Evolution of *E. coli* $tRNA^{Ile}$: Evidence of Derivation from Other tRNAs

Mark P. Staves, David P. Bloch

Department of Botany, University of Texas at Austin, Austin, TX 78713 (USA)

and

James C. Lacey, Jr.

Department of Biochemistry, University of Alabama at Birmingham, Birmingham, AL 35294 (USA)

Z. Naturforsch. **42c,** 129–133 (1987); received June 13/September 4, 1986

Transfer RNA, Evolution, Codon Assignments, Sequence Homology, Hydrophobicity

Two *E. coli* $tRNA^{Ile}$ sequences were compared against those of 36 other *E. coli* tRNAs. $tRNA^{Ile\ 1}$ was found to bear high similarity with $tRNA^{Val\ 1}$ ($E = 1.11 \times 10^{-18}$) while $tRNA^{Ile\ 2}$ had the greatest match ($E = 3.40 \times 10^{-19}$) with $tRNA^{Lys1}$ (E is the expected number of such matches, per search, based on coincidence). These matches, which we consider to represent homologies, extend from base 7 to base 67 in the former and base 7 to the end (76) in the latter pair. These results coupled with others on the lower activity of isoleucine in reactions postulated to be important in primitive protein synthesis (*i.e.* esterification reactions and non-enzymatic activation by ATP [1–3]) lead us to propose that isoleucine was included among the proteinaceous amino acids, and received its anticodonic assignment, relatively late in evolution through mutation of tRNAs previously employed for other amino acids.

Introduction

Earlier studies related to the origin of coding assignments have shown that there are statistically significant correlations between the hydrophobicities of most amino acids and their respective anticodonic dinucleoside (positions YZ of anticodons XYZ) monophosphates [4]. However, there are four anticodonic assignments, isoleucine, tryptophan, tyrosine and serine (the four XGA assignments for serine) which do not fit these correlations. The anticodons for these amino acids have hydrophobicities that overlap considerably those of the anticodons of other amino acids [4]. In addition, the tryptophan and tyrosine anticodonic dinucleotides (CA and UA) are also the anticodonic equivalents of the present day terminator codons. These facts suggest that perhaps these particular dinucleotides were not sufficiently good discriminators in an early coding mechanism and therefore were not used as anticodons. They may have been used as terminators early in evolution and later some of them assigned to newly introduced amino acids [4]. The plausibility of this suggestion hinges on: 1. evidence for the later inclusion of these amino acids among the proteinaceous amino acids and; 2. evidence that their tRNAs might have been derived from preexisting tRNAs. In recent studies, it has been shown that isoleucine is extremely unreactive in several reactions which are relevant to protein synthesis. These reactions include the esterification of 5′-AMP [1] and poly A [2] by Ac–Ile imidazolides, and the non-enzymatic activation of isoleucine by ATP in the presence of Mg^{2+} [3]. In addition, the Ac–Ile imidazolide hydrolysis rate constant was found to be the lowest of those investigated [2]. These data are in accord with the idea that, because of its poor reactivity, isoleucine was late in being included among the proteinaceaous amino acids. To address the second point above, we explore here similarities between the sequences of $tRNA^{Ile}$ and other tRNAs. A high degree of matching would indicate homology, *i.e.* common ancestry, and would be consistent with relatively recent divergence of the two sequences. We focused primarily on *E. coli* because of the large number of sequences available for comparison.

Methods

Sequences of $tRNA^{Ile\ 1}$ and $tRNA^{Ile\ 2}$ from *E. coli* were compared with 36 different *E. coli* tRNA sequences, all taken from Sprinzl and Gauss [5], using the Los Alamos routines described by Goad and Kanehisa [6]. Each $tRNA^{Ile}$ sequence was compared

Reprint requests to M. P. Staves.

Verlag der Zeitschrift für Naturforschung, D-7400 Tübingen
0341–0382/87/0100–0129 $ 01.30/0

with each other *E. coli* tRNA sequence and the expected value (E) of the match was calculated using the formulae of Goad and Kanehisa [6].

$$E = \frac{(L-2)!}{(m-2)!u!g_{t1}!g_{t2}!} q^m(1-q)^u(L_s - L + 1)$$

$$q = fA_{t1}fA_{t2} + fC_{t1}fC_{t2} + fG_{t1}fG_{t2} + fU_{t1}fU_{t2}$$

where L is the overall length of the matching region, including mismatched bases and gaps, L_s the length of the shorter of the two tRNAs, m the number of matched pairs, u the number of mismatched pairs, and g_{t1} and g_{t2} the number of gaps in the two tRNA sequences. A, G, C and U signify the four bases and fA_{t1}, for example, is the frequency (*i.e.* number of A's divided by total number of nucleotides) of A in the first tRNA and fA_{t2} is the frequency of A in the second tRNA. A single term relating the lengths of the RNAs searched to E was used because the matches fell into corresponding regions of the two tRNAs. When two matches in corresponding regions were found on any comparison (*i.e.* an interrupted match) E was calculated as follows:

$$E = (P_1)(P_2)2 \sum_{i=1}^{i=x} i.$$

P_1 and P_2 are the probabilities of the match (calculated by using the Goad and Kanehisa formulae) and X is L_s minus the lengths of the two matching regions plus one.

The calculated value of E is an estimate, with the number of gaps resulting from insertions or deletions determining its accuracy [7]. The matches described in this paper contain between 0 and 2 gaps in matches 10 to 70 bases long. With this many gaps in a match only 15–20 bases long, the E value still falls within a factor of 2 or 3 of the observed E value using scrambled sequences (data unpublished). Matching regions with a larger number of gaps are rarely found because of penalties assigned to mismatches and gaps when using the formulae [6, 8]. The formulae of Goad and Kanehisa as used here provide an estimate of the expected value of a match and, most importantly, an objective way of identifying the pairs of tRNAs with the most extensive homologous regions.

Results and Discussion

The results of the comparisons of $tRNA^{Ile\ 1}$ and $tRNA^{Ile\ 2}$ with one another and with each other *E. coli* tRNA are presented in Fig. 1. Each of the isoleucine tRNAs shows greater matches with the tRNAs of some of the other amino acids than they do with each other. $tRNA^{Ile\ 1}$ shows greatest homology ($E = 1.11 \times 10^{-18}$) with $tRNA^{Val\ 1}$, but also has high homology with $tRNA^{Thr\ 1}$ ($E = 3.96 \times 10^{-18}$). Because threonine is the biosynthetic precursor of isoleucine and the synthesis of valine and isoleucine share common reactions and threonine, valine and isoleucine are near each other in the genetic anticode (Table I) these data support the proposals of Wong [9–11] that the code evolved along the biosynthetic pathways of the amino acids. Additionally, $tRNA^{Ile\ 1}$ has high homology with the other valine-encoding tRNAs. $tRNA^{Ile\ 2}$ shows an even greater homology ($E = 3.40 \times 10^{-19}$) with $tRNA^{Lys\ 1}$, but also very high homology ($E = 6.51 \times 10^{-13}$) with $tRNA^{Asn\ 1}$, which in turn is quite homologous with $tRNA^{Lys\ 1}$. It should be noted that these homologies are very extensive and cover most of the lengths of these tRNA se-

Decreasing hydrophobicity →

5′-End		Middle position A	Middle position G	Middle position C	Middle position U		3′-End
	A	Phe	Ser	Cys	Tyr	A	
	G	Phe	Ser	Cys	Tyr		
	C	Leu	Ser	Trp	T		
	U	Leu	Ser	T	T		
	A	Leu	Pro	Arg	His	G	
	G	Leu	Pro	Arg	His		
	C	Leu	Pro	Arg	Gln		
	U	Leu	Pro	Arg	Gln		
	A	Val	Ala	Gly	Asp	C	
	G	Val	Ala	Gly	Asp		
	C	Val	Ala	Gly	Glu		
	U	Val	Ala	Gly	Glu		
	A	Ile	Thr	Ser	Asn	U	
	G	Ile	Thr	Ser	Asn		
	C	Met	Thr	Arg	Lys		
	U	Ile	Thr	Arg	Lys		

Decreasing hydrophobicity ↓

Table I. The genetic anticode organized according to decreasing hydrophobicity of the mononucleotides. The assignments listed here are the anticodons predicted from Watson-Crick base pairing to the established codons. In reality, the anticodons found in tRNAs are seldom those shown. The variation is primarily in the wobble position (5′-end), where inosine is frequently found, and where A–U, G–C base pairing is not rigidly maintained (Modified from Lacey and Weber [19]).

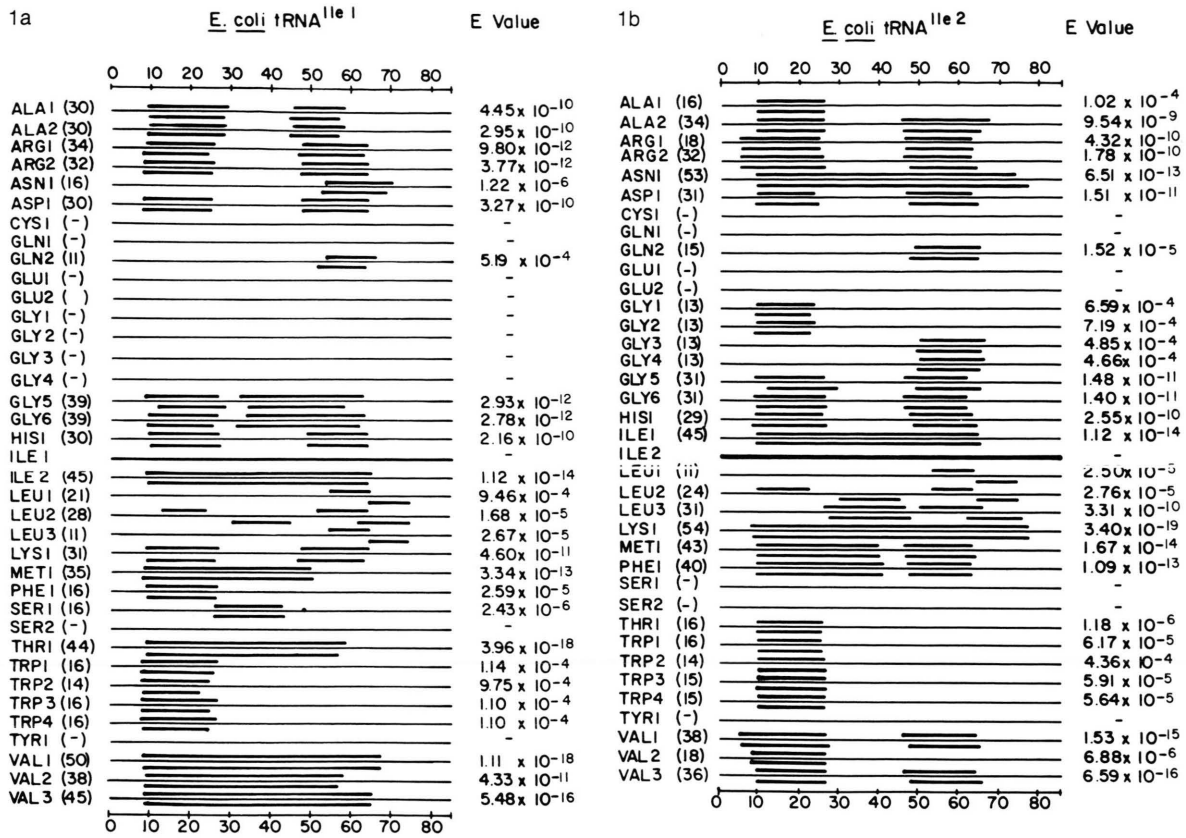

Fig. 1a and b. With each set of lines in this figure, we compare the locations of sequences in the tRNA identified at the left (short lines above the continuous line) which match sequences in a) $tRNA^{Ile\ 1}$ and b) $tRNA^{Ile\ 2}$ (short lines below the continuous line). The positions along the tRNA lengths are shown by the numbers at the top and bottom of the figure. The total number of matching bases in the matching regions is given by the number in parentheses at the left. The symbol (−) does not indicate no matching bases are present, rather that no matching sequence with $E < 1 \times 10^{-3}$, was found. Also it should be noted that the indicated homologous regions are not necessarily identical, *i.e.*, the program of Goad and Kanehisa [6] allows mismatches and gap insertion within certain limits. For example, the $tRNA^{Ile\ 1} - tRNA^{Val\ 1}$ homology from residue 7 to 67 includes no gaps, 50 matches and 11 mismatches. The expected numbers of the matches occuring by chance in a single comparison is shown at the right as calculated from the equations given in Methods. In a small number of comparisons short displaced matches were found. These offset homologies which are included in a longer homologous sequence are not counted in the number of matching bases at left nor in the calculated E values.

quences. The $tRNA^{Ile\ 1} - tRNA^{Val\ 1}$ homology extends from base 7 to 67 (>79% of the molecule) while the $tRNA^{Ile\ 2} - tRNA^{Lys\ 1}$ homology extends from base 7 to 76 (>92%). Matching regions with such low E values were found only rarely. Among comparisons of all *E. coli* tRNA sequences, excluding isoaccepting tRNAs, 20 of 528 or 3.8% of the matches yielded matches of $E < 10^{-15}$.

In addition to these matches found in comparisons among *E. coli* tRNAs, similar matches between corresponding tRNAs were found when comparisons were made among tRNAs from divergent organisms (data unpublished). For example, in comparisons among tRNAs from *Halobacterium volcanii* (an archaebacterium) the best match among non-isoaccepting tRNAs was found between $tRNA^{Ile}$ and $tRNA^{Lys}$. When tRNAs from yeast were compared, the best match with $tRNA^{Ile}$ was with $tRNA^{Val}$. These data show that the relationships among *E. coli* tRNAs reported here are mirrored in distantly re-

lated groups. This in turn makes less plausible the argument that the relationships among *E. coli* tRNAs described here are due to chance or to convergence after the separation of archebacteria, eubacteria and eukaryotes.

Because valine and isoleucine are next to each other in the genetic anticode (Table I) only one nucleotide change would be required to change the anticodon from that specific for valine to one specific for isoleucine. The bulk of the other differences were in the anticodon stem and in the acceptor stem. These are shown in Table IIa. Table IIb shows the changes required to go from $tRNA^{Lys}$ to $tRNA^{Ile\ 2}$ or *vice versa*. Again, aside from the anticodon itself, changes mainly occured in the anticodon and acceptor stems. One might ask whether these regions are responsible for the specificity of charging. This idea is also in accord with the observation that these two tRNAs have high general homology with most other tRNAs in the dihydrouridine (DHU) loop and stem regions (generally residues 9–25) and the TΨC loop and stem regions (generally residues 45–65).

Several evolutionary pathways are possible; for example $tRNA^{Val\ 1} \rightarrow tRNA^{Ile\ 1} \rightarrow tRNA^{Ile\ 2} \rightarrow tRNA^{Lys\ 1}$. However, because the $tRNA^{Ile\ 1}-tRNA^{Val\ 1}$ homology and the $tRNA^{Ile\ 2}-tRNA^{Lys\ 1}$ homology are greater than the $tRNA^{Ile\ 1}-tRNA^{Ile\ 2}$ homology, it seems more plausible that $tRNA^{Val\ 1} \rightarrow tRNA^{Ile\ 1}$ and $tRNA^{Lys\ 1} \rightarrow tRNA^{Ile\ 2}$ occured independently, the two isoleucine tRNAs having different origins. This case of independent origin and convergence of tRNA function may not be unique. There is also evidence for the convergence of serine (XGA) and serine (GCU) tRNAs (Staves, unpublished data). Of related interest is the proposal by Mullins *et al.* [12] that *E. coli* $tRNA^{Tyr}$ and *E. coli* 5S RNA share a common ancestral nucleotide sequence.

The inclusion of a new amino acid in protein synthesis would not only require the generation of a "new" (modified old) tRNA, but also a "new" aminoacyl-tRNA-synthetase. The transition from $tRNA^{Val} \rightarrow tRNA^{Ile}$ seems to be plausible because valine and isoleucine are both hydrophobic and are branched at the β-carbon. However, the transition from $tRNA^{Lys}$ to $tRNA^{Ile}$ would not follow, according to interpretations based on physical characteristics, because lysine is extremely hydrophilic and isoleucine is hydrophobic. Nevertheless, the apparent close relationship between these two molecules and the parsimony involved in the change make this pathway attractive. The aminoacyl-tRNA synthetases for leucine, isoleucine and valine share many properties (enzyme quaternary structure, subunit molecular weight and 2′/3′ charging specificity) whereas the lysyl-tRNA synthetase differs in these properties [13]. Thus, it seems most probable that the isoleucyl-tRNA synthetase evolved from the synthetase for leucine or valine and not from that of lysine.

These data support the hypothesis that the anticodonic assignments of isoleucine came later through

Table 2a. Changes required to convert $tRNA^{Val\ 1}$ to $tRNA^{Ile\ 1}$.

	A		B	C		B′	D	E	D′	F	G	H	G′	A′	
0910 *E. coli* Ile1	AGGCUUG	UA	GCUC	AGGU	GGUU	A	GAGC	GCACCC	CUGAUAA	GGGUG	AGGUC	GGUGG	UUCAAGU	CCACU	CAGGCCU	ACCA
2020 *E. coli* Val1	ĠĊGÙĊĊG	UA	GCUC	AGÙU	GGUU	A	GAGC	ȦCĊȦCC	ÙUGAĊAÙ	GGÙĠG	ĠGGUC	GGUGG	UUCĠAGU	CCACU	CĠGȦCGĊ	ACCA

Table 2b. Changes required to convert $tRNA^{Lys\ 1}$ to $tRNA^{Ile\ 2}$.

0911 *E. coli* Ile2	GGCCCCU	UA	GCUC	AGUG	GUU	A	GAGC	AAGCGA	CUNAUAA	UCGCU	UGGUC	GCUGG	UUCAAGU	CCAGC	AGGGGCC	ACCA
1110 *E. coli* Lys1	GGĠÙCĠU	UA	GCUC	AGUÙ	GĠU	A	GAGC	AĠÙÙGA	CUÙÙUAA	UCȦȦU	UGGUC	GCȦGG	UUCĠAȦU	CCÙGC	AĊGȦĊCC	ACCA

A, A′, Acceptor Stem; B, B′, DHU Stem; C, DHU Loop; D, D′, Anticodon Stem; E, Anticodon Loop; F, Variable Loop; G, G′, TΨC Stem; H, TΨC Loop.

Note: All residues are shown as unmodified. The sequences were taken from Sprinzl and Gauss [5] and the numbers to the left are their identification numbers. Dots between sequences show changes necessary to convert from one tRNA to the other.

mutation of previously existing tRNAs, and may explain the lack of correlation between the hydrophobicities of isoleucine and $tRNA^{Ile}$ described by Lacey and Mullins [4]. This in turn implies that the bulk of the anticodonic assignments involved a more direct interaction of amino acid and nucleotides. While some selective binding constants [14, 15] and chemical reactions [1–3] have been found between amino acids and anticodons, it is not yet clear how the assignments came about.

There are few recent reports on the evolution of tRNAs, partly because many workers believe, as do Holmquist *et al.* [16], that modern tRNAs have reached an equilibrium in a mutational sense and evidence of their evolutionary pathways has been lost because of divergence. While this may be true for some tRNAs, or parts of tRNAs, it does not hold for all cases. Most workers have concentrated on development of tRNAs in a phylogenetic context, although Holmquist *et al.* [16] and Schwartz *et al.* [17] did consider intraspecies comparisons. These workers placed valine and isoleucine on the same branch of an evolutionary tree. An excellent review appeared in 1981 by Cedergren *et al.* [18] on tRNA evolution; however, their treatment was more relevant to broad questions of tRNA evolution than to the particular questions we are asking here.

Also relevant to the study here is the data of Wetzel [13] who, in studying the evolution of aminoacyl-tRNA synthetases, concluded that the enzymes of phenylalanine, isoleucine, leucine, valine and methionine constitute a family which evolved last.

Acknowledgements

This work was supported by grants from the RGK Foundation and the Richard Lounsbery Foundation (to DPB), a grant of computer time from the Cray Corporation to the University of Texas and a grant (NGR 01–010–001) from the National Aeronautics and Space Administration (to JCL).

Note added in proof: We report with sadness the death of our friend, colleague and teacher David P. Bloch (10.2.26–23.10.86).

[1] D. W. Mullins Jr. and J. C. Lacey Jr., J. Mol. Evol., in the press.
[2] J. C. Lacey Jr., L. M. Hall, and D. W. Mullins Jr., Origins Life, in the press.
[3] D. W. Mullins Jr. and J. C. Lacey Jr., Biochem. Biophys. Res. Comm. **96,** 491–497 (1980).
[4] J. C. Lacey Jr. and D. W. Mullins Jr., Origins Life **13,** 3–42 (1983).
[5] M. Sprinzl and D. H. Gauss, Nucleic Acids Res. **10,** r 1–55 (1982).
[6] W. B. Goad and M. I. Kanehisa, Nucleic Acids Res. **10,** 247–263 (1982).
[7] D. J. Lipman, W. J. Wilbur, T. F. Smith, and M. S. Waterman, Nucleic Acids Res. **12,** 215–226 (1984).
[8] M. Kanehisa, Nucleic Acids Res. **12,** 203–213 (1984).
[9] J. T. Wong, Proc. Nat. Acad. Sci. USA **72,** 1909–1912 (1975).
[10] J. T. Wong, Proc. Nat. Acad. Sci. USA **73,** 2336–2340 (1976).
[11] J. T. Wong, Proc. Nat. Acad. Sci. USA **77,** 1083–1086 (1980).
[12] D. W. Mullins Jr., J. C. Lacey Jr., and R. A. Hearn, Nature New Biol. **242,** 80–82 (1973).
[13] R. Wetzel, Origins Life **9,** 39–50 (1978).
[14] M. A. Khaled, D. W. Mullins Jr., M. Swindle, and J. C. Lacey Jr., Origins Life **13,** 87–96 (1983).
[15] M. A. Khaled, D. W. Mullins Jr., and J. C. Lacey Jr., J. Mol. Evol. **20,** 66–70 (1984).
[16] R. Holmquist, T. H. Jukes, and S. Pangburn, J. Mol. Biol. **78,** 91–116 (1973).
[17] R. M. Schwartz, M. J. Gantt, and M. O. Dayhoff, in: Atlas of Protein Sequence and Structure (M. O. Dayhoff, ed.), **Vol. 5,** suppl. 2, pp. 271–291, Nat. Biomed. Res. Found, Washington D.C. 1976.
[18] R. J. Cedergren, B. LaRue, and H. Grosjean, CRC Crit. Rev. Biochem. **11,** 35–104 (1981).
[19] J. C. Lacey Jr. and A. L. Weber, Precamb. Res. **5,** 1–22 (1977).