Evolution of the Transfer RNA Molecule

Semih Erhan, Larry D. Greller, and Barbara Rasco

Department of Animal Biology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia

(Z. Naturforsch. **32 c**, 413–418 [1977]; received January 10/February 17, 1977)

Evolution, Transfer RNA, Sequence Homology, Computers

Base sequences of many transfer RNA (tRNA) species obtained from different sources contain homologous regions. These homologies, which are 6 to 20 nucleotides long, occur both within the same tRNA molecule and between many different tRNA molecules repeatedly. Since it is very unlikely an 80 or so nucleotide long tRNA molecule could have been formed at once, under primordial conditions, we propose that the homologous oligonucleotides found within the tRNA molecules to-day represent the earliest adapter from which tRNA molecules have evolved.

Introduction

Since the pioneering experiments of Miller [1], many steps leading to the emergence of life have been elucidated; however, the evolution of transfer RNA (tRNA) molecule is still not understood. This is in marked contrast with the great breakthroughs that have been achieved in understanding the genetic code, in sequencing several tRNA species and in obtaining three dimensional structure of some tRNA molecules.

This lack of understanding of tRNA evolution is essentially due to two basic problems:

1) The first problem, which relates to the formation of all nucleic acids, is the difficulty in producing nucleosides under promordial conditions [2, 3].

2) The second problem more specifically relates to the origin of the translation phenomenon and thus to the origin of tRNA molecules. The questions to which answers must be found were:

a. How and when did the tRNA molecule emerge, as the adapter we know to-day?, and

b. can an 80 or so nucleotide long polymer be formed at once or does it represent, in its present form and size, the product of many evolutionary steps?

The first question, to which unfortunately no experimental answer can be provided, is really equivalent to asking whether the interaction between protein- and nucleic acid-like molecules occurred directly or there was always, from the beginning, an adapter molecule whose function has always been to pick an amino acid and to bring it to its proper place.

Requests for reprints should be sent to S. Erhan, 2101 Chestnut Street, *Philadelphia*, PA. 19103.

We believe it is more realistic to think that the earliest interactions between these polymers, or more precisely between one polymer and the monomer of the other, occured directly and that an adapter has later become necessary when "proto-organelles", – the earliest structural elements within the primordial cells – emerged. With this event various activities began to occur on well separated, relatively rigid structures and hence the cellular constituents could not stream through the cell as a dilute solution, and some of them had to be transported through the cell by primitive adapters.

The second question is easier to answer, at least theoretically, because it can be broken down into two related questions:

i. What is the probability of the formation of an 80 or so nucleotide long polymer, having a particular base sequence, from 4 nucleotides?

ii. How likely is it to build an 80 or so nucleotide long polymer, starting with one monomer, by the addition of one monomer at a time, even without considering the degradation which most likely will occur?

The probability of forming, from 4 nucleotides, an 80 nucleotide long polymer having a particular base sequence is $4^{-80} = 1.7 \times 10^{-49}$. Considering that more than one of these had to be available to accomodate the then available amino acids, the likelihood of their having been formed within the first few hundred million years appears very doubtful.

Regarding the second question, Simon, in dealing with a related question commented on the unlikelyhood of the formation of a 100 amino acid long polypeptide at once and suggested that if the smaller

peptides were to combine to yield larger and larger polypeptides then the probability of the formation of the final polymer would be much greater [4].

So using a similar argument we propose that the present size of the tRNA molecules represent the product of later stages in the evolution of a primordial tRNA like molecule, which was much smaller, perhaps as small as 1/4 th or even 1/8 th the size of the present tRNA molecule. In this communication we are going to present evidence for this idea on the basis of comparisons of the base sequences of various tRNA molecules.

Methods

A comparison of the base sequences of two nucleic acids is very similar to a comparison of the amino acid sequences of two proteins. For this purpose similarity scores (m-scores) among various nucleic acid bases have to be derived. This was done, using Relatedness Odds Matrix of Dayhoff [5], in the same way amino acid similarity scores were obtained [6, 7]. Table I gives the similarity scores obtained as well as the equivalence of the modified bases.

Table I. Similarity scores (m) represent how similar each base is to one another. They are derived from Relatedness Odds Matrix, as described in the text.

a) Similarity scores (m) for nucleic acid bases

	A	C	G	U
A	9	4	5	6
C	4	9	2	6
G	5	2	9	3
U	6	6	3	9

b) Equivalence of some unusual bases found in tRNA

		Equivalent to
D	5,6 dihydrouridine	U
Q	pseudouridine	U
T	ribosyl thymine	U
S	thiouridine	U
I	inosine	G

The comparison is then performed by a computer using a sliding match between twe two sequences [6]. One sequence, called "the key", is kept stationary while the other, called "the target", moves one nucleotide at a time, the similarity of the vertical base pairs are computed and printed. The printout is then scanned by the eye and any segment where approximately 50% of the bases are found to be identical between the two nucleic acids is considered likely to be significant. For the ease of detection of such regions the score 9, which represents identical vertical base pairs, is printed as a (′).

In order to be able to select those homologies that are significant, a "significance threshold" has to be established [6]. This threshold provides the borderline beyond which the observed homologies could have occured by chance. For homologies among proteins, where similarities among 20 amino acids are to be considered, we had demonstrated that a probability of 10^{-3} or better is a significant cut-off limit [6, 8, 9]. With nucleic acids where only 4 bases are found, the probability level has to be raised by an order of magnitude to 10^{-2} or better.

When a potentially significant homology is found the individual similarity scores (m) are simply added to obtain cumulative similarity score (M′) for the homology. Table II gives the correlation between the length of the homology (span length), the cumulative M score and the $P(M' \geqslant M)$, (the probability of the observed score to happen by chance). Here M′ is the cumulative score and M is the score which corresponds to a certain level of probability. If M′ is larger than a particular M score for a certain span length, say 5, then that homology is significant to the level of the M score exceeded. In other words the higher the M′ score

Table II. Cumulative similarity scores for varying span lengths and different levels of significance. Cumulative similarity score, M, is the sum of individual similarity scores (m), for a certain span-length, which measures how similar two sequences being compared are. The larger the M score the greater is the similarity. Span-length is the length of the homologous sequence. Negative orders of 10 give the levels of probability each M score corresponds to.

Span length	M							
	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}
4	34							
5	43							
6	49	52						
7	56	60						
8	62	67	70					
9	69	74	77	79				
10	75	81	85	88				
11	82	87	92	95	97			
12	88	94	99	103	106			
13	94	101	105	109	112	115		
14	101	107	112	117	120	123	124	
15	107	114	119	124	127	130	133	
16	113	120	126	130	135	138	141	142
17	119	126	132	137	141	145	148	151
18	125	133	139	144	148	152	156	158
19	132	139	145	151	155	159	163	166

the lesser is the probability that the observed homology could occur by chance.

Since we have coined the term "subsequence" for the homologies found among proteins [8], we shall continue to use subsequence also to indicate significant homologies found within and among tRNA molecules.

Results

If the premise that tRNA molecules have evolved from smaller oligonucleotides were true, then one should be able to see homologous subsequence throughout the entire length of tRNA molecules. Indeed homologous subsequences can be seen to occur within the entire length of tRNA molecules by matching the base sequence of a particular tRNA molecule against itself as well as matching the sequences of different tRNA molecules.

Table III shows the results of such a comparison obtained with alanyl tRNA ($tRNA^{Ala}$) against itself and Table IV shows the results obtained with leucyl tRNA ($tRNA^{Leu}$) against itself. It is important to point out here that many other homologies and many with longer span-lengths are also present within these molecules. Tables III and IV show only those with the highest statistical significance.

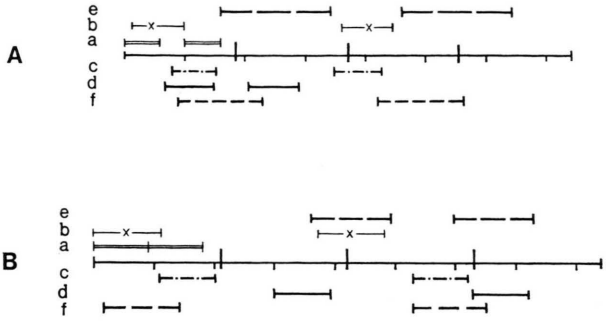

Fig. 1. The position of homologies found within an individual tRNA molecule. The molecule is drawn as a linear sequence beginning with the 5′ end on the left side. Small regular lines below the molecule represent each tenth base while larger lines divide the molecule into quarter lengths. A) $tRNA^{Ala}$, B) $tRNA^{Leu}$. Homologies are identified by letters alphabetically and correspond to the order of pairs given in Tables III and IV such that homologies designated (b) correspond to pairs 1—11 and 38—48 in Table IV.

Table III. Homologies found within $tRNA^{Ala}$. a) Position number indicates the position along the nucleic acid, of the sequence being matched, numbered contiguously from the 5′ end. Contiguous numbering is used to avoid confusion arrising from gaps introduced to obtain best homology. b) The numbers and the primes (′) below the target sequences represent the individual similarity scores (m), between the bases of two nucleic acids. M score shows the cumulative similarity score. Double matching probability gives the probability for such a matching to occur by chance. c) tRNA sequence is arbitrarily divided into four quarter lengths, which are numbered from 1 through 4, starting with the 5′ end of the molecule. 1 : 1 represents, for example, a homology found within the first quarter length of the molecule while 2 : 3 represents a homology found between 2nd and 3rd quarter lengths, see Fig. 1.

Homologous oligonucleotides found within alanyl tRNA ($tRNA^{Ala}$)

Position number of the oligonucleotide [a]	Base sequences of the homologous segment Individual scores [b]	M score of the match	Double matching probability $P(M' \geqslant M)$	Part of the tRNA it is found in [c]	Designation in Fig. 1
10—16	G C G C G U	47	10^{-3}	1 : 1	a
1— 6	G G G C G U ′ 2 ′ ′ ′ ′				
2—10	G G C G U G U G G	71	5×10^{-4}	1 : 3	b
36—44	I G C I Q G G G A ′ ′ ′ ′ ′ ′ 3 ′ 5				
8—14	U G G C G C G	60	10^{-4}	1 : 3	c
35—41	U I G C I Q G ′ ′ ′ ′ ′ 6 ′				
7—14	G U G G C G C G	68	5×10^{-5}	1 : 2	d
21—28	G D A G C G C G ′ ′ 5 ′ ′ ′ ′ ′				
16—33	A G D C G G D A G C G C G C U C C C	125	10^{-3}	2 : 3	e
46—63	A G U C U C C G G T Q C G A U U C C ′ ′ ′ ′ 3 2 6 5 ′ 6 3 ′ ′ 4 ′ 6 ′ ′				
9—22	G G C G C G U A G D C G G D	102	10^{-3}	1 : 3	f
42—55	G G A G A G U C U C C G G T ′ ′ 4 ′ 4 ′ ′ 4 3 6 ′ ′ ′ ′				

Table IV. The same internal homologies seen within $tRNA^{Leu}$.

Homologous oligonucleotides found within leucyl tRNA ($tRNA^{Leu}$)					
Position number of the oligo-nucleotide [a]	Base sequences of the homologous segment Individual scores [b]	M score of the match	Double matching probability P (M′ ≫ M)	Part of the tRNA it is found in [c]	Designation in Fig. 1
1— 9 10—18	G C G A A G G U G G C G G A A D D G ′ ′ ′ 5 ′ 5 3 ′ ′	67	10^{-3}	1 : 1	a
1—11 38—48	G C G A A G G U G G C G Q G Q U A G U G U C ′ 6 ′ 6 6 5 ′ ′ ′ 3 ′	80	10^{-3}	1 : 3	b
11—19 53—61	C G G A A D D G G C G G A C G U G G ′ ′ ′ ′ 4 3 ′ ′ ′	70	10^{-3}	1 : 4	c
31—39 63—71	G C U U C A G G Q G G T Q C A A G U ′ 2 ′ ′ ′ ′ 5 ′ ′	70	10^{-3}	2 : 4	d
37—49 61—73	G G Q G Q U A G U G U C C G G G G T Q C A A G U C C ′ ′ 3 ′ ′ ′ 4 5 6 ′ ′ ′ ′	99	10^{-4}	3 : 4	e
2—13 53—64	C G A A G G U G G C G G C G G A C G U G G G G G ′ ′ 5 ′ 2 ′ ′ ′ ′ 2 ′ ′	90	5×10^{-5}	1 : 3	f

Table V. Homologies found between different tRNA molecules.

Homologies found among different tRNA molecules					
tRNA	Position of the homology	Base sequences of the homologous segment Individual scores	M score of the match	Double matching probability P (M′ ≫ M)	Part of the tRNA it is found in
$tRNA^{Ala}$ (yeast) $tRNA^{Leu}$ (*E. coli*)	13—24 11—22	C G U A G D C G G D A G C G G A A D D G G D A G ′ ′ 3 ′ 5 ′ 6 ′ ′ ′ ′ ′	95	10^{-4}	1 : 1
$tRNA^{Ala}$ (yeast) $tRNA^{Phe}$ (yeast)	28—37 10—19	G C U C C C U U I G G C U C A G D D G G ′ ′ ′ ′ 4 2 ′ ′ ′ ′	78	5×10^{-4}	2 : 1
$tRNA^{Ala}$ (yeast) $tRNA^{Met}$ (yeast)	53—59 18—24	G G T Q C G A G G D D A G A ′ ′ ′ ′ 4 ′ ′	58	5×10^{-4}	4 : 2
$tRNA^{Ala}$ (yeast) $tRNA^{Trp}$ (*E. coli*)	1— 7 3— 9	G G G C G U G G G G C G S A ′ ′ ′ ′ ′ ′ 5	59	10^{-4}	1 : 1
$tRNA^{Ala}$ (yeast) $tRNA^{Ser}$ (brewers yeast)	14—20 76—82	G U A G D C G G U U G U C G ′ ′ 6 ′ ′ ′ ′	60	10^{-4}	1 : 4

An interesting point which emerges from these tables is that the strongest homologies are seen for span lengths which correspond to 1/8th length of the tRNA molecules: approx. 8 nucleotides long for $tRNA^{Ala}$ and approx. 10 – 11 nucleotides long lor $tRNA^{Leu}$. Since CCA occurs in all tRNA molecules, that sequence was not considered within the length of tRNAs.

Another point to be noted is that homologies are found between all quarters of the molecules as well as within the same quarter. Fig. 1 a and b demonstrate this point.

In Table V evidence is presented that these homologies are not only observed within a single tRNA molecule, but are found between many tRNA pairs.

Since there are also complementary bases along these oligonucleotides, a small degree of secondary structure may occur in solution.

Discussion

In the area of chemical evolution three schools of thought have evolved to account for the emergence of life. The first, based on the results of Fox with microspheres [10] and of Oparin with coacervate droplets [11] suggests that protein-like polymers alone were sufficient for life to appear under primordial conditions. According to this view the emergence of life is a gradual event, with competition for nutrients occuring at the microsphere stage and where the first "cells" may not even have to qualify to be alive as we understand life to-day. Most of the experimental evidence supports this view.

The second viewpoint starts with the premise that life without nucleic acids, in particular DNA, is inconceivable. The problem with this idea is that, even though nucleic acid bases can form hydrogen bonded complementary pairs, they can not replicate themselves. Indeed, so far no nucleic acid has been found to possess any catalytic activity, let alone such a complicated one as "replicase" activity. Thus one has to assume, even though it has never been stated clearly, that somehow a (bio)catalyst was conveniently available to perform this function and later to separate the replicated strands. We have recently elaborated on the relative importance of proteins and nucleic acids for a cell today and under prebiotic conditions (Erhan, manuscript submitted for publication).

The third view which is more like a reconciliation effort suggests that polypeptides and polynucleotides coevolved.

Regardless of which one of these alternative views one subscribes to, a direct amino acid-nucleotide interaction, be it one nucleotide per amino acid or three nucleotides per amino acid as we know today, appears to be the likely first step because it is impossible to conceive several 80 nucleotides long nucleic acid molecules to have been ready and waiting, then, to function as adapters. Lacey and Pruitt have shown, by model building, that it is quite feasible to accomodate trinucleotides along an α-helical polypeptide [12]. Woese, too, has produced indirect experimental evidence that aromatic compounds such as pyridine – actually it is estimated that 3 – 4 pyridine molecules are bound to each amino acid [13] – and even 2-picoline (Woese, personal communication) can bind to various amino acids.

How did this recognition occur between amino acids and trinucleotides is impossible to know exactly. However, such a recognition must certainly represent a very early event in the chemical evolution, regardless of which of the three alternatives listed above actually occured. This would correspond to a time early enough when amino acids and nucleic acid precursors, formed under energy impact on the primitive gases of the atmosphere, were still available in the oceans and where microspheres and similar structures provided relative stability from degradation.

The appearance of adapter molecules had to wait for the emergence of more elaborate structural elements within these "protocells", the "protoorganelles".

As the messenger RNA had to evolve to protect the DNA from protein synthesizing activities, so can we envisage the necessity for the emergence of adapter molecules, whose function was to carry the amino acids to the assembly site, from the site they were either synthesized or transported into the "cell". At that time great selectivity was certainly not necessary and the amino acids might simply have been held through hydrophobic interactions.

A few oligonucleotides, 8 – 10 nucleotides long and having even the anticodon region at one end, could easily have been produced under prebiotic conditions and among them those which could bind amino acids might have been concerved. The growth of the chain, then, could have occured at several steps in a way similar to the evolution of MDV-1 phage RNA described in detail by Wells, Kramer, and Spiegelman [14]. Briefly, if a given oligonucleotide is superior because of a particular sequence, a complementary sequence is produced and is incorporated into the chain. As the protein biosynthetic machinery became more elaborate and larger, the size of the tRNA molecule had to increase to accomodate them. The final size was predicated by the emergence of ribosomes, which are quite large organelles.

[1] S. L. Miller, Science **117,** 528 [1953].
[2] S. L. Miller and L. E. Orgel, in The Origins of Life on the Earth, p. 112, Prentice-Hall, Englewood Cliffs, N.J. 1974.
[3] R. A. Sanchez and L. E. Orgel, J. Mol. Biol. **47,** 531 [1970].
[4] H. A. Simon, in Hierarchy Theory (H. H. Pattee and G. Braziller, ed.), p. 3, New York 1973.
[5] M. O. Dayhoff, Atlas of Protein Sequence and Structure, **Vol. 5,** Natl. Biomed. Res. Fdn. Washington D.C. 1972.
[6] L. D. Greller and S. Erhan, Int. J. Peptide Prot. Res. **6,** 165 [1974].
[7] A. D. McLachlan, J. Mol. Biol. **61,** 409 [1971].
[8] S. Erhan and L. D. Greller, Int. J. Peptide Prot. Res. **6,** 174 [1974].
[9] T. R. Marzolf, L. D. Greller, and S. Erhan, Int. J. Biomed. Computing Accepted for publication 1977.
[10] S. W. Fox, in The Origins of Prebiological Systems (S. W. Fos, ed.), p. 361, Academic Press, New York 1965.
[11] A. I. Oparin, in Origins of Life, Dover Publishing Co. Inc., New York 1953.
[12] J. C. Lacey and K. M. Pruitt, Nature **223,** 799 [1969].
[13] C. Woese, Proc. Nat. Acad. Sci. U.S. **54,** 1546 [1965].
[14] D. R. Mills, F. R. Kramer, and S. Spiegelman, Science **180,** 916 [1973].