states that there is a 'rule' in REH theory that the number of one-base differences must be at least as large as the number of two and three base changes. There is no such rule. In REH theory with genetic events equiprobable all REH theory states is that for proteins the *expectation value* of the number of *minimal one-base type amino acid replacements* is greater than 0.956 times the sum of the *expectation values* of the number of *minimal two- and three-base type amino acid replacements*. The corresponding statement for genes, again under the assumption of the equiprobability of genetic events, is (Holmquist 1980, Table 3) that the expectation value of the number of codons showing a single base replacement is greater than 1/6 times the sum of the expected number of codons showing two or three base replacements. For the Tables in Holmquist and Pearl (1980) in which genetic events are nonequiprobable the correpsonding values are 0.992 for proteins and 1/5.48 for genes or mRNAs if the distribution of fixed mutations among the variable codon sites is Poisson.

The analogy (Fitch 1980, p 203) between our own method and the Jukes and Cantor (1969)/Kimura and Ohta (1972) formula is accurate. If the number of varions is small, then a relatively few number of base substitutions can randomize *that portion* of the sequence irrespective of the homology existing between conserved residues elsewhere in the sequence. REH theory will alert one to such situations by indicating a large variance in the estimate (see footnote 1, this paper). Parsimony ignores such situations, On that same page Fitch states that our model has too many varions. His statement is false; properly calculated the number of varions estimated is always at least as large as the number of codons changed, and never greater than the number of homologous codon sites compared, Fitch states that if their are no multiple base differences, our model predicts that the substitution rate is zero, and that the model can give infinite substitution rates. Both statements are incorrect (see discussion of Equation 21 in Holmquist (1978a).

Discussion

The first comparisons of evolutionary divergence in sequences of amino acids and proteins were made with hemoglobins and cytochrome *c*. An early observation (Zuckerkandl and Schroeder 1961) was that human and gorilla hemoglobins differed only by one amino acid substitution in each chain as compared with 18 differences between the amino acid sequences of human and horse alpha globin chains.

Earlier yet, Tuppy (1958) compared amino acid sequences of a peptide obtained from cytochromes *c* of various species. The peptide was 11 residues in length, and in yeast contained four differences from the corresponding peptide in silkworm which in turn differed by one residue from that obtained from three mammals —cow, horse and pig. The concept of a molecular evolutionary clock arose from such simple observations.

The next step forward in comparing amino acid sequences was made possible by preliminary findings with the genetic code, even before the sequences of nucleotides in codons were known. These findings showed that miuational changes in proteins corresponded to single nucleotide substitutions in the codes for the corresponding amino acids. Many such amino acid mutations were discovered in hemoglobins and tobacco mosaic virus coat protein. In contrast, some amino acid substitutions found in evolutionary comparisons of hemoglobins and cytochromes *c* were unexplainable by single nucleotide substitutions, so that it was necessary to conclude that at least two nucleotides had been changed. For example, four of the eighteen amino acid differences between human and horse alpha hemoglobins correspond to at least two base differences per codon. It was thus possible to classify amino acid differences between homologous proteins in terms of minimal one, two and three base differences in the corresponding codons (Jukes 1963). The procedure begged the question of silent substitutions in their codon positions, and also ignored the obvious possibility that replacements involving arginine, leucine and serine might involve more substitutions than could be attributed to the minimum base difference interpretation.

The next step forward was our introduction of the 'random evolutionary hits' REH method of comparison in 1972 (Holmquist et al. 1972; Jukes and Holmquist 1972). This was introduced to deal with the problems of multiple hits at the same site, revertants, multiply hit codons, and silent changes in codons – all of which must occur among the evolutionary changes that take place during the divergence of two genes from a common ancestor – within the context of the genetic code findings of the preceding paragraph. Obviously it could not be expected that these assumptions were correct, but until it was possible to disprove them, we let them stand as a first approximation. We apologize to Professor Fitch for not having had the forsesight in 1972 that hindsight has given him in 1980.

It had for some time been obvious that the arginine content of proteins was far less, on the average, than

was to be expected from the percentage of arginine codons in the amino acid code (King and Jukes 1969). We knew by 1975 (Jukes et al. 1975) that other amino acids, especially lysine, aspartic acid and glutamic acid, had an average distribution in proteins differing from that expected by their respective number of codons. When sequences of mRNAs appeared, it was immediately obvious that silent nucleotide changes in evolution were more frequent than substitutions that produced amino acid changes.

Certainly, further adjustments will be necessary to cope with matters such as the preferential use of some codons for a single amino acid, evolutionary pressures, as for example the Treffers mutator gene that favors transversions over transitions, and selection against CpG doublets. Some of the more important constraints on evolution, and their relevance to evolutionary estimation, have been considered in Holmquist and Pearl (1980).

In contrast, methods of parsimony, grinding out the same mathematical monotone, have made no provision for the newly perceived realities of molecular evolution.

Conclusions

We do not see the purpose in Fitch's (1980) criticism. To the extent it finds that evolutionary estimates made from nucleic acid sequences differ from those made from protein sequence, it simply confirms our own published findings. In finding fault with some of the assumptions used earlier models of molecular evolution it is trivial. To the extent Fitch uses those very assumptions in assessing the completeness of count in his own model, the criticism is less than forthright. And in implying that the variance in our REH estimates of total base substitutions is unduly large, it is wrong.

Our approach has consistently been to incorporate the constraints upon evolutionary pathways delineated by experiment into the formalism of REH theory so that an understanding can be gained as to how these restricted pathways affect our ability to make valid evolutionary inferences. By taking this step-wise approach, one can avoid the worst of those pitfalls that result from imposing a priori prejudices into a theoretical framework. The only available macromolecular evolutionary knowledge is in the pattern of changes observed in nucleic acids and proteins isolated from living organisms. Those changes are statistical in nature even through individually they may be due to a mulitplicity of fully or partly determined evolutionary events.

The concepts we introduced in 1972 to permit a more accurate evolutionary analysis of macromolecular data remain valid today and can be implemented in practice (Holmquist and Pearl 1980; Holmquist et al. 1981). We do not defend the assumptions of the 1972 model (Holmquist et al. 1972; Jukes and Holmquist 1972) beyond the historical context that those assumptions were consistent with the experimental data then available.

Until the method of parsimony frees itself from its present ad hoc abiological algorithmic nature, its results pose a precision that is illusory. (Felsenstein 1978; Homquist 1979). We have tried to present a reasonable alternative to this situation by allowing the structure of the data to determine the structure of the theory rather than imposing upon that data an analysis based upon a nonfalsifiable hypothesis of parsimony. Goodman and his coworkers (Moore et al. 1976; Moore 1977; Baba et al. 1981), despite the criticism of their efforts by Fitch (1980), have done and are doing much of the hard work to develop methods, such as the augmentation algorithm, to make the parsimony method reflect biological information more accurately. Cavender (1978) has made a good start on the very difficult problem of determining confidence limits for taxonomic trees deduced by the method of parsimony.

We should like to end this paper with a caution from the Argentine physicist/philospher Mario Bunge: 'Ockham's razor, like all razors, should be handled with care to avoid beheading in the attempt to shave off some pilosities. In science, as in the barber shop, better alive and bearded than cleanly shaven and dead.'[4]

Acknowledgements. This work was supported by NSF award PCM-76-18627 'Protein and Nucleic Acid Evolution' and by NASA grant NGR 05-003-460 'The chemistry of Living Systems'.

References

Baba ML, Darga L, Goodman M, Czelusniak J (1981) Evolution of cytochrome *c* investigated by the maximum parsimony method. J Mol Evol 17: 197–213

Cavender JA (1978) Taxonomy with confidence. Math Bios 40:271–280

Dayhoff M, Park CM, McLaughlin PJ (1972) Building a phylogenetic tree: Cytochrome *c*. In: Dayhoff M (ed) Atlas of protein sequence and structure, vol. 5. National Biomedical Research Fdn., Georgetown University Medical Center, Washington, DC, p 8

Dayhoff M, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff M (ed) Atlas of protein sequence and structure, vol. 5, suppl. 3. National Biomedical Research Fdn., Georgetown University Medical Center, Washington, DC, p 351

Doolittle RF, Blombäck B (1964) Amino-acid sequnce investigations of fibrinopeptides from various mammals: Evolutionary implications. Nature 202:147-152

4 We are indebted to Dr. Jorge Crisci, Division Plantas Vasculares, Museo de la Plata, for calling our attention to the works of Professor Bunge through his talk 'Parsimony in evolutionary theory: law or methodological prescription?'

Felsenstein J (1978) Cases in which parsimony or compatiblity methods will be positively misleading. Syst Zool 27:401–409

Fitch W (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. Syst Zool 20: 406–416

Fitch W (1973) Aspects of molecular evolution. Ann Rev Genet 7:343–380

Fitch W (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: Comparison of several methods and three beta hemoglobin mRNAs. J Mol Evol 16:153–209

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

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

Friday A (1980) The status of immunological distance data in the construction of phylogenetic classifications: A critique. In: Bisby FA, Vaughan JG, Wright CA (eds) Chemosystematics: Principles and practice. Academic Press, London New York

Holmquist R (1972a) Theoretical foundations of paleogenetics. In: LeCam L, Neuman J, Scott EL (eds) Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol. 5, University of California Press, Berkeley, p 315

Holmquist R (1972b) Empirical support for a stochastic model of evolution. J Mol Evol 1:211–222

Holmquist R (1973) The stochastic model and deviations from randomness in eukaryotic tRNAs: Comparison with the PAM approach. J Mol Evol 2:145–148

Holmquist R (1976) Random and nonrandom processes in the molecular evolution of higher organisms. In: Goodman M, Tashian RE, Tashian JH (eds) Molecular anthopology. Plenum Press, New York, p 89

Holmquist R (1978a) The REH theory of protein and nucleic acid divergence: A retrospective update. J Mol Evol 11: 361–374

Holmquist R (1978b) A measure of the denseness of a phylogenetic network. J Mol Evol 11:225–231

Holmquist R (1979) The method of parsimony: An experimental test and theoretical analysis of the adequacy of molecular restoration studies. J Mol Biol 135:939–958

Holmquist R (1980) Evolutionary analysis of α and β hemoglobin genes by REH theory under the assumption of the equiprobability of genetic events. J Mol Evol 15:149–159

Holmquist R, Cimino JB (1980) A general method for biological inference: Illustrated by the estimation of gene nucleotide transition probabilities. BioSystems: J Mol, Cellular & Behavioral Origins and Evol 12:1–22

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

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

Holmquist R, Jukes TH, Moise H, Goodman M, Moore GW (1976) The evolution of the globin family genes: concordance of stochastic and augmented maximum parsimony genetic distances for α hemoglobin, β hemoglobin and myoglobin phylogenies. J Mol Biol 105:39–74

Holmquist R, Pearl D, Jukes TH (1981) Nonuniform molecular divergence: The quantitative evolutionary analysis of genes and messenger RNAs under selective structural constraints. In: Goodman M (ed) Macromolecular sequences in systematics and evolutionary biology. Plenum Press, New York

Jaynes ET (1979) Where do we stand on maximum entropy? In: Levine RD, Tribus M (eds) The maximum entropy formalism. MIT Press, Cambridge (Massachusetts), London, p 15

Jukes TH (1963) Some recent advances in studies of the transcription of the genetic message. Adv Biol Med Phys 9: 1–41

Jukes TH, Cantor C (1969) Evolution of protein molecules. In Munro HN (ed) Mammalian protein metabolism, III. Academic Press, New York, p 21

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

Jukes TH, Holmquist R, Moise H (1975) Amino acid composition of proteins: Selection against the genetic code. Science 189:50–51

Karon J (1979) The covarion model for the evolution of proteins: Parameter estimates and comparison with Holmquist, Cantor and Jukes' stochastic model. J Mol Evol 12:197–218

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1981) Was globin evolution very rapid in its early stages? A dubious case against the rate-constancy hypothesis. J Mol Evol 17:110–113

Kimura M, Ohta T (1972) On the stochastic model for estimation of mutational distance between homologous proteins. J Mol Evol 2:87–90

King JL (1980) Does the information density of amino acid composition increase? J Mol Evol 15:73–75

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

Lawn R, Efstratiadis A, O'Connell C, Maniatis T (1980) The nucleotide sequence of the human β-globin gene. Cell 21: 647–651

Moore GW (1977) Proof of the populous path algorithm for missing mutations in parsimony trees. J Theor Biol 66:95–106

Moore GW, Goodman M, Callahan C, Holmquist R, Moise H (1976) Stochastic versus augmented maximum parsimony method for estimating superimposed mutations in the divergent evolution of protein sequences. Methods tested on cytochrome *c* amino acid sequences. J Mol Biol 105:15–37

Nei M, Tateno Y (1978) Nonrandom amino acid substitution and estimation of the number of nucleotide substitutions in evolution. J Mol Evol 11:301–310

Ratner VA, Rodin SN, Zharkikh AA (1977) Analysis of the molecular evolution of globins by a more accurate method. In: Ratner VA (ed) Mathematical models of evolution and selection. Academy of Sciences USSR, Novosibirsk, p 67 (Figure 2).

Sattath S, Tversky A (1977) Additive similarity trees. Psychometrika 42:319–345

Shepard RN (1980) Multidimensional scaling, tree-fitting and clustering. Science 210:390–398

Tuppey H (1958) Über die Artspezifität der Proteinstruktur. In: Neuberger A (ed) Symposium on protein structure. John Wiley, New York, pp 66–76

Uzzell T, Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096

Verhoeyen M, Fang R, Min Jou W, Devos R, Huylebroeck D, Saman E, Fiers W (1980) Antigenic drift between the haemagglutinin of the Hong Kong influenza strains A/Aichi/2/68 and A/Victoria/3/75. Nature (London) 286:771–775

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

Zuckerkandl E, Schroeder WA (1961) Amino-acid composition of the polypeptide chains of gorilla haemoglobin. Nature 192:984–985

Received February 23, 1981/Revised April 18, 1981