How Old Is the Genetic Code? Statistical Geometry of tRNA Provides an Answer

Manfred Eigen,* Björn F. Lindemann, Manfred Tietze,
Ruthild Winkler-Oswatitsch, Andreas Dress, Arndt von Haeseler

The age of the molecular organization of life as expressed in the genetic code can be estimated from experimental data. Comparative sequence analysis of transfer RNA by the method of statistical geometry in sequence space suggests that about one-third of the present transfer RNA sequence divergence was present at the urkingdom level about the time when archaebacteria separated from eubacteria. It is concluded that the genetic code is not older than, but almost as old as our planet. While this result may not be unexpected, it was not clear until now that interpretable data exist that permit inferences about such early stages of life as the establishment of the genetic code.

It is a general belief that life originated on Earth, yet there is no direct experimental evidence about when and where life actually came about. Hence, any logically consistent hypothesis would seem to be acceptable (*1, 2*). Recent progress in sequence analysis of nucleic acids, however, opens a new experimental access to the problem. In particular, the nearly 1000 transfer RNA sequences that are known today (*3*) offer, because of their pronounced similarities of primary sequence and common features of secondary and tertiary structure, fertile ground for comparative analysis (Fig. 1).

There are two ways of ordering the wealth of sequence data. First, one may focus on a given species and produce an alignment of all its individual tRNAs. Fifteen such species families and their respective consensus sequences (Table 1) have been identified, with each family consisting of 20 to 40 individual tRNAs. Second, one may specify a tRNA by its anticodon and follow it through various species, as has been done in 24 phylogenies of individual sets of tRNAs with common anticodon (Table 2). Each anticodon family contains 15 to 30 individual tRNA sequences from all kingdoms. All sequences listed in Tables 1 and 2 are explicitly given in (*3*).

Correspondingly, there are two types of evolutionary divergence, each of a different topological nature. They show up as bundle-like diagrams for parallel divergence in species families (Fig. 2, left) and as treelike branching patterns for consecutive divergence in anticodon families (Fig. 2, middle). For sequences with ancient origins, both parallel and reverse mutations have accumulated to yield looped or netlike diagrams (Fig. 2, right). It is therefore important to devise an objective method for establishing the correct topology.

The two forms of divergence project to different origins. Phylogenies are rooted in the last common ancestor of kingdoms, whereas the spread within each species is rooted at the origin of the genetic code, or more accurately, at the time of "fixation" of the various tRNAs, when amino acids became assigned to their codons. There are therefore two essentially different distances that may, depending on a calibration of evolutionary rates, allow an establishment of time spans or their limits. In particular (as is suggested in Fig. 3), if it is possible to assign sequences that refer to ancestors of kingdoms and to determine their spread, we may compare divergence at notably different times of divergence.

Distance as a Measure of Kinship

Kinship relations are revealed by alignment of sequences. Such an alignment is relatively simple for tRNA because of its uniform structure. A measure of kinship then is the distance d_{ik}, the number of positions at which two aligned sequences i and k are occupied by different nucleotides. Such "horizontally" summed (that is, over the length of the tRNA chain) overall distances are meaningful measures of divergence only if they are sufficiently different from zero and from their limiting value of complete randomization (that is, $d_{ik} = 0.5\nu$ or 0.75ν, where ν is the number of nucleotides compared,

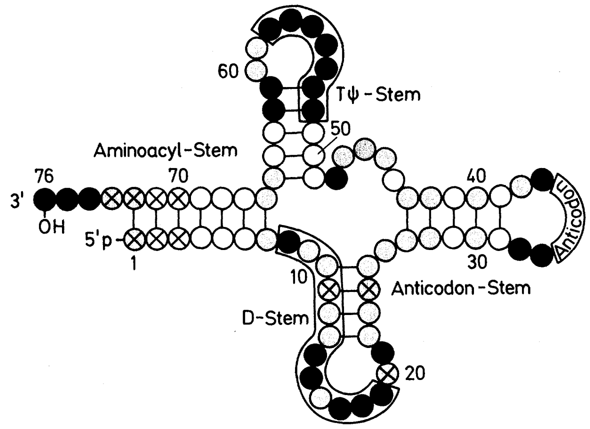
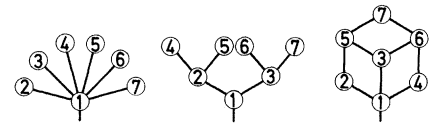


Fig. 1. Secondary structure of the 76 common positions of tRNA. Counting base pairs as one position there are, apart from the anticodon, 52 positions available for comparative analysis. Details described below refer to later parts of the article: three categories of variability [on the basis of distinguishing the purine nucleotides (R = A or G) from the pyrimidine nucleotides (Y = U or C)] are indicated: constant (black), moderately diverged (gray), and highly diverged (white). Two control regions for transcription of eukaryotic genes by polymerase III, 5'-ICR and 3'-ICR, are indicated by frames (*13*). Major identity elements for recognition of four tRNAs by their cognate amino acyl synthetases are marked by crosses (*11, 12*).

M. Eigen, B. F. Lindemann, M. Tietze, and R. Winkler-Oswatitsch are at the Max-Planck-Institut für biophysikalische Chemie, D-3400 Göttingen, Federal Republic of Germany. A. Dress and A. von Haeseler are in the Fakultät für Mathematik, Universität Bielefeld, D-4800 Bielefeld, Federal Republic of Germany.

*To whom correspondence should be addressed.



①	RRYRRY	RRYRRY	RRYRRY
②	YRYRRY	YRYRRY	RYYRRY
③	RYYRRY	RYYRRY	RRRRRY
④	RRRRRY	YRRRRY	RRYYRY
⑤	RRYYRY	YRYYRY	RYRRRY
⑥	RRYRYY	RYYRYY	RRRYRY
⑦	RRYRRR	RYYRRR	RYRYRY

Fig. 2. Examples of "bundle" (left) and "tree" (middle) topologies representing different types of evolutionary divergence, as compared with a "net" (right) resulting from parallel and reverse mutations. (Each line segment represents a change at one position.) In quaternary sequences there are further alternatives of introducing loops.

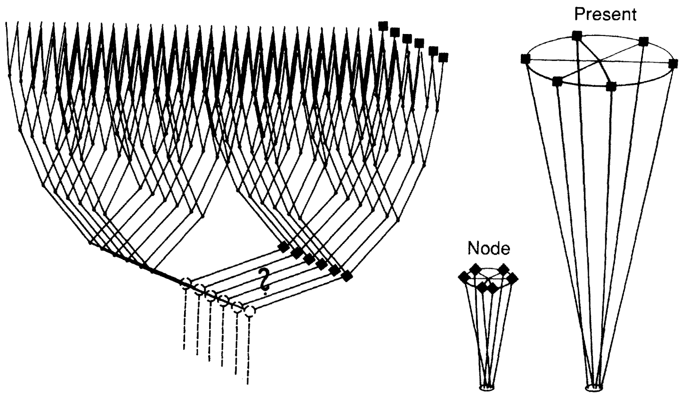


Fig. 3. Construction of phylogenetic trees for several anticodon families and reconstruction of sequences at common branching points (reference nodes) allows one to compare divergence of tRNAs in present species with that in ancestors of present kingdoms, thereby providing data that project to different points on the time axis.

for binary or quaternary sequences, respectively), and if all positions on average have the same probability of changing. In the following, we focus on binary sequences, counting only transversions, that is, changes from purines (R = A or G) to pyrimidines (Y = U or C) or vice versa.

Distances can also be counted vertically (*4–6*). We start from alignment of all sequences of a species family and record for each position the nucleotide (R or Y) that appears most frequently, thereby defining a consensus sequence. As a measure of randomization, we count for every position i the number δ_i of nucleotides that differs from the consensus nucleotide. The frequency distribution $f(\delta)$ for appearance of positional distances δ then provides information about the uniformity of divergence. In a similar way, as the horizontal distance d_{ik} is related to the number of positions ν in the sequences compared, the vertical distance δ_i depends on n, the number of sequences in the alignment. When $\delta_i = 0.5n$ for binary sequences, one sees complete randomization. When more than two symbols are involved (for example, quaternary sequences), vertical distances have to be modified by statistically weighting the different types of substitution, as in information science is done in computing the Shannon-entropy (*7*).

Histograms for $f(\delta)$ as a function of δ were obtained by computer simulation in which uniform probabilities of change were assumed (Fig. 4). A set of $n = 30$ initially identical binary sequences (unshaded vertical box at $\delta = 0$) was subjected to random mutation. The sequences quickly assume a diffuse, but peaked distribution, characterized by an average value δ (compare the distribution at $\delta = 0$ with the shaded histogram at $\delta \approx 3$), that travels along the abscissa until it reaches a limiting value near $\delta \approx n/2 = 15$ (dark histogram), where it signals complete randomization. The experimental values for $n = 28$ (binary) tRNA sequences of *Bacillus subtilis* (open circles) are representative for all species families of tRNA studied so far. Among the 52 reference positions (counting base pairs as one position), 21 ± 2 are universally constant (not just being a tail of a diffuse distribution), whereas the remaining 31 ± 2 positions are spread over the entire abscissa, not matching any of the diagrams typical of uniform probabilities of transversion and fixation.

The diagram allows us to define the category of "constant" positions, positions that are obviously biased by present functional needs and therefore do not necessarily represent ancestral information. This makes it necessary to infer total randomization of ancestral information, if distances d_{ik} come close to $0.5(\nu - \nu_c)$ in binary sequences, where ν_c is the number of constant positions. To classify the 31 ± 2 variable positions further, we have compared: (i) 15 (binary) consensus sequences of present species families (Table 1), (ii) two sets of 24 individual sequences that refer to early nodes (that is, branching points) in the phylogenetic trees of their respective anticodon families (Table 2), and (iii) the (binary) consensus sequences I and II, referring to both sets of individual early nodal sequences (see Table 2 at the bottom).

Table 1. Consensus sequences of present tRNA species families. The $\bar{\delta}$ are average distances (vertical deviations of an individual from species family consensus sequences) determined separately for 21 moderately (*a*, shaded) and 10 highly (*b*, unshaded) diverged positions. For positions where assignment is uncertain, both R and Y being represented by nearly 50%, the first of the two symbols refers to the more abundant one. (Abbreviations: *Bac. subtil.*, *Bacillus subtilis*; Chloro, chloroplasts; *Eugl.*, *Euglena*; Mito., mitochondria; *Asp. nid.*, *Aspergillus nidulans*; *Halob. volc.*, *Halobacterium volcanii*, *Meth. vanni.*, *Methanococcus vannielli*, and Bov., bovine.) The sequences and the numbering of positions are from (*3*). Base pairs are considered one position (listed by the symbol that appears first if counted from 5′ to 3′). Positions that are constant in individual sequences are not included. (Note that constant positions in this table refer to consensus sequences.)

	1	2	3	4	5	6	7	9	10	11	12	13	16	20	26	27	28	29	30	31	38	44	45	46	47	49	50	51	59	60	73	$\bar{\delta}_a/n$ (%)	$\bar{\delta}_b/n$ (%)
Bac. subtil	R	R	R	R	R	R_Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	Y	R	R	R	Y	R	R	Y	R	R_Y	R	R	Y	R	17	41
E. coli	R	R	R	R	R	Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R	Y	R	R	R	R	Y	R	R	R	R	Y	R	18	43
Phage T_4/T_5	R	R_Y	R	R	R	Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R_Y	Y	R	R	R	R	Y	R	Y	R	R	Y	R	23	41
Chloro. (plants)	R	R_Y	R	R	R	Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R	Y	R	R	R	R	Y	R	Y	R	R	Y	R	19	40
Chloro. (*Eugl.*)	R	Y	R	R	R	Y	R	R	R	Y	Y	Y	Y	Y	R	Y	R	R	R	R	R	R	R	R	Y	R	Y	R	R	Y	R	18	42
Mito. Yeast	R	R	R	R	R_Y	Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	Y	R	Y	R	Y	R	R	-	R	Y	R	R	Y	R	19	39
Asp. nid.	R	R	R	R	R_Y	R_Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	Y	R	R	R_Y	R_Y	R	R	-	R	Y	R	R	Y	R	23	40
Human/Bov.	R	R	R	R	R	Y	R	R	R	Y	Y	Y	Y	-	R	Y	Y	R	R	R	R	R	R	R	-	R	Y_R	R	Y	Y	R	21	36
Mouse	R	R	Y	R	R	Y	R	R	R	Y	Y	Y	Y	-	R	Y	Y	R	R	R	R	R	R	R	-	R	R	R	-	Y	R	21	36
Halob. volc.	R	Y	R	Y	Y	R	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R_Y	R	R_Y	R	Y	R	R	Y	R	Y	R	R	Y	R	20	40
Meth. vanni.	R	Y_R	Y_R	Y	Y	R_Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R	R	R	Y_R	R	R	Y	R	Y	R	R	Y	R	15	35
Yeast	R	Y	Y	Y	Y	Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R	R	R	R	R	R	Y	R	Y	R	Y_R	Y	R	17	34
Drosophila	R	Y	Y	Y	Y_R	Y_R	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	Y	R	R	Y	R	R	R	Y	Y	Y	R	R_Y	Y	R	18	33
Mouse/Rat	R	Y	Y	Y	Y	R	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R	R	R	R	R	R	Y	Y	Y	R	R	Y	R	19	31
Human/Bov.	R	Y	Y	Y	Y	R	R	R	R	Y	Y_R	Y	Y	Y	R	Y	Y	R	R	R	R	R	R	R	Y	R	Y	R	R	Y	R	17	29

The consensus sequences in Table 1 record the symbol appearing most frequently in each vertical row. If R and Y appear with (almost) equal frequencies, both symbols are included, the first representing the more abundant one. Phylogenetic trees were constructed by a modified "parsimony" method (*8*) that has been described elsewhere (*9*). The nodes I and II refer to two defined tripods (dendrograms of three sequences). Node I represents the tripodal node of eubacteria, mitochondria, and archaebacteria; node II represents the tripodal node of eubacteria, archaebacteria, and eukaryotes. The individual nodal sequences at all positions in Table 2 where nodes I and II do not agree are shown with the first symbol node I and with the second symbol node II. The two consensus sequences correspondingly refer to the first (I) and second (II) symbol, respectively. In assigning the nodal nucleotide, we gave each of the three kingdoms (including mitochondria) equal weight. In the trees of individual anticodon families, methanogens and thermoacidophiles occur as early diverging members of an archaebacterial kingdom. Likewise, chloroplasts appear as relatives of cyanobacteria in the eubacterial kingdom, while mitochondria form a separate group that is more closely related to eubacteria than to archaebacteria (*10*).

The 31 variable positions represented in Tables 1 and 2 may be subdivided into two groups called *a* (shaded) and *b* (nonshaded). The 21 shaded class *a* positions consistently show a lower degree of variation than do the 10 nonshaded class *b* positions. (This is true for the correspondingly shaded positions in Fig. 1.) This fact is reflected in the δ values quoted in Table 2 for each individual position (δ_i) and in Table 1 and in the lower part of Table 2 for the averages ($\bar{\delta}_a$ and $\bar{\delta}_b$), which are taken—separately for both classes—over all individual sequences of each species (or node) family. Whereas $\bar{\delta}_b$ values almost reach the limiting values characteristic of complete randomization ($0.4n$ to $0.5n$), the $\bar{\delta}_a$ values show only half as much divergence ($\sim 0.2n$). We used this criterion for defining classes *a* and *b*. For early nodal families, $\bar{\delta}_a$ is smaller by up to a factor of 3 than corresponding values for present species families, whereas $\bar{\delta}_b$ remains close to the limit of randomization. We correspondingly call class *a* positions "moderately" and class *b* positions "highly" diverged.

At class *a* positions, all consensus sequences are practically identical, as to be expected from theory (*6*) for binary sequences with $\bar{\delta} < 0.3n$. Therefore, most of the differences among consensus sequences are limited to the ten highly diverged class *b* positions. Nevertheless, the distances between consensus sequences clearly reflect phylogenetic kinship and yield a meaningful evolutionary tree in which nodes I and II, constructed solely from consensus sequences (Table 1), agree exactly with those obtained independently from phylogenies (Table 2). This suggests, and is confirmed by inspecting phylogenetic divergence, that during phylogeny class *b* positions have not changed to an appreciably larger extent than class *a* positions. Their status of divergence obviously reflects a situation that existed before phylogeny.

Variable positions (including class *b*) encode functions that, at a quite early stage of evolution, require discrimination of individual tRNAs (Fig. 1) (*11, 12*). Nondiscriminative functional elements, such as the boxes acting as initiation signals for polymerase III in eukaryotes, are localized preferably in constant or moderately diverged regions (*13*).

Statistical Geometry in Sequence Space

We are now in a position to proceed with comparative analysis according to the program outlined in the first section. For distance comparison, we use a combination of vertical and horizontal analysis introduced recently (*6*) and called "statistical geometry in sequence space."

Consider a space consisting of 2^ν points, each of which represents one of the 2^ν possible binary sequences ν positions long. The points are arranged in a geometry that correctly represents kinship relations. For $\nu = 3$, for example, the sequences would be assigned to the $2^3 = 8$ corners of a cube. For ν positions, we would have a ν-dimensional hypercube, in which the ν one-error mutants of each sequence would occupy the ν corner points of the ν axes starting from the point representing that sequence. The concept of relating messages or (aligned) sequences to one another by counting the number of characters by which they differ was introduced in information theory by Hamming (*14*). In a geometrical representation of it, the measure of separation or metric that counts distances like street blocks is called a "Hamming metric." Throughout this article, we relate distances to the total number of the positions that are under consideration and express them as percentages of this number. For binary sequences, a relative distance of 50% means total randomization, since it is equivalent to the distance expected in any random distribution of binary symbols.

Now consider a quartet of binary sequences A, B, C, and D located at four distinct points in sequence space. The four points can always be connected as to yield diagrams as shown in Fig. 5. The seven distance segments in these diagrams (that is, three box dimensions and four protrusions) represent the sizes of seven classes of positions where the four sequences are not congruently occupied. If we denote the segments as in the examples shown in Fig. 5, the lengths of the protrusions from the box to A, B, C, or D represent the numbers of positions in which, by vertical comparison, one sequence differs from the three others (for example, $A \neq B = C = D$). The box dimensions represent the numbers of positions exhibiting one of the three possible pairwise congruences (that is, $A = B \neq C = D$, $A = C \neq B = D$, and $A = D \neq B = C$).

The method becomes a "statistical" one by the application of this concept to all $\binom{n}{4}$ quartets that can be formed from *n* sequences [for example, for $n = 40$: $\binom{n}{4} = \frac{n!}{4!(n-4)!} = 91390$] and the determination of statistical averages of all geometrical segments. Individual protrusions then cannot be distinguished anymore, and we define d_1 as the number of positions with three congruently occupied sequences (that is, "three of a kind") and d_2 as the number of positions with pairwise congruences ("two pairs"), which leaves $d_0 = \nu - d_1 - d_2$ as the number of homologous positions (that is, "four of a kind"). Statistically sound averages of d_0, d_1, and d_2 can be given for any sufficiently large set of sequences.

An objective decision about the topology of divergence can be obtained by analyzing relations between the three pairwise congruences subsumed under d_2. In the statistical analysis we average separately for the largest ($\bar{l}$), medium ($\bar{m}$), and smallest ($\bar{s}$) of these three box dimensions. For an ideal bundle, $\bar{l}$, $\bar{m}$, and $\bar{s}$ must be zero. If all three box dimensions are of a similar length, we are dealing

Table 2. Early nodal sequences of individual tRNAs. Two sets of data, one referring to the common ancestor (tripodal node) of eubacteria, archaebacteria, and mitochondria (I), the other to that of eubacteria, archaebacteria, and eukaryotes (II), have been established. Where assignments differ, the first of two paired symbols refers to tripod I, the second to tripod II. Moderately diverged (class *a*) positions again are shaded. Consensus sequences of individual early nodal sequences I and II, relative positional deviations δ_i, are shown in the lower part of the table. For the consensus sequences called master I and master II, respectively, $\bar{\delta}_a/n$ is 6.7 and 9.2%, $\bar{\delta}_b/n$ is 37.9 and 38.3%. In spite of its special role in translation, the initiator tRNA (mti) reveals positional class assignments and noise levels that are similar to those of all other tRNAs.

	1	2	3	4	5	6	7	9	10	11	12	13	16	20	26	27	28	29	30	31	38	44	45	46	47	49	50	51	59	60	73
Ala (UGC)	R	R	R	*RY*	Y	Y	R	R	R	Y	Y	Y	Y	R	R	Y	Y	Y	*RY*	Y	R	R	R	R	Y	R	R	R	R	Y	R
Arg (ACG)	R	Y	R	Y	Y	Y	R	R	R	Y	Y	Y	Y	R	R	Y	Y	Y	R	R	R	R	R	R	Y	R	R	R	R	Y	R
Asp (GUC)	R	*RY*	Y	Y	Y	R	R	R	R	Y	*YR*	Y	Y	Y	R	Y	*YR*	R	R	Y	Y	R	R	R	Y	R	Y	R	R	Y	R
Asn (GUU)	R	Y	Y	Y	Y	Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	Y	R	R	R	R	R	R	Y	R	Y	R	R	Y	R
Cys (GCA)	R	*RY*	Y	R	R	*YR*	R	R	R	Y	R	R	Y	Y	R	Y	Y	R	R	*RY*	R	R	R	Y	Y	R	Y	Y	R	Y	Y
Gln (UUG)	Y	R	R	R	Y	Y	R	R	R	Y	Y	R	Y	Y	R	Y	Y	R	R	R	Y	Y	R	Y	R	R	Y	Y	R	Y	R
Glu (UUC)	R	Y	Y	Y	*YR*	*YR*	*RY*	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	*YR*	Y	Y	*YR*	R	R	-	*RY*	R	R	R	Y	R
Gly (GCC)	R	Y	R	R	*RY*	R	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	Y	Y	*RY*	*RY*	R	R	Y	*RY*	Y	R	R	Y	R
Gly (UCC)	R	Y	R	*RY*	*RY*	*YR*	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R	Y	R	Y	R	R	-	*RY*	Y	R	Y	Y	R
His (GUG)	R	Y	*RY*	R	*YR*	*YR*	R	R	R	Y	R	Y	Y	*YR*	Y	Y	Y	R	R	*RY*	R	*YR*	R	R	Y	R	Y	R	Y	Y	Y
Ile (GAU)	R	R	R	Y	Y	*YR*	R	R	R	Y	Y	Y	Y	Y	R	Y	*RY*	Y	R	*YR*	R	R	R	R	Y	R	R	Y	R	Y	R
Leu (CAA)	R	Y	R	R	R	*YR*	R	R	R	Y	*RY*	R	Y	Y	R	R	Y	R	R	R	R	Y	Y	Y	Y	R	Y	R	R	Y	R
Leu (UAG)	R	Y	R	Y	R	*YR*	R	R	R	Y	Y	R	Y	Y	R	Y	Y	R	R	R	R	Y	R	R	Y	R	Y	R	R	Y	R
Leu (UAA)	R	Y	R	R	R	R	R	R	R	Y	Y	R	Y	Y	R	Y	Y	R	R	R	R	Y	*RY*	R	Y	R	Y	R	R	Y	R
Lys (UUU)	R	R	R	Y	Y	R	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	Y	R	R	R	Y	R	R	Y	R	R	R	R	Y	R
Lys (CUU)	R	R	R	Y	Y	R	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	*YR*	R	R	R	R	R	R	Y	R	R	R	R	Y	R
Met (CAU)	R	*RY*	Y	Y	R	*YR*	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R	R	R	R	R	R	Y	R	Y	R	R	Y	R
Mti (CAU)	R	R	Y	R	R	R	R	R	R	R	R	Y	Y	*YR*	R	Y	Y	R	R	R	R	R	R	R	Y	R	*YR*	*RY*	R	Y	R
Phe (GAA)	R	Y	Y	R	R	*RY*	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	*RY*	R	R	R	R	Y	R	Y	R	Y	Y	Y	Y	R
Pro (UGG)	R	R	R	Y	Y	Y	R	R	R	Y	Y	Y	Y	Y	*RY*	Y	*YR*	*YR*	R	R	R	R	R	R	Y	*RY*	Y	*RY*	R	Y	R
Thr (UGU)	R	Y	Y	*RY*	*RY*	R	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	Y	R	R	R	*RY*	R	R	Y	R	R	R	R	Y	R
Trp (CCA)	R	R	R	R	Y	Y	R	R	R	Y	Y	Y	Y	Y	R	*YR*	Y	*RY*	R	*YR*	R	*RY*	R	R	Y	R	R	R	R	Y	R
Tyr (GUA)	R	*RY*	R	R	*RY*	R	R	R	R	Y	Y	Y	Y	Y	R	R	Y	R	R	R	R	Y	*RY*	R	Y	R	Y	Y	R	Y	R
Val (UAC)	R	R	R	R	R	Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	Y	R	Y	R	R	R	R	Y	R	R	Y	R	Y	R
Master I	R	R	R	R	*RY*	Y	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R	R	R	R	R	R	Y	R	Y	R	R	Y	R
δ_i (I)	1	11	8	10	12	9	0	0	0	1	4	5	0	2	1	2	1	9	2	8	3	9	2	3	1	0	9	6	3	0	2
Master II	R	Y	R	Y	Y	R	R	R	R	Y	Y	Y	Y	Y	R	Y	Y	R	R	R	R	R	R	R	Y	R	Y	R	R	Y	R
δ_i (II)	1	9	9	11	10	8	1	0	0	1	4	5	0	4	2	3	2	9	2	8	4	10	4	3	1	4	10	8	3	0	2

with a randomized bundle, for which the relative magnitude of $\bar{d}_2$ with respect to $\bar{d}_1$ is a measure of the degree of randomization.

If one of the box dimensions is significantly larger than the two others we are dealing with a tree-like topology. For an ideal tree, $\bar{m}$ and $\bar{s}$ must be zero. Hence the relative magnitude of $\bar{m}$ and $\bar{s}$ with respect to $\bar{l}$ is a measure of "tree-likeness." If $2\bar{l} - (\bar{m} + \bar{s}) \approx 0$, the tree structure is randomized to an extent that it cannot be distinguished any more from a (randomized) bundle. The examples shown in Fig. 5 include a tree-like diagram with ancient divergence and two bundle-like diagrams where $\bar{l}$, $\bar{m}$, and $\bar{s}$ are of a similar magnitude.

At this point we may note an important difference between sequence and distance space. The classical analysis, aimed at the construction of dendrograms, is based on overall distances (*8, 10*). Four sequences define six distances, but a dendrogram involves only five segments. Hence, there is generally ambiguity in assigning one of the three possible four-sequence trees, requiring a best compromise (for example, maximum parsimony) for constructing the tree. This ambiguity does not exist in assigning quartets in sequence space, whose particular geometry always yields exact assignments, thus quantifying the degree of uncertainty of tree assignment. Concomitant recording of different box segments provides, in addition, an internal calibration that is more sensitive to large divergences than are absolute values of cumulative distances. Moreover, as the segments are averages of a very large number of quartets, they are statistically quite robust. (The values generated are not entirely statistically independent. In a set of 30 sequences, each pair contributes to about 1.6% of the quartets.)

Returning to the data analysis, we show in Fig. 6 a computer simulation of parallel divergence of 30 sequences, each comprising 30 positions, submitted to random mutations with uniform average rate. The mutation distance Δ counts the average number of mutational events per sequence, irrespective of their being of a forward, parallel, or reverse nature. It is to be distinguished from both horizontal and vertical distances d and δ that can be counted in every alignment. The term Δ/ν should be considered some equivalent of (relative) time. [In the sequence analysis literature, it is often related to $\nu = 100$ and called PAM, meaning "accepted point mutations" per 100 residues (*15*).] When $\Delta/\nu = 1$, on average every position had a chance to change, but not all did, while others were reversed. Ancestor families (Table 2) are shown separately for moderately (class *a*) and highly (class *b*) diverged positions.

All class *b* data (including ancestor families) refer to Δ/ν values near one, where kinship relations are almost completely randomized (Fig. 6). The class *a* data, on the other hand, generally show much smaller degrees of randomization. The divergence is bundle-like with a small residual tree-likeness indicated by (small) differences in the Δ positions obtained for d_1 and d_2. All but mitochondrial and T-phage species families show comparable degrees of randomization, defining coincident Δ_a values (the index *a* referring to class *a* positions). Mitochondria and T phages project to somewhat larger Δ_a values. For class *a* positions, the two ancestor families (nodes I and II) show considerably smaller degrees of randomization, defining a smallest value Δ_{min} for the earliest node at $\Delta_{min} \leq (0.35 \pm 0.1)\Delta_a$. (The "earliest node" itself cannot be determined objectively, since we cannot assign a tripod to it. The error limit given refers to the spread of data and the uncertainty of assignment of class *a* and class *b* positions.)

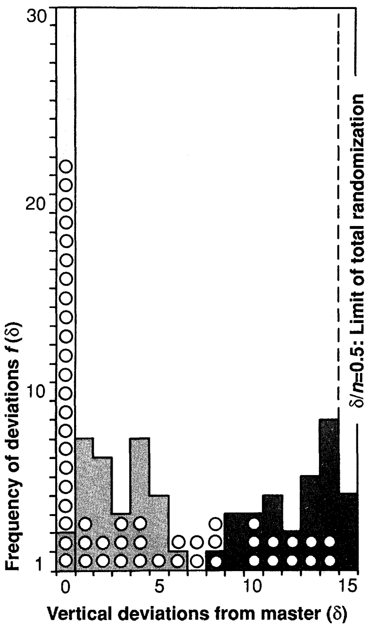


Fig. 4. Superposition of simulated and real histograms of vertical divergence. The abscissa counts the deviations δ from the consensus symbol in vertical rows of an alignment, whereas the ordinate records $f(\delta)$ the frequency of δ values observed. First ignore the open circles representing experimental data. Three stages of bundle-like divergence of 30 binary sequences, each comprising 30 positions, are simulated on a computer as a random-walk model with uniform substitution probabilities. The three histograms refer: (i) to time 0 (unshaded bar at δ = 0, representing the initial stage of identity of all 30 sequences); (ii) to an intermediate time, when divergence still is small (shaded bars); and (iii) to infinite time when divergence is complete, that is, when all kinships are randomized (dark bars). Now consider the experimental data for a typical present-day tRNA species family, namely, 28 binary (RY) sequences of *Bacillus subtilis* (52 reference positions). There are 22 positions (for all species families studied: 21 ± 2 positions) that seem to be still in their initial state (or have been biased to be identical). They do not form any "tail" of the distribution of the remaining 30 (generally 31 ± 2) positions, which themselves do not match any of the stimulated histograms, indicating highly nonuniform probabilities of substitition.

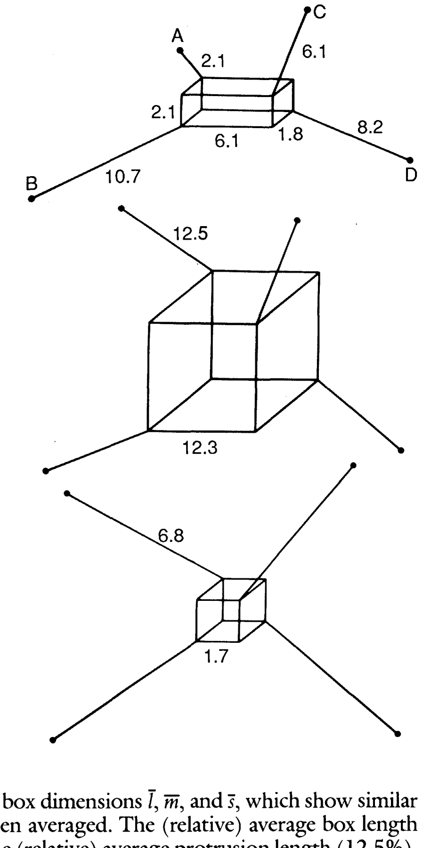


Fig. 5. Examples of average geometries of tRNA in binary sequence space. (**Top**) Average tree-like diagram referring to the anticodon CAU family (methionine initiator) involving four different taxa with early divergence: eubacteria (A), mitochondria (B), archaebacteria (C), and eukaryotes (D). Three members of each taxon were chosen, yielding $3^4 = 81$ dendrograms, from which the average segment lengths (protrusions to A, B, C, or D and box dimensions $\bar{l}$, $\bar{m}$, and $\bar{s}$ given as percentage of $\nu = 31$ positions) were determined. The Hamming distance between any two sequences can be easily constructed (for example, $\bar{d}_{AC} = 2.1 + 6.1 + 1.8 + 6.1 = 16.1$). (**Middle**) Average diagram of $\binom{28}{4} = 20475$ quartet combinations of the highly diverged class *b* positions in the species family of 28 tRNAs of *Bacillus subtilis*. The divergence is almost completely randomized, as indicated by the large average box dimensions $\bar{l}$, $\bar{m}$, and $\bar{s}$, which show similar lengths and therefore have been averaged. The (relative) average box length (12.3%) is about as large as the (relative) average protrusion length (12.5%). The average Hamming distance between two sequences is 2 × 12.5 + 2 × 12.3 = 49.6%. (**Bottom**) Average diagram of $\binom{24}{4} = 10626$ quartet combinations of the 21 moderately diverged class *a* positions in 24 sequences of an ancestor family (node II, the tripod of eubacteria, archaebacteria, and eukaryotes). The diagram represents an almost ideal bundle. The average lengths of protrusions are much larger than the average box dimensions $\bar{l}$, $\bar{m}$, and $\bar{s}$ which themselves turned out to be of similar size, so that a cube with edge length $(\bar{l} + \bar{m} + \bar{s})/3$ was constructed. The average Hamming distance between two sequences is 2 × 6.8 + 2 × 1.7 = 17%.

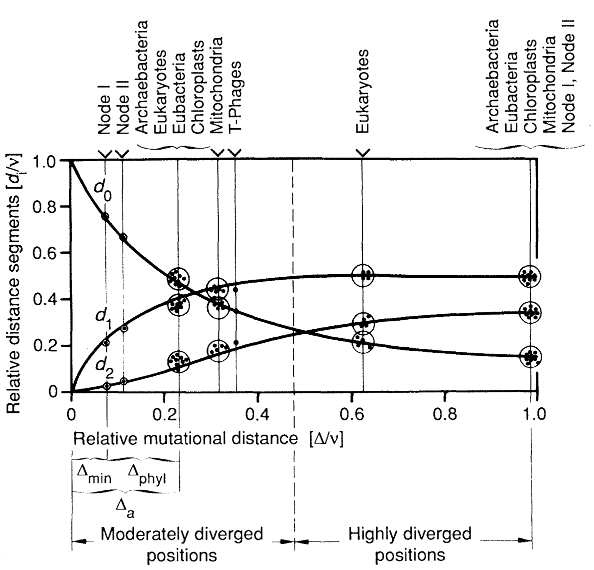


Fig. 6. The three lines describe a computer simulation of the three average distance segments of geometries of quartet combinations of RY sequences (statistical averages), assuming parallel divergence with uniform substitution rates. The experimental values (dots) shown are for the three average segments $\bar{d}_0$, $\bar{d}_1$, $\bar{d}_2$ of present and early tRNA species families. All experimental segments $\bar{l}$, $\bar{m}$, and $\bar{s}$, that contribute to $\bar{d}_2$ were almost of the same magnitude, identifying the divergence as "bundle-like" with small residual tree-likeness. The experimental values in the left half of the picture refer to 21 moderately divergent and those in the right half to 10 highly diverged positions. (The circles refer to the spread of data.) In a simulated bundle divergence, residual tree-likeness expresses itself in too small $\bar{d}_1$ and too large $\bar{d}_2$ values relative to their average and δ/ν positions. A small residual tree-likeness is recognizable for all groups in bundle diagrams. Phylogenetic divergences (Δ_{phyl}) identify themselves by drastic deviations. Values for Δ referring to those divergences had to be determined separately from diagrams simulating tree-like divergence. Note that mitochondria show larger divergences at class *a* positions than eubacteria, whereas chloroplasts do not.

If one of the positions is incorrectly assigned to class *a* or class *b*, respectively, Δ_{min} decreases or increases by less than 10% of its value. Uncertainties of assignment of nodal sequences in Table 2 refer mainly to class *b* positions and hence do not affect Δ_{min} values greatly.

Data for d_0, d_1, and d_2 of the anticodon families do not match the curves simulated for bundle-like divergence in Fig. 6. We have therefore simulated a consecutive tree-like divergence, which could be matched with data of phylogenetic divergence of the kind shown in the upper diagram of Fig. 5. [Eighty-one quartets of eubacterial, mitochondrial, archaebacterial, and eukaryotic sequences yield a Δ_{phyl} value of $(0.65 \pm 0.1)\Delta_a$.] A relation $\Delta_a = \Delta_{min} + \Delta_{phyl}$, as indicated in Fig. 6, is approximately fulfilled. (This is not necessarily the case if any of the Δ values refers to nearly complete randomization, as is true for class *b* positions.)

Quaternary Sequences

So far we have treated nucleic acids as binary sequences of purines and pyrimidines. In part, this was with heuristic intent, for the concepts of sequence space and statistical geometry are most transparent for binary systems. Counting only transversions, that is, changes between purines and pyrimidines, does not bias comparative analysis and is acceptable as long as it provides information about divergence. In fact, the loss of information by neglecting transitions ($G \rightleftharpoons A$ and $C \rightleftharpoons U$) is compensated by more accurate distance assignments; transitional distances, especially for mitochondrial sequences, are highly randomized. In this section the logical basis provided by the binary model is used to anlayze quaternary sequences.

A quaternary sequence space (*6*) may be constructed by assigning to each corner point in the ν-dimensional hypercube (representing a binary sequence) a subspace that itself is a hypercube of dimension ν. Assignment then proceeds in two successive steps. Localization in the first hypercube defines the RY sequence, whereas localization in the second hypercube (subspace) specifies which of the R or Y bases is involved at each position. (The concept could be generalized further so as to include deleted or inserted symbols.)

Quartets of quaternary sequences establish distance relations that involve four different subspaces of the RY hypercube. They include five classes of distance segments. In addition to the three classes present in binary systems, we have "one pair" (d_3) and "no pair" (d_4). The segment d_4 involves all four bases and belongs therefore to the class d_2 in RY space. Segment d_3 has two possible assignments (for example, GGAC = RRRY and GGUC = RRYY) and therefore is split among the RY distance classes d_1 and d_2.

We have shown that one may again use representative geometries to deal with quartets of quaternary sequences. The segments d_3 and d_4 introduce triangular, tetrahedral, or similar polyhedral structure elements (*6*). However, a disadvantage of the five distance classes defined is that they do not differentiate between transitions and transversions, which may contribute with different rates of substitution.

To cope with this problem, we have to suitably combine the five distance classes in AUGC space with the three distance classes in RY space, yielding altogether eight new distance segments. We use symbols d_{qb} where *q* specifies the five distance segments in quaternary (AUGC) space and *b* refers to the three distance segments in binary (RY) space that count transversions only. The binary distance class d_0 (for example, RRRR) thereby splits up into three subclasses: d_{00}, d_{10}, and d_{20} (examples, GGGG, GGGA, and GGAA, respectively). Likewise the d_1 of binary sequence space (for example, RRRY) splits up into d_{11} and d_{31} (examples, GGGU and GGAU, respectively), whereas d_2 (for example, RRYY) has three subclasses: d_{22}, d_{32}, and d_{42} (examples, GGUU, GGUC, and GAUC, respectively).

Statistical geometry in binary sequence space yields the rate of divergence exclusively due to transversions. Is there a similar way to determine a rate referring exclusively to transitional divergence? We must find a hypercube that is uniquely representative of transitional changes, and we must compare quartets of sequences that are localized in this subspace.

For this purpose we define a subspace of RY space that includes only those coordinates that refer to homologous positions (that is, "four of a kind": RRRR or YYYY) of a given quartet. In this subspace the four sequences are represented by the same point. We then specify transitions by assigning a d_0-dimensional transitional subspace to this point in the d_0-dimensional RY-subspace. We call the second subspace "transitional" because any extension in this space is exclusively due to transitional changes. Hence, we may compare geometries in this space with those obtained earlier in RY space. Note, however, that the number of coordinates differs in both spaces. All distances in the transversional hypercube (and hence any Δ value obtained) refer to ν coordinates, while those in the transitional subspace (including the resulting Δ value) refer to d_0 coordinates and have to be normalized correspondingly.

Carrying out such an analysis, we find different mutation distances for transitional and transversional changes. If substitution probabilities for transversion and transition were equal, the Δ/ν values for transversional changes would be twice as large as the Δ/d_0 values describing transitional changes in the d_0 class (because a given nucleotide can mutate two ways as a transversion but only one way as a transition). The experimental data, however, show larger values for transitions than for transversions. The best fits of all experimental data are obtained if a transition is assumed to occur 2.5 ± 0.3 times more frequently than an individual transversion. The simulations whose results are shown in Fig. 7 were carried out on this basis. The Δ/ν values now include two transversions plus one transition that occurs 2.5 times more frequently than one of the two transversions. Class *a* divergence (Fig. 7) therefore appears at Δ/ν values that are about $1/2(2 + 2.5) = 2.25$ times larger than those found in Fig. 6, and they congruently fit all curves, including those that reflect mixed transitional and transversional divergences. A 10% variation of the ratio of transitional to transversional substitution rates yields perceptible deviations. Note that some of the segments, in particular those which pass through maxima, react very sensitively to the ratio of transitional and transversional rates. If the ratio were taken to be one, the relative magnitudes of some distance segments would even invert.

The class *b* positions again appear to be completely randomized. Note that $\Delta/\nu \approx 1$ is equivalent to a value near 0.5 in Fig. 6. Hence, the limiting values for randomization lie outside the frame of Fig. 7. Moreover, in class *b* positions the ratios of transitional and transversional rates may differ from those valid for class *a* positions. The distances d_{00} and d_{42} clearly indicate that randomization is complete for class *b* positions.

Altogether we obtain from the quaternary analysis shown in Fig. 7 a picture of divergence in species families similar to that from the binary analysis shown in Fig. 6. In quaternary sequence space, kinship relations (based on class *a* positions) are appreciably more randomized ($\Delta/\nu \approx 0.48$) than in RY space ($\Delta/\nu \approx 0.23$), yet they still appear to be well defined. Mitochondrial sequences, on the other hand, are appreciably more randomized, to such an extent that it is difficult to reconstruct early ancestors on the basis of the available data. Anticodon families again show Δ/ν values that lie between 0.65 and 0.75 of those for present species families, confirming the results obtained from RY analysis.

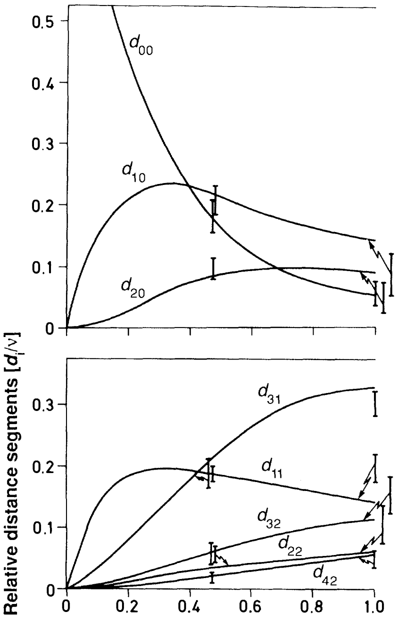


Fig. 7. The eight lines describe a computer simulation of the eight average distance segments of 30 A, U, G, C sequences, each comprising 30 positions, assuming different rates of transition and transversion. The distance segments are defined in the text. The ratio of individual rates of transition to transversion in the simulations is 2.5 ± 0.3. Experimental values match to a common Δ/ν value of 0.48 for all distance segments of moderately variable positions. Distance segments for highly diverged positions show large fluctuations around values that correspond to complete randomization. The bars indicate the maximal spread of data. They are placed so as to yield an optimal compromise for all eight distance segments, which is reached far outside the frame at the right-hand side of the picture.

Conclusions

Individual and master sequences of tRNA reflect kinship relations that are consistent with generally accepted evolutionary patterns (*16–18*) and allow comparative analysis to be extended into a prephylogenetic time range. The evolutionary spread of individual tRNAs such as that found in various species families does not drastically exceed phylogenetic divergence of given tRNAs. The qualitative conclusion follows that the origin of the genetic code did not predate early phylogenetic diversification (for example, of eubacteria and archaebacteria) to a considerable extent. Had the code been much older—and this would be possible only in case of extraterrestrial origin—those changes that clearly can be identified as phylogenetic divergence would previously have become randomized to a large extent. The study reported in this article quantifies this statement by making a detailed comparative analysis based on the method of statistical geometry in sequence space.

Twenty-one "moderately" diverged positions (Fig. 1) are the main source of evolutionary information. Their binary sequences are homologous in nearly all master sequences of present and precursor families. Statistical geometry shows that their evolution was divergent: early precursors appear appreciably less diverged and blurred than present families. This is not true for the ten highly diverged positions whose very early divergence provided the diversity necessary for discriminative code adaptors. All these positions are located in the 3′ half of the chain while only seven of them, being at complementary positions in the aminoacyl and anticodon stem regions, can also be associated with the 5′ half. This fact is consistent with (although not a particularly strong argument in favor of) speculations about the 3′ half having been a precursor of tRNA (*19, 20*). Transitional changes in tRNA have accumulated 2.5 ± 0.3 times as fast as transversions. The ratios of divergence Δ_{min}/Δ_a of early precursors and present families are about 0.35 ± 0.1, while the ratios of phylogenetic divergence to present family divergence Δ_{phyl}/Δ_a are near 0.65 ± 0.1. This result, obtained from statistical geometry, is supported by the δ values from vertical analysis quoted in Tables 1 and 2. The main kingdoms, on the basis of relative distance segments (degree of randomization), appear to be "equally old." Descending from a common ancestor, they had an equal span of time available to accumulate "noise." Their divergence is essentially bundle-like [with a small residual tree character (*21*)].

The sequence data yield relative mutation distances, which could be converted to time ratios only if one could assume substitution rates to be time-independent. In the very early phases of evolution, however, error rates and acceptability of mutations must have been larger than in present organisms, where structures are optimized and error rates are minimized. Hence, ratios of mutation distances that refer to such early periods only represent upper limits, which strengthens our argument that the genetic code is younger than our planet.

Errors of assignment to positional classes (that is, constant, moderately, and highly diverged) are limited in general to ±2 positions, possibly differing in different species families. Furthermore, these assignments as such are only one possible attempt of quantifying the nonuniformity of evolutionary substitution rates. The main source of errors, however, is the uncertainty in establishing and dating earliest nodes. If early nodes of kingdom separation are to be dated around 2.5 ± 0.5 billion years ago (*7, 17, 22*), the code cannot be older than 3.8 (±0.6) billion years.

REFERENCES AND NOTES

1. L. E. Orgel, *The Origins of Life* (Wiley, New York, 1973).
2. F. C. H. Crick, *Life Itself: Its Origin and Nature* (Simon and Schuster, New York, 1981).
3. M. Sprinzl, T. Hartmann, F. Meissner, J. Moll, T. Vorderwulbecke, *Nucleic Acids*

Res. **15**, R53 (1987).
4. M. Eigen and R. Winkler-Oswatitsch, *Naturwissenschaften* **68**, 217 (1981).
5. R. Winkler-Oswatitsch, M. Eigen, A. Dress, *Chem. Scr.* **26B**, 59 (1986).
6. M. Eigen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 5913 (1988).
7. C. L. Manske and D. J. Chapman, *J. Mol. Evol.* **26**, 226 (1987).
8. W. M. Fitch, *Am. Nat.* **111**, 223 (1977).
9. H. J. Bandelt and A. Dress, *Adv. Appl. Math.* 7, 309 (1986).
10. G. E. Fox, K. R. Luehrsen, C. R. Woese, *Zentralbl. Bakteriol. Mikrobiol. Hyg. 1. Abt. Orig. C* 3, 330 (1982).
11. Y. M. Hou and P. Schimmel, *Nature* **333**, 140 (1988).
12. L. H. Schulman and J. Abelson, *Science* **240**, 1591 (1988).
13. S. J. Sharp, J. Schaak, L. Cooley, D. J. Burke, D. Söll, *CRC Crit. Rev. Biochem.* **19**, 107 (1985).
14. R. W. Hamming, *Bell Syst. Tech. J.* **29**, 147 (1950).
15. M. O. Dayhoff, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC, 1979).
16. G. E. Fox *et al.*, *Science* **209**, 457 (1980).
17. H. G. Schlegel, *Allgemeine Mikrobiologie* (Thieme, New York, ed. 6, 1985).
18. J. Hopfield, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4334 (1987).
19. M. Eigen and R. Winkler-Oswatitsch, *Naturwissenschaften* **68**, 282 (1981).
20. W. M. Fitch, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 759 (1987).
21. H. Hori and S. Osawa, *Mol. Biol. Evol.* **4**(5), 455 (1987).
22. G. J. Olsen, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 825 (1987) (compare T. Cavalier-Smith, *ibid.*, p. 807i).
23. We thank Dr. W. C. Gardiner for reviewing and correcting the manuscript. We also thank the referees for their helpful comments.

Stereochemistry of RNA Cleavage by the *Tetrahymena* Ribozyme and Evidence That the Chemical Step Is Not Rate-Limiting

JAMES A. MCSWIGGEN AND THOMAS R. CECH

The intervening sequence of the ribosomal RNA precursor of *Tetrahymena* is a catalytic RNA molecule, or ribozyme. Acting as a sequence-specific endoribonuclease, it cleaves single-stranded RNA substrates with concomitant addition of guanosine. The chemistry of the reaction has now been studied by introduction of a single phosphorothioate in the substrate RNA at the cleavage site. Kinetic studies show no significant effect of this substitution on k_{cat} (rate constant) or K_m (Michaelis constant), providing evidence that some step other than the chemical step is rate-limiting. Product analysis reveals that the reaction proceeds with inversion of configuration at phosphorus, consistent with an in-line, S_N2 (P) mechanism. Thus, the ribozyme reaction is in the same mechanistic category as the individual displacement reactions catalyzed by protein nucleotidyltransferases, phosphotransferases, and nucleases.

THE NUCLEAR PRECURSOR TO RIBOSOMAL RNA IN *Tetrahymena thermophila*, a ciliated protozoan, contains a 413-nucleotide intervening sequence (IVS). The IVS excises itself from the larger RNA in a protein-independent reaction called self-splicing (*1*).

Self-splicing occurs through two transesterification reactions (exchanges of phosphate esters which leave the total number of phosphodiester bonds unchanged). In the first transesterification, the 5′ splice site is cleaved as guanosine is added to the 5′ end of the IVS (*2*); guanosine, GMP, and GTP (guanosine mono- and triphosphate, respectively) have similar activities as substrates. It has been proposed that the guanosine acts as a nucleophile in an in-line, S_N2 (P) reaction (*3, 4*), but there has been no direct test of this mechanism. In the second transesterification step, the 3′ splice site is cleaved as the exons (RNA sequences flanking the IVS) are joined.

The excised IVS RNA retains catalytic activity (*5, 6*). Truncated versions of the IVS RNA act as RNA enzymes (ribozymes) to cleave, join, or dephosphorylate RNA substrates (*7–9*). The sequence-specific endoribonuclease activity of the *Tetrahymena* ribozyme (Fig. 1) is an intermolecular version of the first step of preribosomal RNA self-splicing. The site of substrate cleavage is determined by a base-pairing interaction; the substrate binds to the same sequence within the IVS that specifies the 5′ splice site during self-splicing (*9–11*). The endoribonuclease reaction facilitates detailed studies of the chemistry of guanosine addition because the substrate can be provided as an oligonucleotide. Oligonucleotides are easily synthesized with a variety of sequences and functional group substitutions.

Recently we studied cleavage of substrates that had single-base changes several bases preceding the cleavage site, giving mismatched substrate-ribozyme complexes. Surprisingly, mismatches greatly enhanced the rate of cleavage (*12, 13*). One reasonable explanation was that the mismatches might be facilitating a rate-limiting conformational change (*14*) rather than affecting the chemical step. Such a model gives a strong prediction; an alteration of the phosphate at the cleavage site that greatly decreases its reactivity toward O-P bond cleavage might have very little effect on the rate of the reaction. This prediction has now been tested by substitution of a phosphorothioate at the reaction site. The finding of very little change in cleavage rate, even with the best substrates, supports the model of a non–rate-limiting chemical step.

The phosphorothioate-containing substrates also provide a test for the stereochemical course of the reaction. We find that the reaction proceeds with inversion of configuration at phosphorus, the same result obtained for most proteins that catalyze transesterification of phosphate esters (*15, 16*). This stereochemistry suggests

J. A. McSwiggen is an Associate and T. R. Cech is an Investigator of the Howard Hughes Medical Institute, Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309–0215.