

The Phylogeny of tRNA Sequences Provides Evidence for Ambiguity Reduction in the Origin of the Genetic Code

W.M. FITCH* AND K. UPPER†

*Department of Biological Sciences, University of Southern California, Los Angeles, California 90089-1481;

†Department of Genetics, University of Washington, Seattle, Washington 98195

In 1966, Fitch proposed the ambiguity reduction hypothesis of the origin of the genetic code, based on a view that the origin of life was a process in which local (pre)biological order arose from molecular chaos on the earth, driven by the asymmetric energy budget of the earth's atmosphere, a process in which subsets of random biochemical events gradually became the programmed rule of the system. This in turn led to a view, regarding the origin of the genetic code, that suggests that originally there may have been little specificity regarding which amino acids were charged to the various RNA acceptors that paired to the message. Under such conditions, no messenger RNA is likely to produce exactly the same protein twice. The advantages of obtaining a well-defined protein sequence, however, would have gradually reduced the variability in the assignment of amino acids to codons until the current genetic code emerged. If so, the history of that reduction in ambiguity might be recorded in the phylogenetic history of the tRNAs. This suggests a test of the hypothesis. Does a correspondence exist between the pattern of the genetic code and the inferred phylogeny of the tRNAs? In this paper, we show that, for eight tRNAs, the correspondence is precisely of the type required by the ambiguity reduction hypothesis.

THE HYPOTHESIS

Many of the details of that ambiguity reduction are irrelevant to this study. It does not matter if, at an earlier time, there were more or less than 20 amino acids, nor does it matter to what extent deoxynucleotides might have been involved. Perhaps during this period some of the amino acids that might have been used were gradually excluded. α -Amino butyrate and sarcosine are both common products of attempts to create amino acids abiotically (Miller 1957). If they were common originally, they must have been present in pre-genetic-code proteins and later selected against. The ambiguity reduction hypothesis does require, however, at the time of its occurrence, that the basic method of an acceptor tRNA of about 75 nucleotides, having 3 nucleotides that base-pair with message, had already been reasonably well developed. It further requires that a genetic information-storing system was already present and that the descendant tRNAs are all paralogous, having arisen by gene duplication.

As there are several scenarios that are consistent with the hypothesis, it is useful to elaborate one. The general nature of the reduction can be visualized in the following way. If there were no specificity whatsoever, then any amino acid could be charged to any anticodon acceptor. This could easily be the case since, if the protein sequences were not well specified, the charging enzymes might not be able to distinguish one nucleotide from another. The code would then have been NNN = any amino acid (N = any nucleotide; Fig. 1, top). However, if the system developed a bias that tended to charge hydrophobic amino acids preferentially to tRNAs matching messenger-coding triplets that had a central pyrimidine (Y), and tended to charge hydrophilic amino acids to tRNAs matching messenger-coding triplets that had a central purine (R), then the code would become NYN = hydrophobes, NRN = hydrophiles. Hydrophobes and hydrophiles are broadly descriptive of what the NYN and NRN codons encode today. This is equivalent to separating the codons into two groups, as shown by the column of closed circles in the genetic code (Fig. 1, middle). It is also equivalent to improving the charging enzyme to the point where it can distinguish between the sizes of the two major classes of nucleotides.

It is not important to the test of the hypothesis to know or guess what the first ambiguity reduction step was nor what the selective force furthering that choice was, but it may be useful to suggest how even a first differentiation such as NYN/NRN might have been advantageous. Consider a simple repeat of $(\text{NYNNRN})_n$. Although it does not define a protein, it does define, as suggested by Brack and Orgel (1975), an alternating series of hydrophobic and hydrophilic amino acids, exactly what is required to form β -pleated sheets, one of the two major forms of secondary structure in proteins today. Similarly, α -helical properties would reside in another repeat only slightly more complicated. By so minor a reduction in ambiguity, one may specifically code for the two most important substructures of biologically active proteins.

The second ambiguity reduction step might have been a choice, as shown by the row of filled squares in Figure 1 (bottom), between purines and pyrimidines in the first codon position leading to YYN, RYN, YRN, and RRN, specifying four different groups of amino acids. The cyclic amino acids (tyrosine, histidine, tryptophan, and proline) are also specified by these four

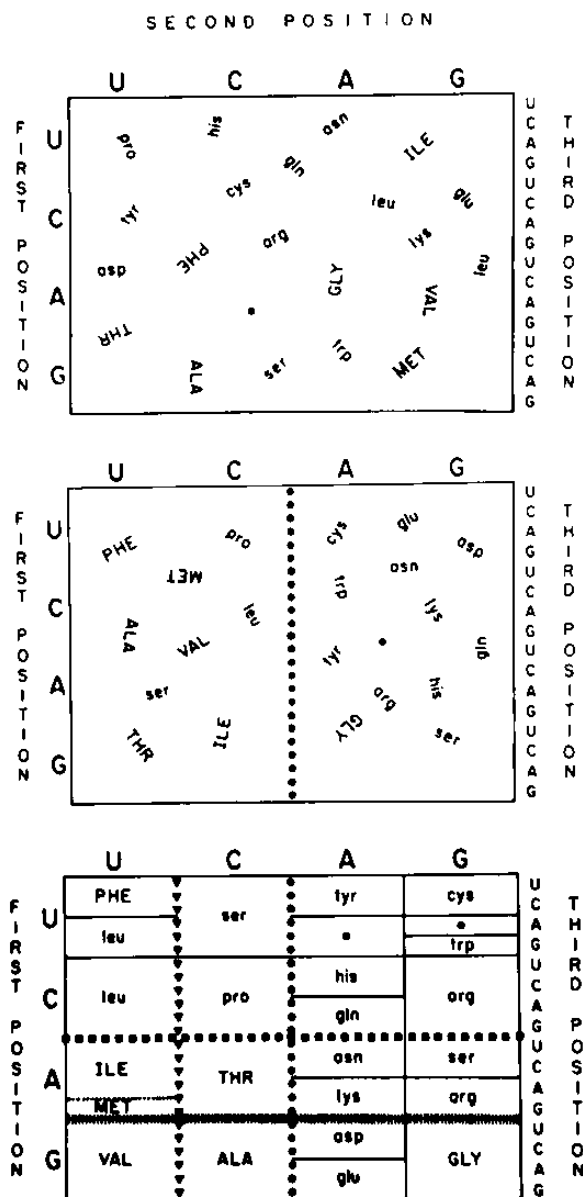


Figure 1. The genetic code. (*Top*) The fully ambiguous "genetic code" as it might have been initially with no particular preference of any amino acid for any particular codon. (*Middle*) The genetic code as it might have been at an early stage if the first reduction in ambiguity had been the institution of a preference for hydrophobic amino acids to be charged to acceptors recognizing a pyrimidine in the middle codon position and for hydrophilic amino acids to be charged to acceptors recognizing a purine in the middle codon position. (*Bottom*) The genetic code as it is today. The various symbols indicate possible evolutionary divergences from an ambiguous ancestral assignment. Squares and circles denote the first differentiation into purine and pyrimidine recognition in the first and second codon positions, respectively; short vertical lines denote the subsequent differentiation into adenine and guanine recognition in the first position purines; triangles denote the subsequent differentiation into cytosine and uracil recognition in the second position pyrimidine; dotted line denotes the differentiation into G and not-G recognition in the third position of the AUN codon. The order of these differentiations is assumed to be independent except that at any one position the purine-pyrimidine differentiation must precede any other differentiation at that position. This restriction was chosen solely on the basis that the purine-pyrimidine differentiation is essentially what is seen in the third position today. Other possible differentiations are omitted because no amino acid reflective of that section of the coding table was used in this study. These differentiation symbols are used to clarify the interpretation of the phylogenies in Fig. 2. All capitalized amino acid abbreviations denote amino acids whose tRNAs are used in this study.

tophan, proline, and phenylalanine) all have codons beginning with a pyrimidine in the first position. In a subsequent step, the pyrimidine might have been separated into the cytidine and uridine specificities that today separate those amino acids with phenyl rings (tyrosine, tryptophan, and phenylalanine) from the heterocyclics histidine and proline (not specifically depicted in Fig. 1).

It is only necessary to continue the process of increasing codon specificity (ambiguity reduction) through additional steps to arrive at a stage where every amino acid has its own set of codons, the ones it uses today (Fig. 1, bottom). This work was started for the purpose of finding evidence that such a process did indeed occur.

The third position of the genetic code is today largely differentiated, if at all, by a purine-pyrimidine division.

Moreover, it could well be that proteins at the earliest stages of evolution were not able to discriminate more finely than between purines and pyrimidines. We have therefore assumed that the purine-pyrimidine division was the first to occur in the first two codon positions as well. There is nothing that forbids, for example, a G-not G (= H) division as in fact is, except for some mitochondria, present today in most methionine-isoleucine codons. These codons must, therefore, either have divided that way initially or else isoleucine later acquired the AUA codon from methionine. Our restriction of the initial division in the first two codon positions into purines and pyrimidines limits the number of possible evolutionary patterns that can be constructed that are consistent with the pattern of the genetic code. This is important to the statistical test to be presented.

METHODS AND RESULTS

General Considerations

Materials. A total of 300 tRNA sequences were examined. For each of the following amino acids there was the accompanying number of tRNAs: Ala, 39; Gly, 43; Ile, 32; Met₁, 48; Met_m, 20; Phe, 29; Thr, 35; and Val, 54. There may have been inadvertent duplicates in the set in that two sequences obtained from different sources that did not agree completely would both be included, there being no effort to see if the difference might not be a sequencing error as opposed to strain differences. Three fungal met_ms are in fact recent duplications of the met, and are genetically in that category, although they were counted here in the functional met_m category. Sources of sequences were GenBank, EMBO, W. McLain, and H. Nicholas (unpubl.), R. Cedergren (unpubl.), and Sprinzl et al. (1985a,b).

It is important that the sequences be in homologous alignment. The alignment of Sprinzl et al. (1985a,b) is a spatially equivalent alignment because the portions that form the secondary structure are given a dominating preference. In this case the homologous and spatially equivalent alignments are almost always congruent, and we therefore adopted their alignment. In the few cases where it seemed obvious that the spatially equivalent alignment was nonhomologous, we altered the former to the homologous alignment.

Tree construction and ancestral sequence estimation. Given a set of sequences homologously aligned, it is possible, for any particular proposed genealogical (phylogenetic) relationship, to determine the minimum number of nucleotide substitutions necessary to obtain these sequences from their common ancestor by the method of Fitch (1971).

We began by examining the set of tRNAs for each amino acid separately to find the most parsimonious tree for each. We used many different starting trees representing different possible phylogenetic relationships. Each of these trees was subjected to the swapping of its neighboring branches in search of better (more parsimonious) trees until a tree was obtained that could not be improved by swapping neighboring branches. The tree requiring the fewest substitutions, after branch swapping, from among all starting phylogenies was presumed to be the best estimate of the true phylogeny. In these procedures, all positions of the tRNA were treated equally, irrespective of their base pairing and irrespective of any posttranscriptional modifications.

The Test of the Ambiguity Reduction Hypothesis

Imagine that one has a good estimate of the nucleotide sequence for each of eight amino acid tRNAs as they existed in the most recent ancestor common to all the organisms that are alive today, that is, in the cenancestor (cen-, from the Greek kainos, meaning recent, and koinos, meaning common). One can im-

agine that the phylogeny of the cenancestor tRNAs for those amino acids might correspond to the pattern of ambiguity reduction shown in the trees of Figure 2. In that case, the tRNA phylogeny and the ambiguity reduction hypothesis would correlate perfectly. One problem is to determine the probability of a favorable outcome.

The test will involve the set of eight tRNAs whose amino acids are shown at the branch tips of Figure 2. The determination of their cenancestor sequences will be the concern in the next section. For now, we need to know how many possible phylogenies there are for eight (cenancestor tRNA) sequences. The number of unrooted trees, t , for n taxa = $(n - 2)!! = \Pi(2i - 1)$ for $i = 1 \rightarrow n - 2$ (Fitch and Margoliash 1968). Thus there are 10,395 unrooted trees. We only count unrooted trees because the procedure we shall use involves a parsimonious estimate of the total number of nucleotide substitutions required for a tree, an estimate that is not affected by the location of the root of the tree.

The next question is, how many of those 10,395 trees are consistent with the genetic code's having a pattern of ambiguity reduction like that in Figure 1? Only the four shown in Figure 2. A favorable outcome for the ambiguity reduction hypothesis would be that at least one of the four ambiguity reduction trees was (among) the most parsimonious of the 10,395 trees.

Constructing the Cenancestor Sequence for Each tRNA

Theory. The test using the eight cenancestor tRNA sequences first requires that we have an estimate of that sequence for each of them. This will require the use of the parsimony method that necessarily has within it, for each node on a tree, an estimate of the ancestral sequence that is consistent with the fewest possible nucleotide substitutions on that tree (Fitch 1971). For that ancestor to be as good an estimate as possible the following must be true. (1) There must be the largest possible number of tRNAs for each amino acid. We used a total of 300 tRNAs. (2) The diversity of the taxa must span the diversity of extant species. If only mammalian species were included, the estimated ancestral tRNA sequence would be only that in the ancestral mammal, not that in the cenancestor. We required that any tRNA have at least one known representative from each of the following five major groups: archaebacteria, eubacteria, eukaryotes (preferably plants, animals, and fungi), chloroplasts, and mitochondria (there is no met_m in mitochondria where one met tRNA serves both purposes and it is of the met₁ lineage). (3) The root of the tree of living species must be known. This is not known but is solvable as shown below.

Result. The most parsimonious tree(s) from all starting trees was examined for each amino acid tRNA separately. From these we discovered that each of the five major groups of sequences consistently appeared as a single group on the tree. That is, for any one amino

could well be that there is value in terms of amino-acyl-tRNA-synthetase efficiency if isoacceptor tRNA pairs differ only in their anticodon. If so, and if a pair of isoacceptor tRNAs had drifted genetically apart, there might be positive selection to replace an older form by a newer one created by gene duplication and a G-U interchange in the anticodon of the duplicated gene.

Rooting the Tree

Theory. The characteristics of the most recent ancestor of a group of organisms are potentially inferable from the characteristics of its descendants. One cannot usually infer the characteristics of organisms more ancient than that ancestor due to the absence of other relatives that would provide information. The tRNAs appear to be an exception, however, in that they are presumed to be paralogous, the gene for these various amino acid acceptors having arisen by gene duplications that occurred *prior* to the cenancestor, before the most recent ancestor of all the organisms alive today. That hypothesis of a paralogous relationship will be tested below and found to be statistically supported, but, for the moment, let us consider the consequences if the hypothesis is indeed correct and if all the gene duplications occurred prior to the cenancestor.

Figure 3 illustrates for a group of three taxa and a set of three genes. The question arises, what would happen if a single tree were to be sought using all nine sequences at once? Of the 135,135 possible unrooted trees, three nonrandom examples are shown.

In the second row, the tree has three subtrees, each for a different taxonomic group. This might reflect reality, but it would imply that all the gene duplications occurred subsequent to the divergence of the taxonomic groups being studied.

In the third row is a possible result in which there are again three subtrees, each for a different gene, the divergence of the species having occurred after the gene duplications. Unfortunately, no two of the three subtrees show the same order of taxon divergence.

In the bottom row is the desired result. The three subtrees separate the genes, implying that the genes duplicated prior to the cenancestor and that the order of the taxon divergence within each subtree is the same. This permits us to identify the cenancestral node as the root of those subtrees. This is shown as a closed circle on each of the three subtrees and corresponds to linking the top row trees to each other at their corresponding closed circles. A similar approach to tree rooting using the internal duplication of bacterial ferredoxin was presented by Schwartz and Dayhoff (1978).

In the actual test, there will be sequences for eight genes from each of the major taxonomic groups giving a result rather better than that shown in the third row, but not as perfect as that shown in the bottom row. If, at the level of the major groups of sequences, the tRNAs for each amino acid are orthologous, then one should get the same phylogeny of the groups for each amino acid tRNA. Moreover, if one constructs a large

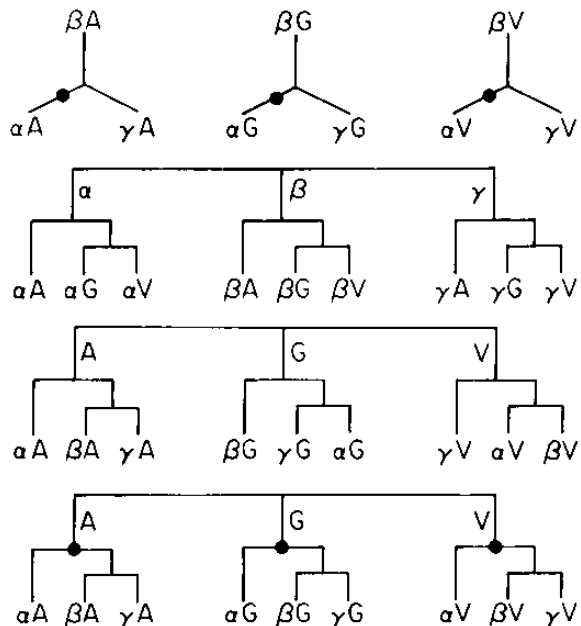


Figure 3. Finding the root of the tree of life, a hypothetical example. The top row shows three trees, each for a separate paralogous gene, A, G, and V. They could be, for example, the tRNAs of alanine, glycine, and valine. For each gene we have a representative from each of three widely divergent taxa, α, β, and γ, that one expects to span the range of all living things. They could be, for example, from archaebacteria, eubacteria, and eukaryotes. The dots represent the roots of that tree. The other three trees show, from the top down, increasing relatedness to their correct phylogeny if all three genes arose via gene duplications prior to the speciation of the organisms in which they reside.

phylogeny that includes simultaneously the five major group ancestral tRNAs (eubacterial, archaebacterial, eukaryotic, chloroplast, and mitochondrial) for all eight amino acid tRNA sets, then one should get a tree with four properties: (1) All the group ancestral tRNAs for any one amino acid set should cluster together; (2) within each such cluster one should observe the same evolutionary relationships among the five major groups; (3) the position possessing the link from any one cluster to the other clusters should locate in the tree where the cenancestral tRNA for that amino acid is to be found; and (4) that position, that amino acid set's root, should be the same for each set. Because the chloroplast and eubacterial sequences always clustered together in the study using all the tRNAs separately, we used the common ancestor of the two, thereby reducing the number of major groups from five to four.

Result. The group ancestors of the tRNA set were examined by the parsimony method with branch swapping with the following results: (1) The four group ancestors in each tRNA set clustered together except for the noninitiating met_m tRNA, which contained the initiating met_i tRNA cluster within it; (2) the tree for each cluster was the same for the seven clustered sets and is shown in Figure 4; (3) the root location for each set (also shown in Fig. 4) is not the same.

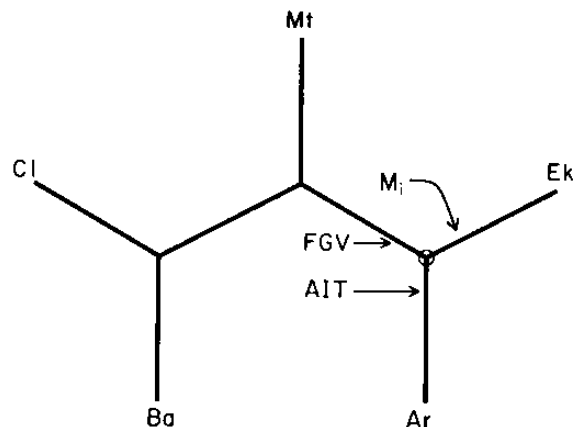


Figure 4. The clustering and rooting of the five major groups for each tRNA. The most parsimonious tree(s) for 31 group ancestral tRNAs was sought. These group ancestral tRNAs were for the eight sets of amino acid tRNAs (the seven shown: A, alanine; F, phenylalanine; G, glycine; I, isoleucine; T, threonine; V, valine; M_i , initiating methionine; plus the not-shown noninitiating methionine) for each of the four groups (Ar, archaeobacteria; Ek, eukaryotic; Mt, mitochondrial; and the composite Ba, eubacterial plus Cl, chloroplast). These 32 (= 8 sets \times 4 groups) ancestors are reduced to 31 because there are no noninitiating met tRNAs in mitochondria. The tRNAs for any one of the seven amino acids on the figure cluster exclusively among themselves. Moreover, within any one amino acid set, the phylogeny of its four group ancestral tRNAs was, in every case, identical to that shown in the figure. The location of the point in the subtree that connects that subtree to all the others (the root) is shown by the arrows labeled according to the amino acid set whose root is located there. Had all the amino acid tRNA sets rooted at the same location, it would have provided strong evidence that the tree of life was rooted there, that the cenacestor occurs at that point. In the absence of a clear choice, the node with a circle around it was chosen to represent the cenacestor.

From the uniformity of these results, we concluded that the basic arrangement of the five major groups is as shown and supports the earlier suggestion from Woese's laboratory (Fox et al. 1980) of three major kingdoms. Unfortunately, it is not possible to determine from these data where the root lies and hence not possible to determine which was the earliest speciation event, that is, which two of the three kingdoms are the most closely related. Nevertheless, it is clear that the circled node in Figure 4 represents that node closest to the cenacestor, and its sequence was therefore chosen to represent the cenacestor in the final test.

The Tree for Eight Cenacestral tRNAs

None of the species phylogenies for an individual set of tRNAs agreed in detail with any other and we therefore took the preceding results only to indicate the relationships required of the five groups and to determine the node whose sequence would represent the cenacestor. Accordingly, we then constructed, using parsimony, a cenacestral sequence for each set of tRNAs, based upon trees that were consistent among all eight tRNAs and that represented our judgment of what biologists would believe to be the correct tree. Only where we could not discover what that belief

might be did we let the parsimony results from the analyses above influence decisions as to the order of bifurcation within the five major groups. Only in the latter instances did we allow parsimony to influence the nature of the tree.

With the eight cenacestral tRNAs in hand, all possible trees for them were examined by the parsimony procedure. The distribution of the resulting 10,395 tree lengths is shown in Figure 5. The ideal result to support the ambiguity reduction hypothesis would be if one of the four most parsimonious trees (length = 100 nucleotide substitutions) were one of the four trees (Fig. 2) that are consistent with the genetic code. That was not the case. One of the code-consistent trees required 104 nucleotide substitutions, the other three required 106. The details for this tail of the distribution are presented in Table 1. As seen there, all four code-consistent trees reside in the lower 3.5% of the distribution.

An important assumption of this work is that the different tRNA genes are paralogous, that is, homologous by virtue of gene duplications, rather than analogous, that is, similar by virtue of convergence from unrelated ancestral genes. The method of Fitch (1970) permits one to decide between these two choices. This test was applied to each of the interior nodes in the upper left tree of Figure 2. For no node was the probability greater than 10^{-8} that the sequence similarity between its two immediate descendants could, by chance, be as great as that observed. Thus, we have direct evidence that the different tRNA genes arose from a common ancestor as required for the ambiguity reduction hypothesis.

As another by-product of the method, one obtains an estimate of the urancestral tRNA sequence that was the gene that, by gene duplication, gave rise to the various paralogous tRNA genes. This sequence is shown in Figure 6. Immediately below it is the ursequence as proposed by Eigen and Winkler-Oswatitsch (1981), what they called the "master sequence." It was formed from a different, but not disjoint, set of tRNAs by a

Table 1. Distribution of Most Parsimonious Trees in Tail

Substitutions	Trees	Σ Trees	$\Sigma/10,395$
100	4	4	0.0004
101	10	14	0.0013
102	21	35	0.0034
103	25	60	0.0058
104	55(1)	115	0.0111
105	89	204	0.0196
106	144(3)	348	0.0335

Column 1 is the number of nucleotide substitutions required to account for the descent of the eight cenacestral tRNAs from their common urancestral sequence. Column 2 is the number of trees that require the number of substitutions shown in column 1. Column 3 is the number of trees that require a number of substitutions equal to or less than the number shown in column 1. Column 4 is the fraction of all trees that require no more than the number of substitutions shown in column 1. (1) and (3) are the number of trees at that number of substitutions that are consistent with the ambiguity reduction hypothesis.

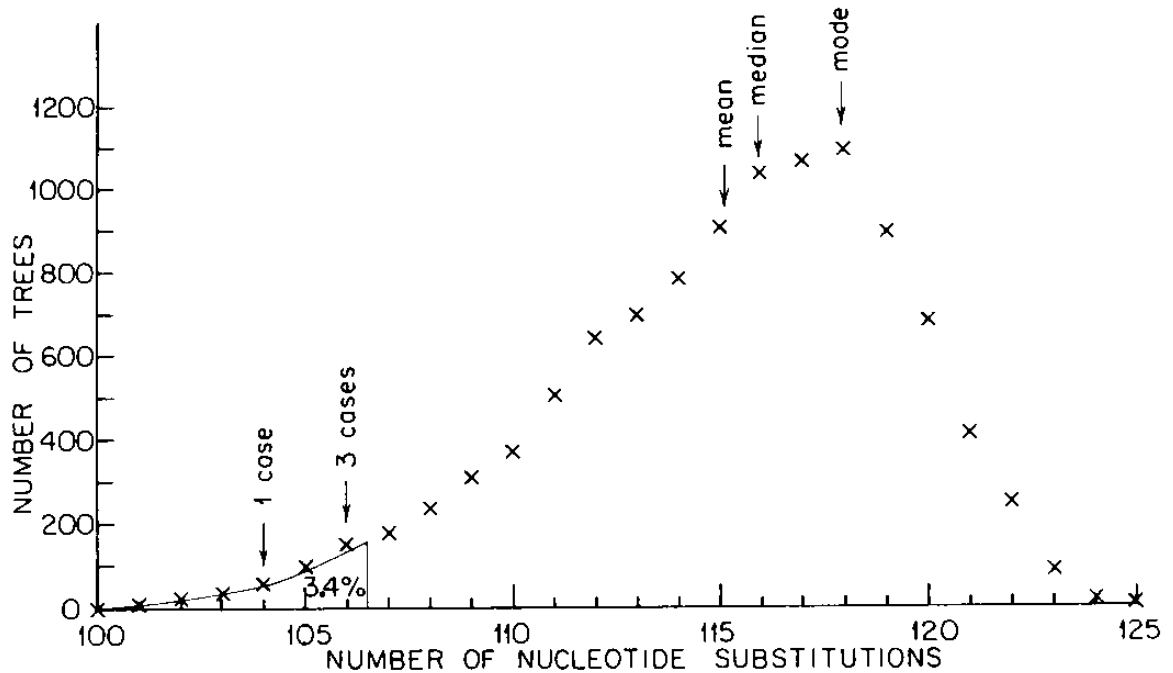


Figure 5. Distribution of tree lengths among the 10,395 possible unrooted trees for the cenacestral tRNAs for eight amino acids. The trees all required between 100 and 125 nucleotide substitutions, inclusive, which number is shown on the abscissa. The number of trees that required any specific number of substitutions is plotted on the ordinate. The locations of the four trees that are consistent with the genetic code as shown in Fig. 2 are indicated at the left of the distribution and they fall, as shown, in the lower 0.0335 of the distribution. That value was found by dividing the number of trees whose length is ≤ 106 substitutions (346) by the total number of trees examined (10,395). The probability that all four trees would fall in the lower 0.0335 of the distribution must be less than 0.0335.

different method. When allowance is made for the four extra positions we have retained from our alignment process, the agreement appears rather good and suggests that the estimates have some validity.

Although the ancestral sequences were reconstructed without regard to whether the secondary structure was retained, as can be seen in Figure 7, the secondary structure has been retained except for the last (fourth)

base pair before the D loop. Only 4 of the 21 base-pairing positions (including the fourth on the D-loop stem) have A or U nucleotides. This is less than the average tRNA but not less than some present-day tRNAs. One should not infer from this that the urancestral sequence was richer in GC pairs to withstand a warmer environment. That might be true but, if the sequences today need a preponderance of GC pairs in

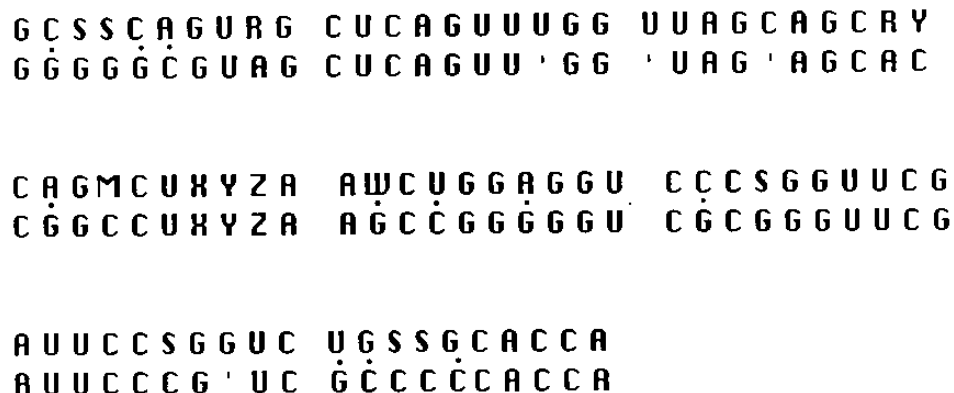


Figure 6. Comparison of our urancestral sequence with the master sequence of Eigen and Winkler-Oswatitsch. The upper sequence is from this work, the lower from Eigen and Winkler-Oswatitsch (1981). The subsequence XYZ represents the anticodon, hyphens are gaps introduced in their sequence to improve the correspondence, and dots between sequences indicate mismatches. The difference in length arises because some tRNAs have extra positions and, since our sequence preserves them all whether they were in fact present in the original ancestral sequence, no disagreement in this respect between the sequences should be inferred. We do, however, believe that an alignment based on homology is preferred to one based on structural equivalence if one is trying to infer ancestral sequences. The single letter IUB code for nucleotides is used, in which Y = C or U, R = A or G, S = C or G, W = A or U, and M = A or C (Nomenclature Committee of the International Union of Biochemistry 1986).

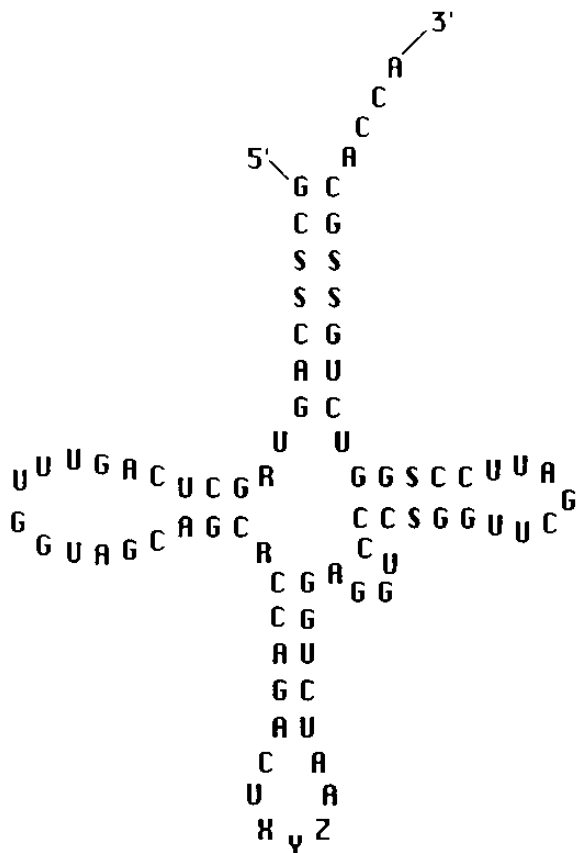


Figure 7. Tertiary structure of the urancestral tRNA. The sequence is the same as the upper sequence of Fig. 6, except that ambiguous nucleotides were made less ambiguous whenever some of the alternative nucleotides were not consistent with normal AU or GC pairing.

their stems but do not care too much where they are, and if through the course of evolution the minority of AU pairs have become distributed across many of the positions, then a simple consensus method must, and a parsimony method is likely to, infer a GC pair to be ancestral in any particular position not functionally required to be an AU pair.

DISCUSSION

Implications of the Result

The shape of the distribution of trees in Figure 5, had it been obtained from random sequences, would have been quite normal looking. Thus its very shape, with its skewed distribution and an extended tail to its lower end, implies the retention of significant biological information in the eight reconstructed cenancestral sequences. This is fortunate since, if the distribution had appeared normal, we would have given less credence to the observation that all four consistent trees were in the lower 3.5% of the distribution.

The probability that any one tree would lie in the lower 3.5% of the distribution is, of course, 0.035. That at least one of the four trees would lie there is consid-

erably greater. That all four would lie there is considerably less than 0.035 but certainly not as low as $(0.035)^4$, since the trees are not independent of each other. It is, nevertheless, a conservative estimate to say that there is less than a 0.035 probability that the null hypothesis is true, that there is no relation between the pattern of the genetic code and the phylogeny of the tRNAs. Said differently, the phylogeny of the tRNAs is consistent with the pattern of the genetic code with better than a 96.5% level of confidence, thus supporting the ambiguity reduction hypothesis.

Limitations to the Significance of the Result

We regard the result as significant in that it rejects the null hypothesis. That is not the same as proving the ambiguity reduction hypothesis that motivated the test in the first instance. Any hypothesis that required this kind of a relationship between the pattern of the genetic code and the phylogeny of the tRNAs is equally supported by our result.

Moreover, the observation of the pattern says nothing about the forces, if any, that might create such a pattern. For example, Woese (1967) has proposed what has been called the *instructive* or *direct recognition* theory (Fitch 1973). In its extreme form, this means that there is a specific interaction between the amino acid and the anticoding triplet so strong that if the evolutionary process were to occur a second time, the identical coding assignments would be made. We are not fond of this version, but in a version closer to Woese's original proposal one can imagine, for example, that there might be preferences or tendencies that made it more likely that hydrophobes would be associated with a second position adenine (A) in the anticodon (that is, with a uracil in the messenger RNA).

Woese (1965) also proposed that his direct recognition process was probably faulty in the early stages of evolution of the translation apparatus. His proposal is not materially different from mine, except that he proposes a mechanism, reduction in the erroneous recognition of the fit between amino acid and adapter.

On the other hand, perhaps the assignment was, as Crick (1968) has suggested, a largely random assignment, a *frozen accident* such that, if that evolutionary process occurred a second time, the coding assignments would have little relation to the present one. Thus, our result may constrain physical theories to forms consistent with an evolutionary development of increasing specificity, but it cannot be in conflict with any physical theory so constrained.

Our results are also not in conflict with special theories such as Jukes's *intruder hypothesis* (1973), which suggests that proteins originally used ornithine and only later was arginine substituted for it. The observation of the pattern does not depend on the presence or absence of subsequent switching of one amino acid for another. Such switching, if it has occurred and is unrecognized, might confound attempts to infer

physical bases for the original assignment of amino acids to the tRNAs, but it does not affect the conclusion of a relationship between the genetic code and the history of its tRNAs.

Our results do appear, however, to contradict the proposal of Sheppard (1981) that the primitive code was for only the eight amino acids coded today by the subset of 16 coding triplets defined by RNY. This appears to us an unlikely proposal in that it also requires the recognition system to specifically differentiate all four nucleotides but not use that ability fully in any position.

ACKNOWLEDGMENT

This work was supported by National Science Foundation grant BSR-8796183.

REFERENCES

Brack, A. and L.E. Orgel. 1975. β -Structures of alternating polypeptides and their biological significance. *Nature* **256**: 383.
 Crick, F.H.C. 1968. Origin of the genetic code. *J. Mol. Biol.* **38**: 367.
 Eigen, M. and R. Winkler-Oswatitsch. 1981. Transfer-RNA: The early adapter. *Naturwissenschaften* **68**: 217.
 Fitch, W.M. 1966. Evidence suggesting a partial, internal duplication in the ancestral gene for heme-containing globins. *J. Mol. Biol.* **16**: 1.
 ———. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**: 99.
 ———. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* **20**: 406.

———. 1973. Aspects of molecular evolution. *Annu. Rev. Genet.* **7**: 343.
 Fitch, W.M. and E. Margoliash. 1968. The construction of phylogenetic trees. *Brookhaven Symp. Biol.* **21**: 217.
 Fox, G.E., E. Stackebrandt, R.B. Hespell, J. Gibson, J. Maniloff, T.A. Dyer, R.S. Wolfe, W.E. Balch, R.S. Tanner, L.J. Magrum, L.B. Zablen, R. Blakemore, R. Gupta, L. Bonene, B.J. Lewis, D.A. Stahl, K.R. Luehrsen, K.N. Chen, and C.R. Woese. 1980. The phylogeny of prokaryotes. *Science* **209**: 457.
 Jukes, T.H. 1973. Arginine as an evolutionary intruder into protein synthesis. *Biochem. Biophys. Res. Commun.* **53**: 709.
 Margulis, L. 1981. Photosynthesis in plastids (chap. 11). In *Symbiosis in cell evolution*. Freeman Publications, San Francisco.
 Miller, S.L. 1957. The mechanism of synthesis of amino acids by electric discharges. *Biochem. Biophys. Acta* **23**: 480.
 Nomenclature Committee of the International Union of Biochemistry. 1986. Nomenclature of incompletely specified bases in nucleic acid sequences. *Mol. Biol. Evol.* **3**: 99.
 Schwartz, R.M. and M.O. Dayhoff. 1978. Origins of prokaryotes, mitochondria, and chloroplasts. *Science* **199**: 395.
 Sheppard, J.C.W. 1981. Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. *J. Mol. Evol.* **17**: 94.
 Sprinzl, M., T. Voderwulbecke, and T. Hartman. 1985a. Compilation of sequences of tRNA genes. *Nucleic Acids Res.* **13**: r51.
 Sprinzl, M., J. Moll, F. Meissner, and T. Hartman. 1985b. Compilation of tRNA sequences. *Nucleic Acids Res.* **13**: r1.
 Woese, C. 1965. On the evolution of the genetic code. *Proc. Natl. Acad. Sci.* **54**: 1546.
 ———. 1967. On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **31**: 723.