Proc. Natl. Acad. Sci. USA
Vol. 77, No. 5, pp. 2791–2795, May 1980
Evolution

Convergence and minimal mutation criteria for evaluating early events in tRNA evolution

(molecular evolution/genetic code/phylogeny)

R. J. CEDERGREN*, BERNARD LARUE†, DAVID SANKOFF‡, GUY LAPALME§, AND HENRI GROSJEAN¶

*Départements de Biochimie et §d'Informatique et ‡Centre de Recherches Mathématiques, Université de Montréal, Montréal, Québec, Canada H3C 3J7; †Département de Chimie-Biologie, Université du Québec à Trois-Rivières, C.P. 500, Trois-Rivières, Québec; and ¶Laboratoire de Chimie Biologique, Université Libre de Bruxelles, 67 rue des Chevaux, 1640 Rhode-St-Genèse, Belgium

Communicated by John J. Hopfield, February 29, 1980

ABSTRACT **The convergence of ancestral sequences independently constructed from different branches of a phylogenetic tree can be used as a test of homology of data sequences. This criterion has shown that all phenylalanine tRNAs are related to a common ancestor, whereas eukaryotic and prokaryotic tyrosine tRNAs may have independent origins. All glycine tRNAs share a common ancestor. The glycine tRNA family splits according to the purine or pyrimidine nature of the first anticodon base prior to the divergence of eukaryotes and prokaryotes. The structural similarity between some prokaryotic glycine and valine tRNAs is the result of their derivation from a common ancestor that existed previous to the divergence of the different glycine tRNAs. These results support models of genetic code evolution involving the incremental elaboration of earlier, simpler codes.**

Hypotheses on the evolution of the protein translation apparatus often assume that the present-day genetic code has emerged from simpler, less precise versions in which fewer than the current 20 amino acids were implicated (1–8). Although rather speculative at the moment, it is conceivable that the identity of only one and later two nucleotides of the triplet codon specified amino acids in the earlier codes leading eventually to a code where some amino acids required all three letters of the codon to be unambiguously assigned. This scenario implies the emergence of different tRNA species that could transport the newly added amino acids. Therefore, in addition to providing a molecular record of species divergence, the comparison of tRNA sequences may permit the determination of relationships between amino acids which led to the expanded and refined present-day genetic code.

Even though well over 120 tRNA sequences have been determined, the hopes of deducing the early evolutionary events described above remain largely unfulfilled; a random evolutionary fluctuation or convergent evolution (or both) have seemingly introduced enough noise to obliterate many phylogenetic relationships (9, 10). Fig. 1 is indicative of this noise effect; the distribution of structural differences between tRNA specific for the same amino acid and tRNAs of different specificities overlaps extensively. It follows that phylogenetic trees based solely on difference matrix methods may be misleading because the common origins of distantly related sequences will be largely hidden among random similarities (10). On the other hand, most comparisons of tRNAs within the same kingdom (i.e., eukaryotic or prokaryotic) having the same anticodon clearly indicate that this restrained group of tRNAs are evolutionarily related.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

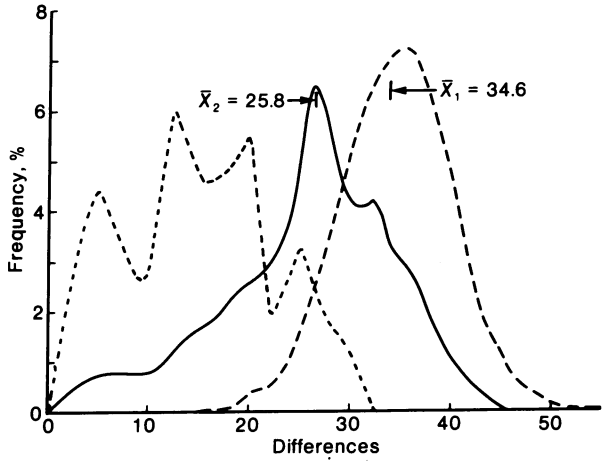


FIG. 1. Distribution of nucleotide differences in pairwise comparisons of tRNAs. —, Distribution in comparisons of tRNAs specific for the same amino acid; average value, 25.8. – – –, Distribution of nonisoacceptor tRNA pairwise combinations; average value, 34.6. - - -, Distribution of differences from pairwise comparisons of tRNA specific for the same amino acid, containing the same anticodon, and within either the eukaryotic or prokaryotic group. The dashed curves were smoothed by averaging adjacent values.

Thus, phylogenetically relevant information has been at least partly conserved and, consequently, we examine sequence data here with the idea of a *stepwise* approach to the deduction and evaluation of early evolutionary events in genetic code elaboration (11): tRNAs are progressively clustered into families according to their anticodon, their amino acid specificity, and, finally, the sequence similarities when different acceptor activities are involved. At each step, the evolutionary relationships are evaluated by a "convergence" criterion, described below, in conjunction with minimal mutation methods to see whether, in a phylogenetic tree constructed for a tRNA family, the intermediate ancestral sequences show significantly more similarity than data sequences.

Convergence criterion

Given the nucleotide sequence of a set of tRNA molecules, there are various methods for construction of the most likely phylogenetic tree and ancestral or nodal sequences (12, 13); "most likely" in this sense means the tree or sequence that is compatible with the minimal number of mutational events necessary to account for the data. However, even with a set of totally random sequences, it is always possible to construct a minimal tree.

Abbreviation: $tRNA^{AA}_{XXX}$, tRNA specific for the amino acid, AA, and having the anticodon XXX.

Implicit in any representation of divergent evolution by a tree diagram is that, as one moves toward the root, the nodal positions should become more similar, as is shown in [1],

[1]

if a homologous relationship (representing a common line of descent) exists between the four sequences; the nodal sequences A and B, *independently* constructed from 1–2 and 3–4, respectively, should be closer to each other than the average of the 1–3, 1–4, 2–3, and 2–4 distances. No such convergence towards a unique root, however, should exist if the two groups of sequences are evolutionarily unrelated. We have incorporated this principle into a procedure for assessing the relationships between eukaryotic and prokaryotic branches of the phenylalanine, tyrosine, and glycine tRNA families as well as between isoaccepting species of $tRNA^{Gly}$ and $tRNA^{Val}$.

Data treatment

Most sequences were taken from Gauss *et al.* (14), although some newly determined sequences were also available for this analysis. Sequences were aligned according to their common structural elements as described (11, 14). Positions in the variable loop region that are often unfilled and the three anticodon nucleotides were not considered in the analysis. Ancestral sequences were determined with the aid of a computer as described (15). Mutational distances between sequences were established by counting the number of positions that are not identical as a result of either a replacement or an insertion/deletion. Because the reconstruction of ancestral sequences leads to uncertainty in some positions, the contribution of each position is taken to be the average distance of each nucleotide combination from the two sequences.

Phenylalanine tRNA family

The first type of question we examine is whether our convergence criterion can confirm that prokaryotic and eukaryotic branches of a given tRNA family are related. Based on the availability of data, the obvious first choice to study is the phenylalanine family which has a single anticodon type (GAA) and is represented by sequences from five true prokaryotes, two chloroplasts, five eukaryotic cytoplasms, and one mitochondrion. The matrix of differences between each tRNA pair is given in Table 1. All intrabranch comparisons except for those involving the $tRNA^{Phe}$ from *Schizosaccharomyces pombe* indicate a significant relationship (see Fig. 1). This matrix was then used to derive the eukaryotic and prokaryotic phylogenetic trees in Fig. 2 by the minimal mutation method, with the exception of the *Euglena* tRNA position (see below). In a first experiment, the protoprokaryotic sequence was constructed and compared to the protoeukaryotic sequence, which was calculated without the use of the *Schizo. pombe* sequence (judged to be too distant from the other eukaryotes to be certain of its lineage). The two protosequences constructed differed by 25.6 positions, whereas the average distance between present-day prokaryotic and eukaryotic tRNAs is 27.0. Although these values are consistent with the notion of convergence, a dramatic improvement is obtained when the *Schizo. pombe* sequence is incorporated in the eukaryotic tree as in Fig. 2. In this revised tree the eukaryotic and prokaryotic protosequences differ by only 22.2 positions and the average interbranch distance is 27.7. The full significance of this figure appears when it is realized that the minimal distance calculated for the entire tree between the prokaryotic and eukaryotic ancestors (convergence is calculated independently from the two branches) differs by 19 positions; this value represents the best possible convergence and illustrates the large divergence between the two branches. Furthermore, we calculated the distance between the ancestral nodes of 100 pairs of prokaryotic and eukaryotic subtrees; each subtree was obtained by randomly permuting the species labels in the original tree. Only one of the 100 permuted trees converges better than the tree of Fig. 2 (22.1 positions different between ancestral sequences). Incidentally, this single exception is not at all credible biologically and requires far more than the minimal number of mutations to account for the data sequences. This illustrates the *balanced* use of the minimal mutation and convergence criteria to assess evolutionary hypotheses.

Another use of the convergence test is seen in consideration of the position of *Euglena*. Based only on minimal mutation

Table 1. Difference matrix for comparisons among $tRNA^{Phe}$ isolated from various sources

		13	12	11	10	9	8	7	6	5	4	3	2	1
Escherichia coli	1	28	24	33	22	27	28	12	17	19	21	19	18	
Bacillus stearothermophilus	2	33	28	30	28	29	33	13	20	21	17	3		
B. subtilis	3	33	27	27	27	26	30	14	21	22	16			
Mycoplasma	4	32	29	33	29	27	28	18	25	24	pro-pro			
Euglena chloroplast	5	28	27	30	25	23	29	13	6					
Bean chloroplast	6	31	27	29	25	25	31	11						
Anacystis nidulans	7	33	25	32	23	23	31	eu-pro						
Saccharomyces cerevisiae	8	28	15	27	17	13								
Wheat	9	29	16	26	14									
Mammals	10	29	4	21										
Schizo. pombe	11	32	25	eu-eu										
Euglena (cytoplasm)	12	31												
S. cerevisiae mitochondria	13													

eu and pro, eukaryotic and prokaryotic sequences, respectively.

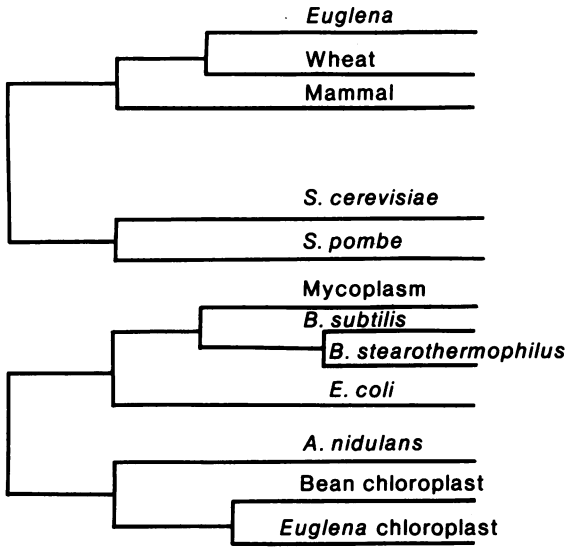


FIG. 2. Phylogenetic tree derived from the $tRNA^{Phe}$ sequences.

criteria, the *Euglena* sequence should diverge from the mammalian branch (see Table 1). The convergence test provides a quantitative counterargument to this biologically odd solution: when, in fact, *Euglena* is attached near mammals, the resulting protoeukaryotic sequence converges less (24.7 compared to 22.2) to the protoprokaryotic sequence than does the tree in Fig. 2. Although *Euglena* is placed here with the plants, there is little to choose between this position and a branching prior to the divergence of plants and animals based on convergence or number of mutations. Similarily, *Schizo. pombe* could be branched prior to the yeast divergence without changing the results significantly.

Attempts were made to integrate the mitochondrial sequence into the phenylalanine tree. First, from mutational distances in Table 1, this tRNA is not obviously related to either eukaryotes or prokaryotes. In order to establish a possibly very ancient relationship, we experimented with various possible topologies. When positioned as the first divergence in either the prokaryotic or the eukaryotic branch, the mitochondrial sequence did not improve convergence between the two branches. In addition, the mitochondrial sequence is no closer to the sequence resulting from the combination of the prokaryotic and eukaryotic ancestral sequences than it is to all other actual $tRNA^{Phe}$ sequences on the average.

Tyrosine tRNA family

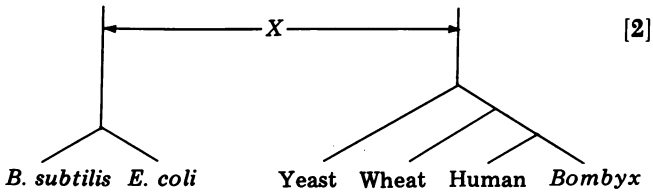
By use of the seven known sequences of $tRNA^{Tyr}$, which also has only one possible anticodon (four eukaryotic and three prokaryotic), the ancestral sequences were constructed and compared. Although there is an average of 32.0 differences in interbranch comparisons, the ancestral sequences differ by 32.9 positions. This lack of convergence could indicate that the eukaryotic and prokaryotic branches have independent origins. In support of this proposal, we note that there is no heterologous aminoacylation of either the prokaryotic or eukaryotic species by aminoacyl-tRNA synthetases isolated from the opposite branch (16); more significantly, the prokaryotic sequences possess a large variable loop‖ not found in eukaryotes. This feature is unique to $tRNA^{Tyr}$; no major length differences exist between the prokaryotic and eukaryotic branches of any other tRNA family having the same anticodon. Again, as in the preceding example, the mitochondrial $tRNA^{Tyr}$ appears unrelated to the other tyrosine tRNAs although it is structurally somewhat closer to the prokaryotic sequences and possesses the large variable loop typical of prokaryotic $tRNA^{Tyr}$. It is possible that when other mitochondrial sequences become available a convergence of mitochondrial and prokaryotic ancestors will be evident.

‖ Differences in the variable loop cannot account for the lack of convergence because these positions are omitted from the analysis.

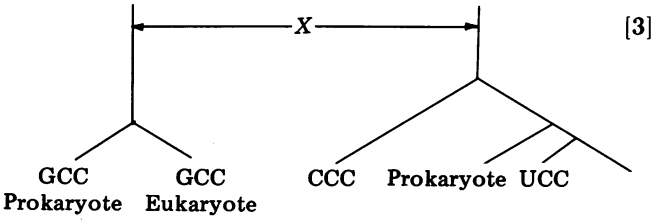
Glycine tRNA family

Unlike the preceding families, the analysis of the glycine tRNA family is complicated by the existence of three isoacceptor species having GCC, CCC, or UCC as the anticodon sequence. In the following studies, we have not considered the $tRNA^{Gly}$ sequences from *Staphylococcus epidermis*, which do not participate in protein synthesis and are not, therefore, subject to the same evolutionary pressure as other tRNAs. The eukaryotic $tRNA^{Gly}_{CCC}$ and $tRNA^{Gly}_{UCC}$ were eliminated as well because these sequences are distant from other $tRNA^{Gly}$ and, since they are single sequences, no ancestral sequence can be calculated. The single prokaryotic $tRNA^{Gly}_{CCC}$ was used, however, because it is closely related to the three known prokaryotic $tRNA^{Gly}_{UCC}$ sequences and could be converged with them. In all, the 10 remaining glycine tRNA sequences are rather well distributed among prokaryotes and eukaryotes as well as representing the remaining isoacceptor activities.

Our analysis started with the prokaryotic and eukaryotic $tRNA^{Gly}_{GCC}$ as arranged in [2]. These sequences are shown to be related; the distance X between the prokaryotic and eukaryotic ancestor is 23.7, whereas the average interbranch distance is 26.7.



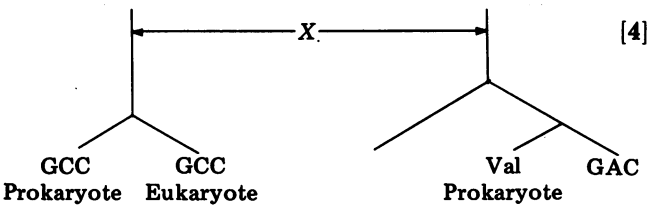
Although the $tRNA^{Gly}_{CCC}$ and $tRNA^{Gly}_{UCC}$ sequences could be related to this tree in two ways, if at all, a first experiment showed that the ancestor of these tRNAs does not converge toward either the protoprokaryotic or the protoeukaryotic $tRNA^{Gly}_{GCC}$ sequences considered separately. The alternate topology in [3], however, showed significant convergence:



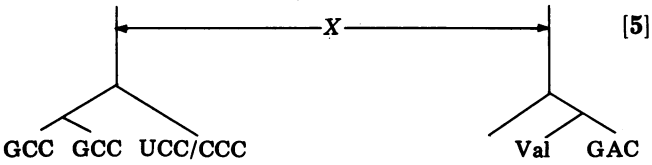
The ancestral sequences shown above in [3] differ in 23.9 positions, whereas the average distance between these two groupings is 26.8. This topology clearly shows that all $tRNA^{Gly}$ share a common ancestor and that the differentiation at the first anticodon position occurred between purines and pyrimidines, before the divergence of prokaryotes and eukaryotes. Because not enough data from eukaryotic $tRNA^{Gly}_{UCC}$ and $tRNA^{Gly}_{CCC}$ are available, it is impossible to assign the correct chronology to the distinction between the eukaryotic UCC and CCC anticodon-containing tRNAs with respect to the divergence of eukaryotes and prokaryotes.

Relationship of $tRNA^{Gly}$ to $tRNA^{Val}_{GAC}$

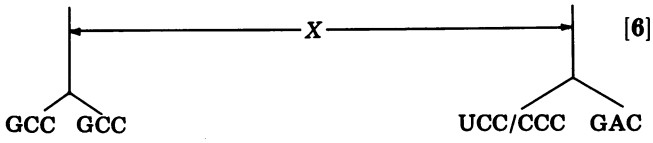
Because it had been previously noted (9, 17) that prokaryotic $tRNA^{Val}_{GAC}$ and $tRNA^{Gly}_{GCC}$ were quite similar (falling in the lowest 1% of the comparisons of tRNAs having different activities shown in Fig. 1), the ancestral sequence of the $tRNA^{Val}_{GAC}$ was compared with those previously constructed for the $tRNA^{Gly}$ tree. Here again many different topologies are possible if, indeed, there is a relationship between members of these two families. First, the convergence of the $tRNA^{Val}_{GAC}$ ancestor to either eukaryotic or prokaryotic $tRNA^{Gly}_{GCC}$ ancestors was tested, but gave a negative result. However, the experiment depicted in [4] showed that the ancestral proto-$tRNA^{Gly}_{GCC}$ does converge toward the $tRNA^{Val}_{GAC}$ ancestor, $X = 23.2$ in contrast to 28.0 average.



A comparable amount of convergence is obtained as in [5], when all $tRNA^{Gly}$ are first used to construct a common ancestor which, when compared to the $tRNA^{Val}_{GAC}$ ancestor, differs in only 25.5 positions in contrast to 29.6 average.



A corollary of this arrangement is that the combination of the $tRNA^{Val}_{GAC}$ ancestor and the $tRNA^{Gly}_{CCC}/tRNA^{Gly}_{UCC}$ ancestor should converge toward the $tRNA^{Gly}_{GCC}$ ancestor as in [6]. Indeed, the most significant convergence is found, $X = 21.2$ in contrast to 27.7 average distance.



We conclude therefore that $tRNA^{Gly}$ and $tRNA^{Val}_{GAC}$ are members of one homologous superfamily (Fig. 3). Any speculations on the origin of $tRNA^{Val}_{IAC}$ must await the availability of new sequence data because preliminary tests with available data do not indicate any clear option.

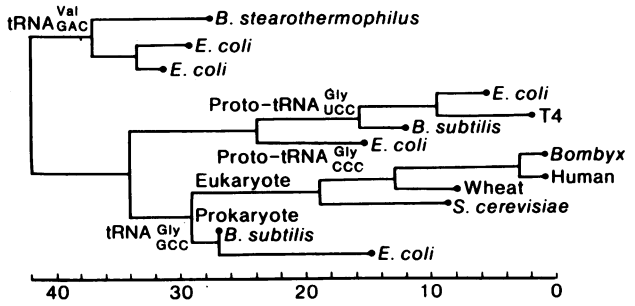


FIG. 3. Phylogenetic tree of the $tRNA^{Gly}/tRNA^{Val}$ superfamily. The horizontal distance represents the actual number of structural changes between nodal or actual sequences.

Discussion

The convergence test, although developed here for the specific purpose of examining tRNA evolution, is universal in its scope and design because it is based on the essentially divergent nature of molecular evolution. In principle, it can be applied to detect homology between any group of related sequences of the same type, be they RNA, DNA, or protein. The choice of tRNAs as the subject of this investigation, justified by their intimate link to the genetic code evolution, is a hazardous one. Indeed, tRNAs most likely constitute a macromolecular family that approaches mutational equilibrium and whose history involves speciation and gene duplications. Consequently, their phylogeny should be most difficult to decipher. Nevertheless, the identification of $tRNA^{Gly}$ and $tRNA^{Val}_{GAC}$ as members of the same homologous superfamily shows that some aspects of the evolutionary history of the genetic code may be deduced. The multifamily tree (Fig. 3) can still be interpreted in many ways because it is not possible to rigorously determine the position of the root. It seems most probable, however, to assume that the earliest evolutionary event was the divergence of the valine (anticodon GAC) from the glycine tRNA family. The two amino acids have quite different properties, and it is hardly conceivable that they could be interchanged in a modern-type protein without destroying its function. Alternately, it is always possible that one of the two tRNAs changed both its anticodon and acceptor activity after the initial gene duplication, thus avoiding the misreading of an already existing code. Also, it is possible, even probable, that other tRNA families are related to the same root. Unfortunately, not enough sequences of other likely families are available for this analysis.

One major finding is that the three present-day $tRNA^{Gly}$ isoacceptor types are derived from two ancient protosequences (which themselves have a common ancestor): one branch has produced the $tRNA^{Gly}$ with the GCC anticodon which, according to the wobble hypothesis (18), should read the two glycine codons ending in a pyrimidine; the second branch has produced a prototype which subsequently diverged into two $tRNA^{Gly}$ each of which, again following the wobble rules, should read only one glycine codon having a terminal purine.

This filiation nicely complements recent protein translation studies showing that all four valine (19) or glycine (20) codons can be read by a single isoacceptor species *in vitro* and that third-codon position misreading can occur *in vivo* (21). These results have been analyzed in terms of a "two out of three" reading mechanism (22) which may, in fact, be a vestige of a previous genetic code. The distinction between the purine- and pyrimidine-reading tRNAs, as indicated by the $tRNA^{Gly}$ tree, would seem to have been the first step in the refinement of this early genetic code.

Consideration of the $tRNA^{Val}_{GAC}/tRNA^{Gly}$ tree and, in particular, the $tRNA^{Gly}_{GCC}$ part, where the eukaryotic/prokaryotic divergence is established, leads to another major conclusion: namely, that the mutational fixation rate varies in each branch of the tree, since the branch lengths in Fig. 3 are drawn proportional to the mutational distance. Disregarding these different rates during sequence comparisons can lead to erroneous conclusions: for example, the $tRNA^{Val}_{GAC}$ differs very little from the prokaryotic $tRNA^{Gly}_{GCC}$. One would thus be tempted to cluster these families together, but, in fact, the closeness of these two groups of tRNA is more representative of extremely slow mutation rates in these two branches than a recent common ancestor. Different mutation rates may also explain the inconsistencies in the eukaryotic $tRNA^{Phe}$ tree where *Euglena* is close to the mammalian sequence and *S. pombe* is inordi-

nately distant from other eukaryotic sequences. As shown in a recent work (23), *S. pombe* $tRNA^{Phe}$ is the only member of this family in which multiple mutational changes have touched the D stem and extra loop regions: in fact, the rest of the sequence is more highly conserved than in other eukaryotes. Finally, the effect of gene splicing in eukaryotic tRNA is difficult to evaluate because, at least for the $tRNA^{Phe}$ family, no particular anomalies have been found that would indicate independent origin of the two branches of the family even though this tRNA is known to be spliced (24).

From the results presented here we would like to propose a general model based on the $tRNA^{Val}/tRNA^{Gly}$ tree which, although tentative, does explain the available tRNA structure data. We propose that each branch and each tRNA family has a different mutation fixation rate. Eukaryotic tRNAs seem to have incorporated natural mutations much faster than prokaryotic tRNAs. The many structural differences between eukaryotic and prokaryotic tRNAs emanate, therefore, mostly from the rapid tRNA evolution in the emerging eukaryotes. A corollary of this hypothesis is that tRNA evolution may be most dramatic among the lower eukaryotes or Protista. Although based on sketchy data, mitochondrial tRNA could stabilize mutations even faster than eukaryotic tRNA. This high rate would necessarily make it extremely difficult to place mitochondrial tRNAs in evolutionary schemes as attempted above. A cursory comparison between mitochondrial and cytoplasmic initiator methionine tRNAs from *Neurospora crassa* and *S. cerevisiae* lends some credibility to the above hypothesis; mitochondrial sequences differ at 26 positions and the cytoplasmic sequences at only 16 positions.

In conclusion, tRNA sequences, if used with care in a stepwise approach, can yield significant information about the origins of the translation apparatus. Particularly interesting will be the analysis of the leucine, serine, and arginine tRNA families. These amino acids are encoded by six codons, and important knowledge about the basic structure of the genetic code may be obtained.

We thank Ester Cloutier for technical assistance and the many investigators who have furnished tRNA sequence data for this and other studies, in particular, Drs. S. Clarkson, J. Dahlberg, G. Dirheimer, H. Ishikura, G. Keith, W. McClain, K. Murao, J. Olins, U. L. RajBhandary, and D Söll. This work was funded by the National Research Council of Canada. H.G. received a travel grant from the Ministère de l'Education Nationale and the Fonds National de la Recherche Scientifique Belge.

1. Crick, F. H. C. (1968) *J. Mol. Biol.* **38**, 367–378.
2. Jukes, T. H. (1974) *Nature (London)* **246**, 22–26.
3. Wong, J. T.-F. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 2336–2340.
4. Bauer, K. (1976) *Int. J. Pept. Protein Res.* **8**, 13–19.
5. Barricelli, N. A. (1977) *J. Theor. Biol.* **67**, 85–109.
6. Ishigami, M., Nagamo, K. & Tonotsuka, N. (1977) *Biosystems* **9**, 229–243.
7. Eigen, M. & Schuster, P. (1978) *Naturwissenschaften* **65**, 341–369.
8. Wetzel, R. (1978) *Origins Life* **9**, 39–50.
9. Holmquist, R., Jukes, T. H. & Pangburn, S. (1973) *J. Mol. Biol.* **78**, 91–116.
10. Hasegawa, N. (1978) *Origins Life* **9**, 495–500.
11. Cedergren, R. J., Cordeau, R. J. & Robillard, P. (1972) *J. Theor. Biol.* **37**, 209–215.
12. Margoliash, E. & Fitch, W. M. (1967) *Science* **155**, 279–284.
13. Dayhoff, M. O., Park, C. M. & McLaughlin, P. J. (1972) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 5, p. 7.
14. Gauss, D., Grüter, F. & Sprinzl, M. (1979) *Nucleic Acids Res.* **6**, 81–89.
15. Sankoff, D. & Rousseau, P. (1975) *Math. Programming* **9**, 240–251.
16. Jacobson, K. B. (1971) *Prog. Nucleic Acid Res. Mol. Biol.* **11**, 461–488.
17. Clarke, L. & Carbon, J. (1974) *J. Biol. Chem.* **249**, 6874–6885.
18. Crick, F. H. C. (1966) *J. Mol. Biol.* **19**, 548–555.
19. Mitra, S. K., Lustig, F., Akesson, B. & Lagerkvist, V. (1977) *J. Biol. Chem.* **252**, 471–478.
20. Bergquist, P. L., Burns, D. J. & Plinston, C. A. (1968) *Biochemistry* **7**, 1751–1760.
21. Parker, J., Pollard, J. W., Friesen, J. D. & Stanners, C. D. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1091–1095.
22. Mitra, S. K., Lustig, F., Akesson, B., Axburg, T., Elisa, P. & Lagerkvist, U. (1979) *J. Biol. Chem.* **254**, 6397–6401.
23. LaRue, B., Cedergren, R. J., Sankoff, D. & Grosjean, H. (1979) *J. Mol. Evol.*, in press.
24. Valenzuela, P., Venegas, A., Weinberg, F., Bishop, R. & Rutter, W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 190–194.