

An Improved Method for Determining Codon Variability in a Gene and Its Application to the Rate of Fixation of Mutations in Evolution

Walter M. Fitch¹ and Etan Markowitz²

Received 15 Jan. 1970—Final 1 April 1970

If one has the amino acid sequences of a set of homologous proteins as well as their phylogenetic relationships, one can easily determine the minimum number of mutations (nucleotide replacements) which must have been fixed in each codon since their common ancestor. It is found that for 29 species of cytochrome c the data fit the assumption that there is a group of approximately 32 invariant codons and that the remainder compose two Poisson-distributed groups of size 65 and 16 codons, the latter smaller group fixing mutations at about 3.2 times the rate of the larger. It is further found that the size of the invariant group increases as the range of species is narrowed. Extrapolation suggests that less than 10% of the codons in a given mammalian cytochrome c gene are capable of accepting a mutation. This is consistent with the view that at any one point in time only a very restricted number of positions can fix mutations but that as mutations are fixed the positions capable of accepting mutations also change so that examination of a wide range of species reveals a wide range of altered positions. We define this restricted group as the concomitantly variable codons. Given this restriction, the fixation rates for mutations in concomitantly variable codons in cytochrome c and fibrinopeptide A are not very different, a result which should be the case if most of these mutations are in fact selectively neutral as Kimura suggests.

INTRODUCTION

Kimura (1968) has suggested that during the evolution of cytochromes *c*, hemoglobins, and triosephosphate dehydrogenase, the vast majority of mutations fixed were selec-

Paper number 1382 from the Laboratory of Genetics. Work performed in part at the University of Iowa, Department of Preventive Medicine and Environmental Health and Department of Statistics, Iowa City, Iowa. Computing supported by the Graduate College, University of Iowa.

¹ Department of Physiological Chemistry, University of Wisconsin, Madison, Wisconsin.

² Departments of Physiological Chemistry, Medical Genetics, and Statistics, University of Wisconsin, Madison, Wisconsin.

tively neutral mutations. King and Jukes (1969) agree but for different reasons. Corbin and Uzzell (1970) also agree and go on to calculate the absolute mutation rate on the assumption that all mutations are neutral in the codons for the most variable positions of fibrinopeptides A and B. The rate they found is approximately the same as the rate of mispairing of bases during replication as computed by Watson (1965). Maynard Smith (1968) takes issue with Kimura by choosing to dispute the validity of using Haldane's "cost of natural selection" (1957). Maynard Smith then devises a model which permits selection at many loci simultaneously and thereby escapes the dilemma. O'Donald (1969) agrees with Smith's conclusion regarding Haldane but also regards Maynard Smith's model as biologically unrealistic. O'Donald then proposes his own model and comes to the same ultimate conclusion as Kimura regarding the cost.

But regardless of how the cost is accounted, the conclusion that most of the mutations fixed in the structural genes for these proteins were selectively neutral may nevertheless be correct. However, if the vast majority of those mutations fixed are selectively neutral, then in those positions where mutations may be fixed, they should be fixed with equal probability and therefore at equal rates.³ This should be a strong consistency criterion for the validity of the neutrality of most fixed mutations. The difficulty lies in a proper estimate of the number of those positions. This paper will first show how to calculate the number of codon positions capable of fixing nondeleterious mutations and then, using this number, show that mutations are indeed being fixed at approximately equal rates in different genes.

DETERMINATION OF CODON VARIABILITY

A previously published procedure (Fitch and Margoliash, 1967a) showed that, provided a few recognizable coding positions were excluded from the data, the remainder of the mutations found in cytochrome *c* distributed themselves as if there were a number of invariant positions and the rest were equally likely to be the position fixing the next mutation; that is, the number of mutations fixed in the variable positions followed a Poisson distribution. That procedure is improved upon here. First of all, no coding positions are arbitrarily excluded. Secondly, an iterative procedure obtains the maximum likelihood estimates for the parameters. This leads to an expected distribution which may be compared to the observed distribution by the Pearson chi-square test. The procedure examines three possible models. Model 1 assumes that all positions are equally variable. Model 2 assumes that there are some invariant positions and that the rest are equally variable. Model 3 assumes that there are some invariant positions and that the rest constitute two classes of variability. Details are presented in the accompanying paper (Markowitz, 1970).

CODON VARIABILITY IN CYTOCHROME *c*

Figure 1 shows the phylogeny of 29 species based upon their cytochrome *c* sequences using the method of Fitch and Margoliash (1967b). When this particular phylogeny is

³ We recognize that there may exist more acceptable alternatives for some codons than for others at any one point in time. For the present study, we assume that this variability averages out when a collection of codons is examined over long periods of time.

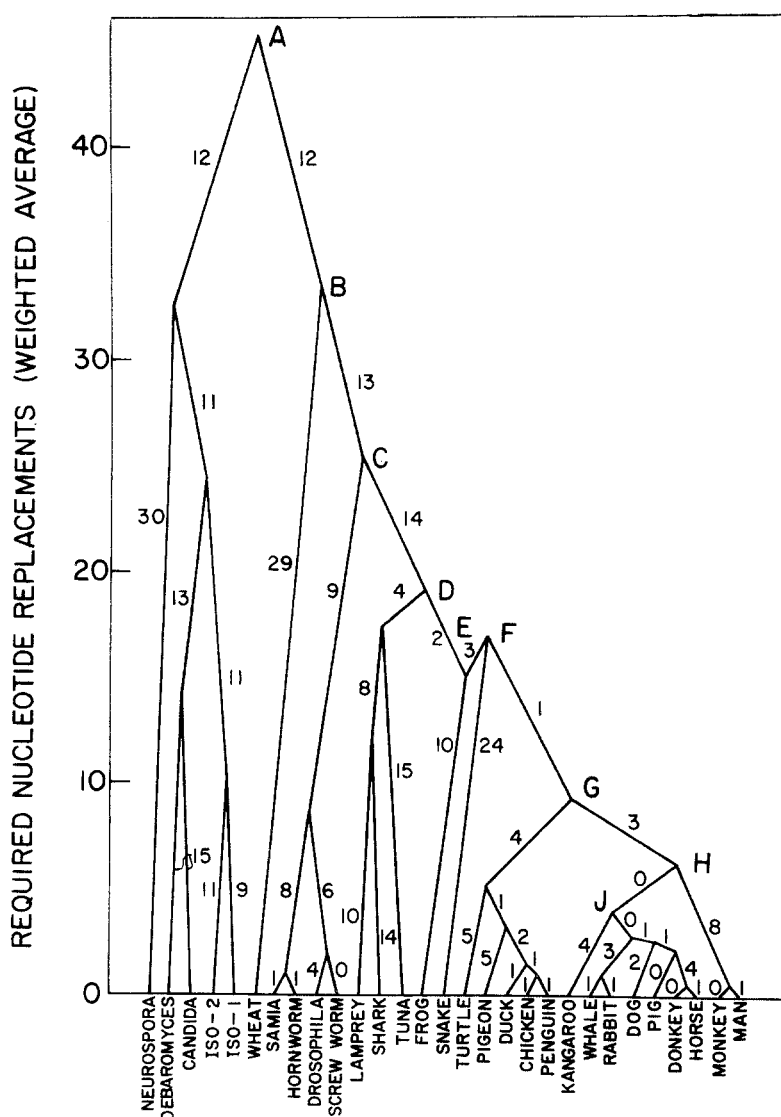


Table I. Distributions of Fixations According to Sites with Various Numbers of Fixations^a

F/S	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	χ^2	df	p
Group A	35	8	17	10	15	7	3	4	4	3	0	0	2	1	3	1	> 453*	> 112	< 10 ⁻²⁶
Model 1	4.4	14.3	23.2	25.1	20.3	13.2	7.1	3.3	1.3	0.5	0.2	0.05	< 0.02				> 100	> 10	< 10 ⁻¹⁴
Model 2	35	0.8	8.2	12.6	14.7	13.6	10.6	7.0	4.1	2.1	1.0	0.4	< 0.25				15	11	0.18
Model 3	35	8.7	13.8	14.5	11.4	7.2	4.7	3.1	2.4	2.2	2.0	1.9	1.6	1.2	0.8	1.4			
Group J	102	6	4	1													189*	112	< 10 ⁻⁶
Model 1	97.2	14.6	1.1	0.06													0.3	1	0.58
Model 2	102	6.6	3.1	1.2															
Group K	2	1	4	2	0	3	3	2	1	0	1								
Model 3	2	1.9	2.2	2.3	2.3	2.2	2.0	1.6	1.1	0.7	0.78						5.8	5	0.45

^a For each of three groups of amino acid sequences, the number of codon sites found to have fixed zero, one, two, etc., mutations is given under the column whose heading is the number of fixations per site (F/S). For each model is indicated the expected distribution of mutations in the group on the assumption that there are no invariant codons (model 1), that there are invariant codons (model 2), and that there are not only invariant codons but also that the variable codons are a mixture of two Poisson-distributed classes with different variability (model 3). Group A are the cytochromes *c* from the 29 species shown in Fig. 1. A total of 366 fixations is distributed over 113 codons. In determining χ^2 values, the rightmost expected value of the distributions contains all of the residual expectation lumped together. Model 1 χ^2 values, shown with an asterisk, are based upon Fisher's variance test or index of dispersion (*cf.* Markowitz, 1970). If all the data for fixations/site > 9 are lumped together, model 2 still gives a $\chi^2 < 51$ which, for 8 degrees of freedom, gives a probability $< 10^{-6}$ of occurring by chance. However, the same lumping of the data for model 3 reduces χ^2 from 15 down to 5.6 with an associated probability of 0.36. The Poisson parameters for the two variable classes of model 3 are 3.2 and 10.1. The 35 codons in model 3 that have never varied are composed of two groups: 32 invariant codons plus three variable but unvaried codons. Group J are the data from the cytochromes *c* of the nonprimate mammals of Fig. 1. A total of 17 fixations is distributed over 113 codons. The 102 codons of model 2 that have never varied are composed of two groups: 95 invariant codons plus seven variable but unvaried codons. No computation was performed for model 3 since there are insufficient degrees of freedom. Group K are data based upon the fibrinopeptides A from 25 mammalian species. A total of 80 fixations is distributed over 19 codons. The estimate of the number of invariant codons for group K is 1.1, with an estimated standard deviation of 1.9.

assumed, the 29 present-day sequences may be accounted for by a minimum of 366 fixations [i.e., nucleotide replacements; cf. Fitch (1970) for method]. Their distribution over the 113 codons is shown by group A in Table I, which gives the number of codons that were found to have fixed the number of nucleotide replacements indicated at the top of each column. Models 1, 2, and 3 immediately below group A are the best-fitting solutions according to the three possible models, respectively, outlined in the preceding paragraph. The χ^2 values are also shown in Table I for each of the three possible models. It can be seen that only the third model gives a reasonable fit to the data. For this distribution, the number of invariant codons is estimated at 32, with the 81 remaining variable codons being divided into two groups of size 65 and 16 with Poisson parameters of 3.2 and 10.1, respectively, which is to say that mutations, on a per codon basis, are being fixed in the latter small group at 3.2 times the rate of fixation in the former large group.

This substantiates the earlier conclusion (Fitch and Margoliash, 1967a) that there is a small group of codons which, relative to the others, is hypervariable,⁴ and contradicts the conclusion of King and Jukes (1969) that the "so-called 'hypermutable sites' . . . are predictable in terms of [a single] Poisson distribution." Since we are using the same basic data, the discrepancy needs to be accounted for. King and Jukes found no need to postulate a second Poisson distribution because they used a less sensitive procedure in determining the number of mutations in each coding position. They simply examined all the amino acids in each position and determined the number of nucleotide replacements necessary to accomplish their interconversion [their method is more explicitly given in Jukes (1969)]. This fails to take proper account of the total number of replacements required as a consequence of the phylogeny. For example, in position 47 of cytochrome *c* of the 20 species examined in the earlier study (Fitch and Margoliash, 1968) there is an alanine in *Neurospora*, a threonine in the screw worm fly, the kangaroo, and the horse, and a serine in two fungi, one insect, one fish, two reptiles, four birds, and six other mammals. There is no reasonable explanation of these data other than to postulate that on three separate occasions, a mutation from a codon for serine to one for threonine has been fixed, and we would thus count it as three fixations rather than one. King and Jukes have not followed this practice. This is of some importance since, of the 230 mutations in that paper (Fitch and Margoliash, 1968), there were 41 such parallel mutations as well as three back-mutations comprising 20% of the total mutations. These would not have been assessed by King and Jukes.

A model with precisely two classes of variable codons is not as biologically attractive as a model with only one. If the variable positions are not all equally variable, and this is clear from the data, then it is more reasonable to assume that there is a range of variability. Our using only two classes of variability is a simplification that merely reflects a combination of the limited sensitivity of the method, the limited amount of sequence data, and the generally well-behaved nature of the property we are measuring. A model that predicted how the range of variability was distributed could be tested,

⁴ "Hypervariable" refers to the fact that, relative to the majority of the codons fixing mutations, this group is fixing them more rapidly. Since one cannot assume that these codons are more mutable, the term "hypermutable" (Fitch and Margoliash, 1967a) has been replaced by "hypervariable" (Fitch and Margoliash, 1968).

but for the present the limited approximation of only two degrees of variability will suffice.

CONCOMITANTLY VARIABLE CODONS

In what was to have been an attempt to show that the same distribution of variability among the codons is obtained regardless of the range of organisms examined, the fungi were omitted from group A to form group B and the computations repeated. As a result, the number of invariant codons increased from 32 to 51. The computation was

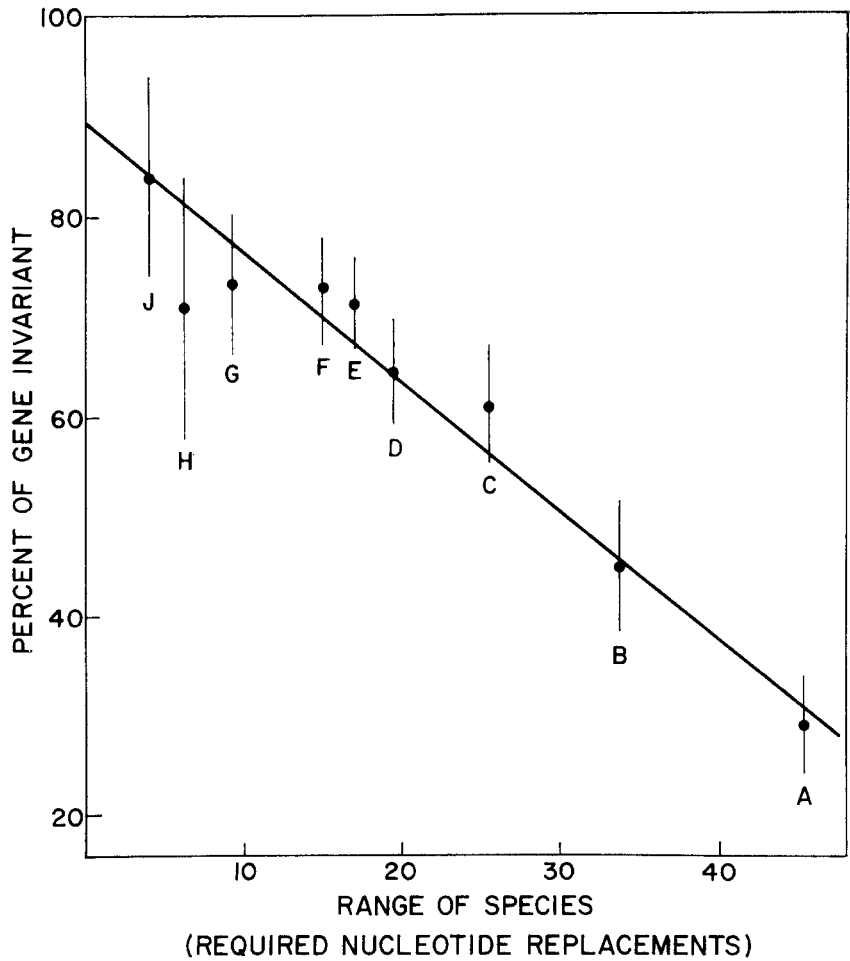


Fig. 2. Concomitantly variable codons. The percent of the gene found to be invariant (y) is plotted as a function of the range of species examined (i.e., peak height in Fig. 1). A least-squares fit to the results is extrapolated to the ordinate, which estimates the fraction of the gene for which all mutations are lethal or malefic. The rightmost and leftmost values plotted are for the groups A and J, respectively, of Table I. Each vertical line extends one estimated standard deviation above and below its y value.

successively repeated, each time using a new group which was formed by omitting the most remote members of the previous group. These groups are shown in Fig. 1 by a letter at the appropriate apex, the members of a particular group being all descendants from the apex bearing that group's letter designation. The reduction in range of species continued until, for group J, only nonprimate mammals remained. The 17 mutations required to account for their differences were distributed as shown in Table I by Group J. Also shown are the best-fitting distributions for models 1 and 2 (there are too few degrees of freedom to fit model 3). For model 2, the estimate of the number of invariant codons is 95, a sharp contrast to the 32 calculated for trial group A. The percentage of the gene which is estimated to be invariant increases as the range of species narrows. That range may be represented by the height of the apex of the group shown in Fig. 1. Figure 2 shows that the inverse relationship between the range of species examined and the percentage of the gene that is invariant is linear in the region for which we have data.⁵ The line shown is a weighted least-squares fit; each point's weight is the reciprocal of its approximate variance. It is vital to recognize that the invariant codons are only a subset of the codons that have not varied. As the range of species narrows, the number of codons that have not varied necessarily increases but the same is not necessarily true for the invariant codons.

The line in Fig. 2 may be extrapolated upward to the y axis. We interpret this intercept to mean that when the range of species being examined is reduced to zero (i.e., when only one species' cytochrome c gene is considered), over 90% of the codons in the mammalian cytochrome c gene are invariant. Although this appears to contradict our experience that, if a broad enough range of species is examined, most of the codons will have varied, there is a rational biological explanation available. This explanation asserts that because of the structural restraints imposed by functional requirements, mutations that will not be selected against are available only for a very limited number of positions. We shall use the term acceptable for such mutations. However, as such acceptable mutations are fixed they alter the positions in which other acceptable mutations may be fixed. Thus, only about 10 codons, on the average, in any cytochrome c may have acceptable mutations available to them but the particular codons will vary from one species to another. We shall term those codons at any one instant in time and in any given gene for which an acceptable mutation is available as the *concomitantly variable codons*.⁶

King and Jukes (1969) state that "74 to 81 residues are variable" and that "there is very little restriction on the type of amino acid that can be accommodated at most of the variable sites." We would interpret our results as indicating that, contrariwise, there is much restriction in the cytochrome c gene most of the time with very few of the codons having acceptable mutations available and we see no reason to assume other-

⁵ Although the curve is linear in this region, we would assume that it must, as the range of species increases, eventually approach, as a lower limit, an asymptote parallel to the x axis at a height indicative of the truly invariant positions. If prokaryotic cytochromes c of *Pseudomonas fluorescens* (Ambler, 1963) and *Rhodospirillum rubrum* (Dus *et al.*, 1968) are considered to be homologous to the eukaryotic horse cytochrome c as shown in Table II, there are only seven of the 75 common positions identical so that we may put an upper limit for the asymptote at 9.4%.

⁶ The phrase *concomitantly variable codons* is awkward and we would suggest the term "covarions" to describe this particular set of codons.

Table II. Comparison of Prokaryotic and Eukaryotic Cytochromes *c*^a

<i>Ps. fluor.</i>					GLU	asp	pro	GLU	val	leu	PHE
<i>Rh. rubrum</i>	glu	GLY	ASP	ala	ala	ala	GLY	GLU	LYS	—	—
Horse		GLY	ASP	val	GLU	lys	GLY	lys	LYS	ile	PHE
	lys	asn	LYS	gly	CYS	val	CYS	HIS	ala	ile	ASP
	VAL	ser	LYS	lys	CYS	leu	CYS	HIS	THR	phe	ASP
	VAL	gln	LYS	—	CYS	ala	CYS	HIS	THR	val	glu
	—	—	—	thr	lys	met	VAL	GLY	PRO	ala	lys
	gln	GLY	GLY	ala	asn	LYS	VAL	GLY	PRO	ASN	phe
	lys	GLY	GLY	lys	his	LYS	thr	GLY	PRO	ASN	his
	asp	VAL	ala	ala	lys	phe	ALA	GLY	GLN	ALA	ala
	GLY	VAL	PHE	glu	asn	thr	ALA	ala	his	lys	asn
	GLY	leu	PHE	gly	arg	lys	thr	GLY	GLN	ALA	gly
	—	—	—	—	—	—	—	—	—	—	—
	tyr	ala	TYR	ser	glu	ser	tyr	thr	glu	met	ala
	phe	thr	TYR	thr	asp	ala	asn	—	—	—	asn
	—	—	—	—	—	—	GLU	ALA	glu	LEU	gln
	LYS	GLY	leu	THR	TRP	thr	GLU	ALA	asn	LEU	ala
	LYS	GLY	ile	THR	TRP	lys	GLU	glu	thr	LEU	glu
	arg	ile	LYS	ASN	gly	ser	gln	gly	VAL	trp	GLY
	TYR	val	LYS	ASN	PRO	LYS	ala	phe	VAL	leu	glu
	TYR	leu	glu	ASN	PRO	LYS	lys	tyr	ile	pro	—
	—	—	—	—	—	—	—	—	—	—	—
	ile	pro	—	—	—	—	—	—	—	—	—
	ser	gly	asp	pro	lys	ala	lys	ser	LYS	MET	pro
	—	—	—	—	—	—	—	thr	LYS	MET	PHE
	asn	ala	val	ser	—	ASP	ASP	GLU	ala	gln	LEU
	—	lys	leu	thr	LYS	ASP	ASP	GLU	ile	GLU	val
	ala	gly	ile	lys	LYS	lys	thr	GLU	arg	GLU	LEU
	ala	lys	trp	val	leu	—	—	ser	gln	LYS	—
	ILE	ALA	TYR	LEU	LYS	—	—	THR	leu	LYS	—
	ILE	ALA	TYR	LEU	LYS	lys	ala	THR	asn	glu	—

^a A dash (—) indicates a gap inserted to maintain the homologous alignment. Underlining indicates those 7 positions which are still invariant. Except for the invariant asparagine (ASN), all of these residues interact with the heme or with residues which do in bovine cytochrome *c* (R. E. Dickerson, personal communication). Positions with a dot under them are invariant among the eukaryotic cytochromes *c* so far examined.

wise than that the range of acceptable mutations in those codons may be similarly limited.

SPATIAL CORRELATIONS IN AMINO ACID REPLACEMENTS

The implication that events in one coding position may be dependent upon events in other positions may be related to the interesting observations of Wyckoff (1968) who noted a spatial relationship between many pairs of the substitutions observed

between rat and bovine ribonucleases. This was possible because the three-dimensional structure of bovine RNase is known. For example, the amino acids in positions 38 and 39 of rat RNase are glycine and serine, respectively. Now glycine could mutate to aspartate but, as Wyckoff points out, presumably this would be damaging because it could interact with lysine 41 and pull this necessary residue out of the active site. Also, the serine could mutate to arginine, with there being no particular reason to suspect that this might not be acceptable. In bovine RNase, the groups are indeed aspartate and arginine but the positively charged arginine neutralizes the negatively charged aspartate and, perhaps, prevents any deleterious effect of the aspartate on the critical lysine 41. If this is true, we presume that the fixation of arginine at 39 must have preceded the fixation of aspartate. This illustrates well how the positions belonging to the group of concomitantly variable codons may change since before the arginine fixation position 38 might not tolerate an aspartate, whereas after both fixations position 39 might not tolerate the return of arginine to a neutral residue.

Another example is provided by positions 57 and 79, which in bovine RNase are valine and methionine and in the rat are isoleucine and leucine, respectively. Although the individual amino acids are of differing volumes, the pairs have the same total length. If there is no further room in the site filled by these residues, we would have to assume that in the ancestral pair, the longer of the amino acids (say isoleucine) had to be replaced by a shorter descendant (say valine) before the other ancestral amino acid (say leucine) could tolerate a mutation changing it to a longer descendant (in this illustration, methionine). This avoids trying to fit both methionine and isoleucine into the restricted space at the same time. For this to be a neutral process, it must be assumed that there is tolerance for the loss of a methylene group in this region. Wyckoff presents evidence for such tolerance at another site in RNase.

At position 14 is an aspartate that is charge-bonded to an arginine in position 33 of the cow and position 32 of the rat. Presumably, in the intermediate form, aspartate 14 has an option between the two arginines (rather than none) but since only one was needed, it was acceptable for one to be replaced subsequently and it was. A summary of these three cases is shown in Fig. 3.

This discussion of paired replacements in RNase is extended because the very specific nature of the replacements observed necessarily focuses one's attention upon the selective requirements of the positions. Nevertheless, as we hope the preceding indicates, it is possible to provide a rational explanation for these events in terms of neutral mutations even though the latitude permitted to what is acceptably neutral may be severely limited by selective forces and even though the neutrality of the intermediates has not been demonstrated. It also shows that we are not yet forced to postulate any deleterious intermediates to account for such paired fixations. And finally, it shows with a biological system how the positions that can accept mutations may change as mutations are fixed. These views encompass assertions regarding the order of the mutational events. As a consequence of knowledge of phylogeny, some orderings have already been postulated for cytochrome *c* (Fitch and Margoliash, 1968) without recourse to such considerations as presented here. It will be interesting to see if similar predictions based upon the forthcoming three-dimensional structure of cytochrome *c* will agree with the phylogenetic orderings.

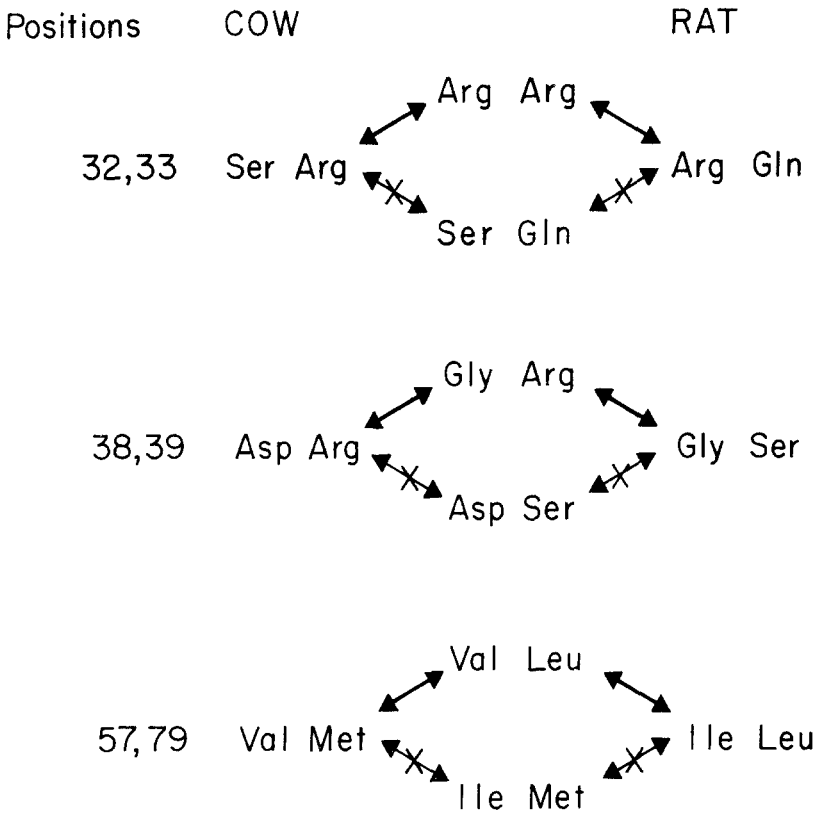


Fig. 3. Transition states in ribonuclease. The figure shows the bovine and rat ribonuclease residues in three selected pairs of positions together with two alternative intermediate forms. In each case, the upper alternative is predicted to have been the true ancestral intermediate on the basis of information supplied by Wyckoff (1968) and elaborated on in the text.

There is corroborative evidence for the view that the number of concomitantly variable codons is very restricted for cytochrome *c*. It was previously noted that the number of codons that had fixed nucleotide replacements in two of its three nucleotides in the interval between two successive nodes of the phylogenetic tree was greatly in excess of expectation if the mutations were being fixed at random (Fitch and Margoliash, 1968). These so-called double mutations appeared to imply a non-randomness in the fixations observed and therefore to suggest the possible operation of selective forces. However, that computation was performed assuming 82 variable codons so that if four mutations were fixed between two successive nodes in the tree, the probability that two of them would have occurred in the same codon was about 0.04. When this computation was carried out for each of the 26 internodal segments containing two or more fixations in that tree for 20 species of cytochrome *c*, there was an expectation of only 11.3 "double" mutations. In fact, there were 32. The number of

expected double mutations would, of course, be increased if the number of concomitantly variable codons were considerably less. A repeat of that calculation assuming 10 rather than 82 concomitantly variable codons yields a value of 32 expected “double” mutations, the number actually found. This means first of all that there is an independent basis (the larger number of “double” mutations) for concluding that there is a restricted number of concomitantly variable codons in the gene for cytochrome *c*, and the agreement of the two methods at a value of 10 such codons is encouraging. Secondly, it means that the “double” mutation data can now be accounted for strictly on the basis of random events and thus selective forces are not necessarily involved.

It should be noticed that the agreement between the expected and found number of “double” mutations is too good in that the calculation of the expected number is inadequate in two ways. One involves the absence of a correction for the possibility that the second fixation in the same codon may be in the same nucleotide. The second involves the absence of a correction for the fact that the 10 concomitantly variable codons represent a sort of moving average and, after a series of fixations, are not the same codons that were concomitantly variable at the beginning. That this second effect is surely present can be seen from Table III, where the preceding data are divided into two groups according to the internodal segment length as determined by the number of mutations thereon. We underestimate the number of “double” mutations in the segments with only a few mutations and overestimate that number in the segments with many mutations. This is a result that presumably would be corrected by using fewer concomitantly variable codons while simultaneously correcting for the change in the codons that are concomitantly variable. Work designed to accomplish

Table III. Distribution of “Double” Fixations by Segment Length^a

Internodal mutations (<i>m</i>)	No. of such internodal segments	Number of “double” fixations		$(E - F)^2$ <i>E</i>
		Expected (<i>E</i>)	Found (<i>F</i>)	
$2 \leq m \leq 7$	16	5	10	5.00
$9 \leq m \leq 30$	10	27	22	0.93
Total	26	32	32	5.93

^a Data are from Fitch and Margoliash (1967*a*), which gives the codon position and the location on the phylogenetic tree of all postulated mutations in the gene for cytochrome *c* during the evolution of 20 eukaryotic species. “Double” mutations are nucleotide replacements in two of the three nucleotides of a single codon on a tree segment between two successive ancestral nodes or between the present-day species and its most recent ancestor. The table shows that while the total expected and found “double” mutations agree, their distribution by segment length does not. χ^2 is 5.93 for 1 degree of freedom. Such a departure from expectation would occur by chance less than 2% of the time if the underlying assumptions about the distribution of “double” mutations were correct.

these corrections is currently in progress and, in addition to increasing the reliability of this estimate of the number of concomitantly variable codons, it should also give information on the degree to which a fixation affects other positions. This, of course, is subject to the qualification that fixations in other genes with gene products that interact with cytochrome *c* (e.g., cytochrome oxidase) may also affect which codons will belong to the group of concomitantly variable codons.

The number of invariant residues of fibrinopeptide A was determined using the data from 25 mammalian sequences and is shown in Table I as group K. The estimate of the number of invariant positions is only one. Further restriction of the range, unlike the case for cytochrome *c*, did not change the estimated number of invariant codons. Presumably, that one invariant codon is the carboxyl-terminal arginine that is necessary for the specificity of thrombin which acts on fibrinogen to form fibrin in the blood-clotting mechanism. In any event, extrapolation to a zero range of species gives a single invariant codon and therefore 18 concomitantly variable codons, a result consistent with the view that this fragment is important simply by virtue of its existence as a blocking group and that its specific structural configuration may be of little importance.

EVOLUTIONARY RATES

The cytochromes *c* and the fibrinopeptides A were chosen in part because of the difference in the rate at which they appear to be evolving. Both groups contain sequences for the horse and pig and, since their common ancestor, there have been five fixations in the 104 codons for cytochrome *c* and 13 fixations in the 19 codons for fibrinopeptide A. This gives rates of 0.048 and 0.684 fixations/codon in the two lines since their divergence from a common ancestor. Thus it would appear that fixations are occurring in the part of the gene encoding fibrinopeptide A at more than 14 times the rate they occur in the cytochrome *c* gene. If, however, we assume that the rate should be calculated on the basis of the number of codons that can accept a mutation, the rates become, respectively, $5/10 = 0.50$ and $13/18 = 0.72$ fixations/concomitantly variable codon. Given the limits of the estimate of the number of concomitantly variable codons, these two fixation rates cannot be said to differ. Furthermore, if the number of concomitantly variable codons in cytochrome *c* is in fact less than 10, as suggested above by the "double" mutation data, then the agreement may be even better.

The calculation of the number of concomitantly variable codons does not presuppose the nature of any selective forces involved. It only requires that a mutation, to be acceptable, must not be selected against, which leaves open the question of its possible neutrality. However, if the mutations are indeed neutral, then the fixation rates in concomitantly variable codons should be the same.³ Since the rates are the same, then either the mutations are, as Kimura suggested, mostly neutral or else the agreement is fortuitous. There is a third possibility, namely, that selection proceeds to fix mutations at essentially similar rates in all genes in all species at all times. But to contrive a selective theory so as to mimic the results expected of a neutral mutation theory is philosophically unpalatable. In view of the diverse nature of these two polypeptides, we would reject the explanation based upon fortuity. In any event,

when similar rates can be estimated for other gene products, we will know whether the agreement is fortuitous.⁷

There is additional reason to believe that the accumulation of mutations has been of a random rather than a selective nature. If selective forces were greatly affecting the nature of cytochrome *c*, one would expect that, say, wheat, *Drosophila*, shark, duck, and human cytochromes *c* would show greatly varying degrees of differentness when compared to *Neurospora* cytochrome *c*. In fact, when the number of nucleotide replacements required to account for the differences in these five pairs is determined, they are found to be 69, 60, 71, 63, and 64, respectively. Is it really reasonable to suggest that the proper interpretation of these data is that, since the time of their common ancestor, the shark's cytochrome *c* has been evolving under selection more than 20% faster⁸ than the human cytochrome *c*? Or do these data simply reflect sampling variability? Nolan and Margoliash (1968) have given many other examples of how, for any one gene, approximately the same number of mutations separate the members of two groups having the same common ancestor. This multiplication of such examples lends further support to the idea that most of the mutations fixed in the genes for cytochromes *c*, fibrinopeptides, and hemoglobins have been selectively neutral. This conclusion, regarding fibrinopeptide A, stands in sharp contrast to that of King and Jukes (1969), who state, "within the short fibrinopeptide A fragment, however, some positions are notably less changeable than others. It is quite likely that only a minority of the changes that occur in this portion of the fibrinogen gene are selectively neutral." Group K of Table I clearly depicts the changeable character the authors speak of, but the row below also clearly shows that the most variable positions "are predictable in terms of the Poisson distributions," to use the authors' own words from another context.

The difficulty perhaps lies in an unwillingness of King and Jukes to assume the existence of a group of codons more variable than another. We might ask therefore how a second, more variable group might arise. Possibilities include: (1) their codons are hot spots, i.e., they are hypermutable; (2) a greater variety of alternatives is acceptable at these positions; (3) an alternative is acceptable over a greater portion of the total time span.

While we are not as yet prepared to choose among these alternatives,⁹ they appear sufficiently reasonable to permit acceptance of the possibility of detecting two groups of codons with differing degrees of variability.

If we assume that the fixation rate is 0.72 and that the common ancestor of the horse and pig existed 80 million years ago (Romer, 1966), then the fixation rate would

⁷ The most obvious next choice for examination would be hemoglobin, but only about a half dozen species of each of the α and β chains have been completely sequenced whereas at least a dozen for any one orthologous gene should be available. This emphasizes the importance of doing further sequences on orthologous gene products from closely related species.

⁸ The 20% comes from the fact that 71 is 10% greater than 64, that the difference must necessarily be related to mutations which occurred since the common ancestor of the shark and man, and that presumably not more than half (32) of the mutations separating the vertebrate and fungal lines have occurred since the time of the common ancestor of the shark and man.

⁹ The third possibility cannot apply to fibrinopeptide A since the number of concomitantly variable codons is constant at 18 regardless of the range of species examined and hence all 18 are always variable.

appear to be 4.5 fixations/concomitantly variable codon/10⁹ years in any one line of descent. The fixation rate should be relatable to the mutation rate, but the calculation is complicated by the fact that not all mutations in a concomitantly variable codon are necessarily acceptable.

Finally, we would make a point of noting the importance of verifying the idea that most mutations fixed in structural genes have been neutral and that on the basis of the number of concomitantly variable codons, fixation rates are uniform because, given one good paleontological reference point we shall then have for the first time a very excellent evolutionary time scale at our disposal. With luck, this could extend back into pre-Cambrian times and permit the temporal location of many events such as speciation and gene duplications which now can only be guessed at.

ADDENDUM

Since this paper was submitted and reviewed, the sequence of porcine ribonuclease has appeared (Jackson and Hirs, 1970). The amino acids at positions 57 and 79 are identical to those in the cow. The amino acids at positions 32, 33 and 38, 39 are ArgArg and GlyArg, respectively, and are the intermediates between the bovine and rat forms predicted on the basis of Wyckoff's data and shown in Fig. 3.

ACKNOWLEDGMENTS

This project received support from National Science Foundation grant GB-7486. The University of Wisconsin Computing Center, whose facilities were used, also receives support from NSF and other government agencies.

REFERENCES

- Ambler, R. P. (1963). The amino acid sequence of *Pseudomonas* cytochrome c-551. *Biochem. J.* **89**: 349.
- Corbin, K. W., and Uzzell, T. (1970). Natural selection and mutation rates in mammals. *Amer. Nat.* **104**: 37.
- Dus, K., Sletten, K., and Kamen, M. K. (1968). Cytochrome *c*₂ of *Rhodospirillum rubrum*. *J. Biol. Chem.* **243**: 5507.
- Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zool.* **19**: 99.
- Fitch, W. M., and Margoliash, E. (1967*a*). A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case. *Biochem. Genet.* **1**: 65.
- Fitch, W. M., and Margoliash, E. (1967*b*). The construction of phylogenetic trees. *Science* **155**: 279.
- Fitch, W. M., and Margoliash, E. (1968). The construction of Phylogenetic trees. II. How well do they reflect past history? *Brookhaven Symp. Biol.* **21**: 217.
- Haldane, J. B. S. (1957). The cost of natural selection. *J. Genet.* **55**: 511.
- Jackson, R. L., and Hirs, C. H. W. (1970). The primary structure of porcine pancreatic ribonuclease. *J. Biol. Chem.* **245**: 637.
- Jukes, T. H. (1969). Evolutionary pattern of specificity regions in light chains of immunoglobins. *Biochem. Genet.* **3**: 109.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**: 624.
- King, J. L., and Jukes, T. H. (1969). Non-Darwinian evolution. *Science* **164**: 788.
- Markowitz, E. (1970). Estimation and testing goodness-of-fit for some models of codon fixation variability. *Biochem. Genet.* **4**: 595.

- Maynard Smith, J. (1968). "Haldane's dilemma" and the rate of evolution. *Nature* **219**: 1114.
- Nolan, C., and Margoliash, E. (1968). Comparative aspects of primary structures of proteins. *Ann. Rev. Biochem.* **37**: 727.
- O'Donald, P. (1969). "Haldane's dilemma" and the rate of natural selection. *Nature* **221**: 815.
- Romer, A. S. (1966). *Vertebrate Paleontology*, 3rd ed., University of Chicago Press, Chicago.
- Watson, J. D. (1965). *Molecular Biology of the Gene*, Benjamin, New York.
- Wyckoff, H. W. (1968). Discussion. *Brookhaven Symp. Biol.* **21**: 252.